Video Surveillance Event Detection Track
The TRECVID 2009 Evaluation
Jonathan Fiscus, Martial Michel, John Garofolo, Paul Over
NIST
Heather Simpson, Stephanie Strassell
LDC
Motivation

• **Problem:** automatic detection of *observable* events of interest in surveillance video

• **Challenges:**
  – requires application of several Computer Vision techniques
    • segmentation, person detection/tracking, object recognition, feature extraction, etc.
  – 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.
Evaluation Source Data

- UK Home Office collected CCTV video at a busy airport
  - 5 Camera views: (1) controlled access door, (2) waiting area, (3) debarkation area, (4) elevator close-up, (5) transit area
- Development data resources:
  - 100 camera hours of video from the 2008 VSED Track
    - Complete annotation for 10 events on 100% of the data
- Evaluation data resources:
  - 45 camera hours of video from the iLIDS Multiple Camera Tracking Scenario Training data set
    - Complete annotation for 10 events annotated on 1/3 of the data
    - Also used for the AVSS 2009 Single Person Tracking Evaluation
TRECVID VSED
Retrospective Event Detection

• Task:
  – Given a textual description of an observable event of interest in the airport surveillance domain, configure a system to detect all occurrences of the event
  – Identify each event observation by:
    • The temporal extent
    • A detection score indicating the system’s confidence that the event occurred
    • A binary decision on the detection score optimizing performance for the primary metric
TRECVID
VSED Freestyle Analysis

• Goal is to support innovation in ways not anticipated by the retrospective task

• Freestyle task includes:
  – rationale
  – clear definition of the task
  – performance measures
  – reference annotations
  – baseline system implementation
Event Annotation Guidelines

• Jointly developed by NIST, Linguistic Data Consortium (LDC), Computer Vision Community
  – Event Definitions left minimal to capture human intuitions

• Updates from 2008 guidelines:
  – Based on annotation questions from 2008 annotation
  – End Time Rule:
    • If Event End Time = a person exiting the frame boundary, frame for end time should be the earliest frame when their body and any objects they are carrying (e.g. rolling luggage) have passed out of the frame. If luggage remains in the frame not moving, can assume person left the luggage and tag at person leaving the frame.
  – People Meet/Split Up rules:
    • If people leave a group but do not leave the frame, the re-merging of those people do not qualify as PeopleMeet
    • If a group is standing near the edge of the frame, people are briefly occluded by frame boundary but under RI rule have not left the group, that is not PeopleSplitUp
  – Some specific case examples added to Annotator guidelines
Annotation Tool and Data Processing

• No changes from 2008
  – Annotation Tool
    • ViPER GT, developed by UMD (now AMA)
    • NIST and LDC adapted tool for workflow system compatibility
  – Data Pre-processing
    • OS limitations required conversion from MPEG to JPEG
      – 1 JPEG image for each frame
    • For each video clip assigned to annotators
      – Divided JPEGs into framespan directories
      – Created .info file specifying order of JPEGs
      – Created ViPER XML file (XGTF) with pointer to .info file
    • Default ViPER playback rate = about 25 frames (JPEGs)/second
Annotation Workflow Design

• Clip duration about same or smaller than 2008
• Rest of workflow revised based on 2008 annotations and experiments
  – 3 events per work session for 9 events
  – 1 pass by senior annotator over ElevatorNoEntry for Camera 4 only
    • ElevatorNoEntry very infrequent, only 1 set of elevators which are easy to see in Camera 4 view
    • Camera 4 ElevatorNoEntry annotations automatically matched to corresponding timeframe in other camera views
  – 3 passes over other 9 events for 14 hours of video
    • (2008 – 1 pass over all 10 events for 100 hours of video)
  – Additional 6 passes over 3 hour subset of video
• Adjudication performed on 3x and 9x annotations
  – 2008 Adjudication performed on system + human
Event Sets

- 3 sets of 3 events, ElevatorNoEntry separate set
- Goal to balance sets by event type and frequency

<table>
<thead>
<tr>
<th>Event Type</th>
<th>Tracking</th>
<th>Object</th>
<th>Gesture</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 1</td>
<td>OpposingFlow</td>
<td>CellToEar</td>
<td>Pointing</td>
</tr>
<tr>
<td>Set 2</td>
<td>PeopleSplitUp</td>
<td>ObjectPut</td>
<td>Embrace</td>
</tr>
<tr>
<td>Set 3</td>
<td>PeopleMeet</td>
<td>TakePicture</td>
<td>PersonRuns</td>
</tr>
</tbody>
</table>
Visualization of Annotation Workflow

Events

- E1
- E2
- E3
- E4
- E5
- E6
- E7
- E9
- E8

Annotators

- A1
- A2
- A3
- A1
- A1
- A1
- A1

Video

<= ~5 minute video clip

Senior Annotator

ElevatorNoEntry

A1 (Camera 4 only)

E10
Annotation Challenges

• Ambiguity of guidelines
  – Loosely defined guidelines tap into human intuition instead of forcing real world into artificial categories
  – But human intuitions often differ on borderline cases
  – Lack of specification can also lead to incorrect interpretation
    • Too broad (e.g. baby as object in ObjectPut)
    • Too strict (e.g. person walking ahead of group as PeopleSplitUp)

• Ambiguity and complexity of data
  – Video quality leads to missed events and ambiguous event instances
    • Gesturing or pointing? ObjectPut or picking up an object? CellToEar or fixing hair?

• Human factors
  – Annotator fatigue a real issue for this task
  – Lower number of events per work session helps

• Technical issues
### 2009 Participants

**11 Sites (45 registered participants)**

**75 Event Runs**

<table>
<thead>
<tr>
<th>Single Person</th>
<th>Single Person + Object</th>
<th>Multiple People</th>
</tr>
</thead>
<tbody>
<tr>
<td>ElevatorNoEntry</td>
<td>OpposingFlow</td>
<td>PersonRuns</td>
</tr>
<tr>
<td>SJTU</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>UAM</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>CMU</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>NEC-UIUC</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>NHKSTRL</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>BUPT-MCPRL</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>BUPT-PRIS</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>PKU-IDM</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>SFU</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>TITGT</td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Toshiba</td>
<td></td>
<td>x</td>
</tr>
</tbody>
</table>

**Total Participants per Event**

<table>
<thead>
<tr>
<th>2008</th>
<th>2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>

**New**
Observation Durations and Event Densities
Comparing 2008 and 2009 Test Sets

Average Duration of Instances

- Elevator
- Opposing Flow
- Person Runs
- Pointing
- Cell to Ear
- Object Put
- Take Picture
- Embrace
- People Meet
- People Split Up

<table>
<thead>
<tr>
<th>Seconds per Instance</th>
<th>2008</th>
<th>2009</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elevator</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>Opposing Flow</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>Person Runs</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Pointing</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Cell to Ear</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Object Put</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>Take Picture</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Embrace</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>People Meet</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>People Split Up</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Rates of Event Instances

- 95% more for Cam2 (Waiting Area)
- 50% more for Cam3 (Debarkation Area)
Evaluation Protocol Synopsis

• NIST used the Framework for Detection Evaluation (F4DE) Toolkit
  • Available for download on the VSED Web Site
  • http://www.itl.nist.gov/iad/mig/tools
• Events are scored independently
• Five step evaluation process
  • Segment mapping
  • Segmented scoring
  • Score accumulation
  • Error metric calculation
  • Error visualization
Segment Mapping for Streaming Media

• Mapping kernel function
  – The mid point of the system-generated extent must be within the reference extent extended by 1 sec.
  – Temporal congruence and decision scores give preference to overlapping events
Segment Scoring

1 Hour of Video

Ref. Obs.

Sys. Obs.

Time

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
Compute Normalized Detection Cost

Ref. Obs.

Sys. Obs.

1 Hour of Video

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

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

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

\[ Rate_{FA}() = \frac{1}{1Hr} = 1FA/Hr \]
Compute Normalized Detection Cost Rate

\[ NDCR() = P_{Miss}() + \frac{Cost_{FA}}{Cost_{Miss} * R_{Target}} * R_{FA}() \]

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

Range of NDCR() is \([0:\infty)\)
NDCR() = 1.0 is a system that outputs nothing

Event Detection Constants
\[ Cost_{Miss} = 10 \]
\[ Cost_{FA} = 1 \]
\[ R_{Target} = 20 \]
Decision Error Tradeoff Curves

$\text{Prob}_{\text{Miss}} \ vs. \ Rate_{\text{FA}}$

Decision Score Histogram

Count of Observations

Decision Score

Full Distribution
Decision Error Tradeoff Curves

\[ \text{Prob}_{\text{Miss}} \text{ vs. Rate}_{\text{FA}} \]

**Decision Score Histogram Separated wrt. Reference Annotation s**

Count of Observations

- **Non-Targets**
- **Targets**

System Decision Score

\[ P_{\text{Miss}}(\theta) = \frac{\#\text{MissedObs}}{\#\text{TrueObs}} \]

\[ \text{Rate}_{\text{FA}}(\theta) = \frac{\#\text{FalseAlarms}}{\text{SignalDuration}} \]

Normalization by \# of Non-Observations is impossible for Streaming Detection Evaluations
Decision Error Tradeoff Curves

Prob\_Miss vs. Rate\_FA

Compute Rate\_FA and P\_Miss for all θ

MinimumNDCR(θ) = \arg\min_θ \left[ P\_Miss(θ) + \frac{Cost\_FA}{Cost\_Miss * R\_Target} * R\_FA(θ) \right]
Decision Error Tradeoff Curves

**Actual vs. Minimum NDCR**

\[
\text{MinimumNDCR}(\theta) = \arg\min_{\theta} \left[ P_{\text{Miss}}(\theta) + \frac{\text{Cost}_{FA}}{\text{Cost}_{Miss} \ast R_{Target}} \ast R_{FA}(\theta) \right]
\]

\[
\text{ActualNDCR}(\text{Act.Dec.}) = P_{\text{Miss}}(\text{Act.Dec.}) + \frac{\text{Cost}_{FA}}{\text{Cost}_{Miss} \ast R_{Target}} \ast R_{FA}(\text{Act.Dec.})
\]

**Event Detection Constants**

\[
\begin{align*}
\text{Cost}_{Miss} &= 10 \\
\text{Cost}_{FA} &= 1 \\
R_{Target} &= 20
\end{align*}
\]
2009 Event Detection Results
2009 Minimum and Actual NDCRs (Set 1)
2009 Minimum and Actual NDCRs (Set 2)

Normalized Detection Cost Rate

- PeopleMeet
- PeopleSplit
- PersonRuns
- Pointing
- TakePicture

Minimum
Act. Dec
2009 Best DET Curves for Events
2008 Best DET Curves for Events

Events comparison

PMiss (in %)

0.01 0.1 1 10 100 1000
RFA (in Events/Hour)

2008
2009 Best DET Curves for Events

- Did performance really decrease?
- Did 2nd year participants improve?
- Is this a test set difference?
- Did 3-Way annotation make a “harder” test set?
Embrace Event
Best Submission per Site

Cross System DET for Embrace Event

Iso-DCR lines
CMU / VCUBE
NEC/UIUC / none
PKU-IDM / eSur
SFU / match
SJTU / baseline
RandomDET
Human Pass 1
Human Pass 2
Human Pass 3
Eval08 Best

Min NDCR
Act. NDCR
2008 vs. 2009 Minimum NDCRs
Conditioned by Selected Events and Cameras

NDRC

Cam1 Cam2 Cam3 Cam5 Cam1 Cam2 Cam3 Cam5 Cam1 Cam2 Cam3 Cam5 Cam1 Cam2 Cam3 Cam5
CellToEar Embrace ObjectPut PeopleMeet PeopleSplitUp PersonRuns Pointing

Range
Max(08,09) Min(08,09) 2009
2008 vs. 2009 Minimum NDCRs

Limited to 2\textsuperscript{nd} Year Participants
Conditioned by Selected Events and Cameras

- Max(08,09)
- Min(08,09)
- 2009
CMU 2008 and 2009 Embrace Event Submissions

Cross Year Comparison of the Embrace Event across cameras for

Iso-DCR lines

All Cameras TransitArea 2008
All Cameras TransitArea 2009

Pmiss (in %)

RFA (in Events/Hour)
CMU 2008 and 2009 Embrace Event
Submissions Split By Cameras

Why is the ALL Camera Curve worse than each SINGLE Camera Curves?
Conclusions and Lessons Learned

• Improvement can be seen in 2 of the events on specific cameras
• Multiple-year participants have shown improvement on 3 events
  – Decision score normalization is important
  – Non-optimal normalization obscures performance gains
• The change in annotation scheme improved the number of found event instances
  – We will be studying the effect on scoring
• Next year’s evaluation should re-use this year’s test set but in what manner
End of Talk

Back up slides to follow
PeopleMeet Event
Best Submission per Site

Cross System DET for PeopleMeet Event

- Iso-DCR lines
- CMU / VCUBE
- NHKSTRL / NHK-SYS1
- PKU-IDM / eSur
- SJTU / baseline
- TITGT / EVAL
- RandomDET
- Human Pass 1
- Human Pass 2
- Human Pass 3
- Eval08 Best

Min NDCR
Act. NDCR
PersonRuns Event
Best Submission per Site

Cross System DET for PersonRuns Event

Min NDCR
Act. NDCR
Person Runs Limited to Participants of Both 2008 and 2009
Pointing Event
Best Submission per Site

Cross System DET for Pointing Event

- Iso-DCR lines
- BUPT-MCPRL / baseline
- CMU / VCUBE
- NEC/UIUC / N2
- SFU / match
- SJTU / baseline
- RandomDET
- Human Pass 1
- Human Pass 2
- Human Pass 3
- Eval08 Best

- Random system
  - $R_{targ} = XXX$
  - MeanDur=XXs
  - TestDur=XXH

- Min NDCR
- Act. NDCR
ObjectPut Event
Best Submission per Site

Cross System DET for ObjectPut Event

- Random system
- $R_{\text{targ}} = \text{XXX}$
- MeanDur=XXs
- TestDur=XXH
OpposingFlow Event
Best Submission per Site

- BUPT-MCPRL / baseline
- BUPT-PRIS / baseline
- CHU / VCUBE
- NHKSTRL / NHK-SYS1
- SJTU / baseline
- Toshiba / cohog
- UAM / baseline
- Human Pass 1
- Human Pass 2
- Human Pass 3
- Eval08 Best

Phase (in 2)

RFA (in Events/Hour)
PeopleSplitUp Event
Best Submission per Site

Cross System DET for PeopleSplitUp Event

- Random system
- $R_{\text{targ}} = XXX$
- MeanDur=XXs
- TestDur=XXH
TakePicture Event
Best Submission per Site

Cross-System DET for TakePicture Event

Iso-DCR lines
- BUPT-MCPRL / baseline
- CMU / VCUBE
- SJTU / baseline
- Human Pass 1
- Human Pass 2
- Human Pass 3
- Eval08 Best

Min NDCR
Act. NDCR
CellToEar Event
Best Submission per Site

Cross System DET for CellToEar Event

- Random system
  - \( R_{\text{targ}} = XXX \)
  - MeanDur=XXs
  - TestDur=XXH

- CMU / VCUBE
- NEC/UIUC / N1
- RandomDET
- Human Pass 1
- Human Pass 2
- Human Pass 3
- Eval08 Best

- Min NDCR
- Act. NDCR

- Iso-DCR lines

- Time (in H)
  - 0 2 4 6 8 10
ElevatorNoEntry Event
All Submissions