# Surveillance Event Detection (SED)

<table>
<thead>
<tr>
<th>Time</th>
<th>Presentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>11:10 – 11:30</td>
<td>Task Overview (NIST)</td>
</tr>
<tr>
<td>11:30 – 11:50</td>
<td>University of Ottawa (VIVA_uOttawa)</td>
</tr>
<tr>
<td>11:50 – 12:10</td>
<td>Dublin City University, CLARITY (dcu_savasa)</td>
</tr>
<tr>
<td>12:10 – 1:20</td>
<td>Lunch is served in the NIST West Square Cafeteria (please have your lunch ticket ready)</td>
</tr>
<tr>
<td>1:20 – 1:40</td>
<td>Carnegie Mellon University; IBM Research (CMU +IBM)</td>
</tr>
<tr>
<td>2:40 – 2:00</td>
<td>Discussion</td>
</tr>
</tbody>
</table>
2012 TRECVID Workshop:
Interactive Surveillance Event Detection (iSED) Task Overview

Jonathan Fiscus (NIST)
Martial Michel (Systems Plus, Inc.)

November 27, 2012
NIST, Gaithersburg, MD, USA
Motivation

• Surveillance Event Detection Motivation
  – SED addresses the need for automatic detection of events in large amounts of surveillance video
  – SED Challenges
    • Requires application of several Computer Vision techniques
    • Involves subtleties that are readily understood by humans, difficult to encode for machine learning approaches
    • Can be complicated due to clutter in the environment, lighting, camera placement, traffic, etc.

• Interactivity Motivation
  – SED remains a difficult task for humans and systems
  – Interactivity/relevance feedback have been effectively employed in other tasks
Surveillance Event Detection Tasks

• Interactive (iSED) Task: Given a textual description of an observable event of interest, at test time allow a searcher 25 minutes to filter incorrect event detections in a non-segmented corpus of video.

• Retrospective SED (rSED) Task: Given a textual description of an observable event of interest, automatically detect all occurrences of the event in a non-segmented corpus of video.

• Identify each detected event observation by:
  • The temporal extent (beginning and end frames)
  • A decision score: a numeric score indicating how likely the event observation exists with more positive values indicating more likely observations (normalized)
  • An actual decision: a boolean value indicating whether or not the event observation should be counted for the primary metric computation
Evaluation Source Data

- Reused same test data as SED ‘09, ‘10, and ‘11 evaluations
- UK Home Office collected CCTV video from 5 camera views at a busy airport
- Development Set
  - 100 hours of video
  - 10 events annotated on 100% of the data
- Evaluation Set
  - “iLIDS Multiple Camera Tracking Scenario Training set”
  - An identified 15-hours of the 45-hour set evaluated (NEW)
  - 10 events annotated on 1/3 of the data
    - 7 events evaluated
**Events and Instances per Hour (IpH)**

<table>
<thead>
<tr>
<th>Event Type</th>
<th>IpH</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Single Person events</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PersonRuns</td>
<td>7.02</td>
<td>Someone runs</td>
</tr>
<tr>
<td>Pointing</td>
<td>69.74</td>
<td>Someone points</td>
</tr>
<tr>
<td><strong>Single Person + Object events</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CellToEar</td>
<td>12.73</td>
<td>Someone puts a cell phone to his/her head or ear</td>
</tr>
<tr>
<td>ObjectPut</td>
<td>40.74</td>
<td>Someone drops or puts down an object</td>
</tr>
<tr>
<td><strong>Multiple People events</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Embrace</td>
<td>11.48</td>
<td>Someone puts one or both arms at least part way around another person</td>
</tr>
<tr>
<td>PeopleMeet</td>
<td>29.46</td>
<td>One or more people walk up to one or more other people, stop, and some communication occurs</td>
</tr>
<tr>
<td>PeopleSplitUp</td>
<td>12.27</td>
<td>From two or more people, standing, sitting, or moving together, communicating, one or more people separate themselves and leave the frame</td>
</tr>
</tbody>
</table>
Evaluation Protocol & Scoring Process

• Evaluation Plan
  http://www.nist.gov/itl/iad/mig/trecvid.cfm

• Framework for Detection Evaluation (F4DE) Toolkit
  http://www.nist.gov/itl/iad/mig/tools.cfm

• Four step evaluation process (for each event)
  1. Segment mapping
  2. Segment scoring
  3. Error metric calculation
  4. Error visualization
Step 1: Segment Mapping

1 Hour of Video

Reference Observations

System Observations

Utilizes the Hungarian Solution to Bipartite Graph Matching
Step 2: Segment Scoring

1 Hour of Video

Reference Observations

System Observations

Correct
Detections
When reference and system observations are mapped

Missed
Detections
When a reference observation is NOT mapped

False Alarms
When a system observation is NOT mapped
Step 3: Error Metric Computation
Compute Normalized Detection Cost Rate (NDCR) (1/2)

\[
P_{Miss} = \frac{\# MissedObs}{\# TrueObs}
\]

\[
Rate_{FA} = \frac{\# FalseAlarms}{SignalDuration}
\]

\[
P_{Miss} = \frac{2}{4} = .50
\]

\[
Rate_{FA} = \frac{1}{1Hr} = 1 FA / Hr
\]
Step 3: Error Metric Computation
Compute Normalized Detection Cost Rate (NDCR) (2/2)

Time

Reference Observations

System Observations

1 Hour of Video

Primary Metric

\[ NDCR = P_{Miss} + \frac{Cost_{FA}}{Cost_{Miss} * R_{TARGET}} * R_{FA} \]

\[ NDCR = 0.5 + \frac{1}{10 * 20} * 1 = 0.505 \]

Range of NDCR() is \([0:\infty)\)

- NDCR = 0.0 is a perfect system
- NDCR = 1.0 is equivalent to a system that outputs nothing

\[ Cost_{Miss} = 10 \]
\[ Cost_{FA} = 1 \]
\[ R_{TARGET} = 20 \]
Step 4: Error Visualization

Detection Error Tradeoff (DET) Curves \( (\text{Prob}_{\text{Miss}} \text{ vs. Rate}_{\text{FA}}) \)

- Compute \( \text{Rate}_{\text{FA}} \) and \( \text{P}_{\text{Miss}} \) for all \( \theta \)

\[
\begin{align*}
\text{MinimumNDCR}(\theta) &= \arg \min_{\theta} \left[ \text{P}_{\text{Miss}}(\theta) + \frac{\text{Cost}_{\text{FA}}}{\text{Cost}_{\text{Miss}} * R_{\text{TARGET}}} * \text{R}_{\text{FA}}(\theta) \right] \\
\text{ActualNDCR}(\text{Act.Dec.}) &= \text{P}_{\text{Miss}}(\text{Act.Dec.}) + \frac{\text{Cost}_{\text{FA}}}{\text{Cost}_{\text{Miss}} * R_{\text{TARGET}}} * \text{R}_{\text{FA}}(\text{Act.Dec.})
\end{align*}
\]

# 12 2012 SED Participants

(with number of systems per event)

<table>
<thead>
<tr>
<th>Single Person</th>
<th>Person + object</th>
<th>Multiple People</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person Runs</td>
<td>Pointing</td>
<td>CellToEar</td>
</tr>
<tr>
<td>SED</td>
<td>SED</td>
<td>SED</td>
</tr>
<tr>
<td>5 years in a row</td>
<td>Carnegie Mellon University &amp; IBM [CMU-IBM]</td>
<td>5</td>
</tr>
<tr>
<td>4 years in a row</td>
<td>Multimedia Communication and Pattern Recognition Labs, Beijing University of Posts and Telecommunications [BUPT-MCPRL]</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Peking University, NEC Laboratories [PKUNEC]</td>
<td>3</td>
</tr>
<tr>
<td>3 years in a row</td>
<td>Beijing Jiaotong University [BJTU-SED]</td>
<td></td>
</tr>
<tr>
<td>NEW</td>
<td>Brno University of Technology [BrnoUT]</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Dublin City University [dcu-savasa]</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>The City College of New York Media Team [MediaCCNY]</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Institute of Computer Science and Technology, Peking University [PKU-OS]</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Queensland University of Technology [sait]</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shanghai Jiaotong University, Center for Brain-like Computing and Machine Intelligence [SJTUBCMI]</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Video Computing Group, University of California Santa Barbara [UcsbUcrVcg]</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>University of Ottawa [VIVA-uOttawa]</td>
<td>1</td>
</tr>
<tr>
<td>Total Interactive Event Runs</td>
<td>108</td>
<td></td>
</tr>
<tr>
<td>Total Retrospective Event Runs</td>
<td>95</td>
<td></td>
</tr>
</tbody>
</table>
Event-Averaged, Lowest NDCR by Site
ISED vs. rSED

- 8 sites submitted both iSED and rSED runs
- 5 reduced NDCR
  - BJTU-SED 1%
  - BrnoUT 19%
  - BUPT-MCPRL 29%
  - CMU-IBM 5%
  - dcu-savasa 1%
BUPT PersonRuns

PersonRuns iSED  vs. rSED - BUPT
PersonRuns Event
iSED vs. rSED
ObjectPut Event

iSED vs. rSED - BUPT
ObjectPut Event
iSED vs. rSED

BUPT

BrnoUT

CMU-IBM
PeopleMeet Event
iSED vs. rSED - BUPT
PeopleMeet Event
iSED vs. rSED

BUPT

BrnoUT

CMU-IBM
Person Runs
SED ‘09-’12
rSED + iSED
PeopleSplitUp
SED ’09-’12
rSED + iSED

Human Group
Embrace SED ‘09–’12  
rSED + iSED
Cell To Ear
SED ‘09-’12
rSED + iSED
Pointing
SED ‘09-‘12
rSED + iSED
PeopleMeet
SED ‘09-’12
rSED + iSED
Object Put
SED ‘09-’12
rSED + iSED
Conclusions

• Interactive systems for 5 of 12 participants yielded more accurate detections
  – BUPT Actual NDCR reduction of 29%
• Single-person and multi-person events show evidence of yearly improvements
  – Single-person: PersonRuns, PeopleSplitUp, Pointing
  – Multi-person: PeopleMeet, Embrace
  – But... still not approaching human performance
• Person+object events remain difficult
  – ObjectPut, CellToEar
• Last year for the SED track
  – Thanks to all who participated!
  – Thanks to the Linguistic Data Consortium for annotations!
  – Special thanks to the iLIDS team for the data!