IBM-Northwestern@TRECVID 2013: Surveillance Event Detection (SED)

Yu Cheng ɨ*, Lisa Brown ɨ, Quanfu Fan ɨ, Rogerio Feris ɨ, Alok Choudhary *, Sharath Pankanti ɨ

ɨ IBM T. J. Watson Research Center
* Northwestern University
Outline

• Retrospective Event Detection
  – System Overview
  – Temporal Modeling for Event Detection
  – Performance Evaluation

• Interactive Event Detection
  – Interactive Visualization
  – Risk Ranking
  – Performance Evaluation
System Overview (CMU-IBM 2012)

**Offline Training**
- Sliding Window
- Training Sequence 1
- Training Sequence n

**Extract MoSIFT [1] features**
- Fisher Vector coding [2]

**Online Testing**
- Sliding Window
- Testing Sequence

**Linear SVM training**
**Hard samples mining**

**Classification**

**Model**

**Detection Result**

**NMS**
System Overview (IBM 2013)

Test Sequence

\[ y_{i-1}, y_i, y_{i+1} \] 


Training Sequence 1

\[ \text{Training Sequence } n \] 


Sequence Temporal Learning
Temporal Modeling

• Motivation:
  – Rich temporal patterns exhibit among visual events.
  – Exploiting temporal dependencies to enhance event detection.
Joint Segmentation and Detection

• Overall Framework:
  – A quadratic integer programming approach combining classification and temporal dependencies between events.
  – For an arbitrary segmentation \( S = \{S_1, S_2, \ldots, S_n\} \) of \( X \) where \( S_i = X(t_i, t_{i+1}) \) \( (t = \{t_1, t_2, \ldots, t_{n+1}\} \) are transition points, the quality of the segmentation can be measured by:

\[
\mu \sum_{i=1}^{n} \sum_{k=1}^{K} \zeta_i^k \varphi_i^k(S_i) + (1 - \mu) \sum_{j=1}^{n'} \sum_{j'=j+1}^{n} \sum_{k=1}^{K} \sum_{k'=1}^{K} p(k, k') \zeta_j^k \zeta_{j'}^{k'}
\]

\[n' \leq n\]
\[\forall i : \sum_{k=1}^{K} \zeta_i^k \leq 1\]
\[\forall i, \forall i', \forall k, \forall k' : \zeta_i^k + \zeta_i^{k'} \leq 1 \quad \text{if} \quad S_i \cap S_j = 0\]
Joint Segmentation and Detection

• Classification Model:
  – Trained discriminatively using multiclass SVM [3] at different window sizes (30, 60, 90 and 120 frames)
  – Non-event is treated as a special null class

• Model Solver:
  – If only first-order dependency is considered, the objective function can be re-written as:
    \[ f(X, K) = \mu \sum_{i=1}^{n} \varphi^k(S_i) + (1 - \mu) \sum_{j=1}^{n} p(j - 1, j) \]
  – The problem can be solved by dynamic programming [4], given any vide flip \( X_{(0,u)} \) with length \( u \):
    \[ f(X_{(0,u)}, K) = \arg\max_{l_{\text{min}} < l < l_{\text{max}}} f(X_{(0,u-1)}, K) + f(X_{(u,u-1)}, K) \]
    \( l_{\text{max}} \) and \( l_{\text{min}} \) are the detection length of video frames.

## Performance Evaluation

<table>
<thead>
<tr>
<th>Primary Runs Results</th>
<th>Ranking</th>
<th>IBM 2013</th>
<th>Others’ Best 2013</th>
<th>CMU-IBM2012</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>ActDCR</td>
<td>MinDCR</td>
<td>ActDCR</td>
</tr>
<tr>
<td>CellToEar</td>
<td>1</td>
<td>0.9985</td>
<td>0.9978</td>
<td>1.0069</td>
</tr>
<tr>
<td>Embrace</td>
<td>1</td>
<td>0.7873</td>
<td>0.7733</td>
<td>0.8357</td>
</tr>
<tr>
<td>ObjectPut</td>
<td>2</td>
<td>1.0046</td>
<td>1.002</td>
<td>0.9981</td>
</tr>
<tr>
<td>PeopleMeet</td>
<td>2</td>
<td>1.0267</td>
<td>0.9769</td>
<td>0.9474</td>
</tr>
<tr>
<td>PeopleSplitUp</td>
<td>1</td>
<td>0.8364</td>
<td>0.8066</td>
<td>0.8947</td>
</tr>
<tr>
<td>PersonRuns</td>
<td>2</td>
<td>0.7887</td>
<td>0.7792</td>
<td>0.7708</td>
</tr>
<tr>
<td>Pointing</td>
<td>3</td>
<td>1.0045</td>
<td>0.9904</td>
<td>0.9959</td>
</tr>
</tbody>
</table>

- Compared to our last year’s results based on FV (CMU-IBM 2012):
  - this year’s system got improvement over 6/7 events (primary run).

- Compared to other teams’ results (Others’ Best 2013):
  - our system leads in 3/7 events (primary run).
Outline

• Retrospective Event Detection
  – System Overview
  – Temporal Modeling for Event Detection
  – Performance Evaluation

• Interactive Event Detection
  – Interactive Visualization
  – Risk Ranking
  – Performance Evaluation
Interactive Visualization

• Motivations:
  – How can we present events to the users more effectively?
    • E.g. two events “peoplemeet” and “pointing” may exist successively. Looking at them together are more beneficial than checking one at each time alone.

  – How can we present more informative events to the users for correction/verification?
    • E.g. correcting mis-detected events is more rewarding. for example, “embrace” → “peoplemeet” vs. “pointing” → “nonevent”.
Event-specific Detection Visualization

- ObjectPut
- Embrace
- Pointing
- PeopleMeet
- PeopleSplitUp
- CellToEar
Event-specific Detection Visualization

Embrace

Pointing

ObjectPut
Risk Ranking of Detected Events

• Approach
  – To measure the adjudication risk of each event detected by considering: 1) the margin of the top two candidates in classification; 2) temporal relations and 3) potential gain of DCR

  – Ranking events by their risk scores

  – Checking and re-labeling events from high risk to low risk.
Risk Ranking of Detected Events

– Considering our classification results: for each segmentation $S_i$, we have its top two candidates $\phi^k(S_i)$ and $\phi^{k'}(S_i)$, and their priors $p(k)$ and $p(k')$

\[ R(S_i) = \frac{1 - (\phi^k(S_i)p(k) - \phi^{k'}(S_i)p(k'))}{||S_i||} . \left\{ \begin{array}{c} w_m \\ w_f \\ w_m + w_f \end{array} \right\} \]

$w_m$ is the cost of a mis-detection and $w_f$ is the cost of a false alarm, $\sum$ is the normalizer. ($w_m = 1, w_f = 0.005$ were set based on DCR)
Risk Ranking of Detected Events

- Pair-wise events: for $S_i$ and $S_{i+1}$, we have $\varphi^k_j(S_i)\varphi^{k_{j+1}}(S_{i+1})$ and their priors $p(k_j, k_{j+1})$ and $p(k'_j, k'_{j+1})$

$$R(S_i, S_{i+1}) = \frac{1 - ((\varphi^k(S_i) + \varphi^k(S_{i+1}))p(k_j, k_{j+1}) - (\varphi^{k'}(S_i) + \varphi^{k'}(S_{i+1}))p(k'_j, k'_{j+1}))}{||S_i \cup S_{i+1}||} \begin{cases} 2 \cdot w_m \\ 2 \cdot w_f \\ 2 \cdot (w_m + w_f) \\ ... \end{cases}$$
Risk Ranking of Detected Events

more informative

less informative
## Performance Evaluation

<table>
<thead>
<tr>
<th>Actual DCR</th>
<th>Retro</th>
<th>Risk-1 (primary)</th>
<th>Risk-2</th>
<th>Risk-3</th>
</tr>
</thead>
<tbody>
<tr>
<td>CellToEar</td>
<td>0.9985</td>
<td>0.9956</td>
<td>0.994</td>
<td>1.0013</td>
</tr>
<tr>
<td>Embrace</td>
<td>0.7873</td>
<td>0.7337</td>
<td>0.6551</td>
<td>0.6705</td>
</tr>
<tr>
<td>ObjectPut</td>
<td>1.0046</td>
<td>0.9928</td>
<td>0.987</td>
<td>1.0053</td>
</tr>
<tr>
<td>PeopleMeet</td>
<td>1.0267</td>
<td>0.9584</td>
<td>0.9145</td>
<td>0.9684</td>
</tr>
<tr>
<td>PeopleSplitUp</td>
<td>0.8364</td>
<td>0.8489</td>
<td>0.8304</td>
<td>0.8924</td>
</tr>
<tr>
<td>PersonRuns</td>
<td>0.7887</td>
<td>0.7188</td>
<td>0.6865</td>
<td>0.7588</td>
</tr>
<tr>
<td>Pointing</td>
<td>1.0045</td>
<td>0.9781</td>
<td>0.974</td>
<td>0.9877</td>
</tr>
</tbody>
</table>

- **Retro**: retrospective event detection
- **Risk-1**: independent evaluation by risk ranking (25 mins for each event type)
- **Risk-2**: joint evaluation by risk ranking (a total of 175 mins)
- **Risk-3**: independent evaluation using classification scores

Risk-2 > Risk-1 > Risk-3 > Retro
Discussions

• A few thoughts
  – ground truth (automatic, crowdsourcing,...)?
  – Independent and/or dependent evaluation?
Conclusions

• Retrospective System:
  – Joint-segmentation-classification provides a promising schema for surveillance event detection.
  – Modeling temporal relations between events can boost the detection performance.

• Interactive System:
  – Event visualization with strong temporal patterns is a more efficient presentation for an interactive system.
  – Risk-based ranking demonstrates its effectiveness in relabeling events.
References:


Future Works

• **Retrospective System:**
  – Exploiting long distance temporal relations into this joint-segmentation-detection framework.
  – Exploring the performance trade-offs between localization and categorization.

• **Interactive System:**
  – Better visualization layout need to be developed, E.g. time layout.
  – Various risk ranking methods need to be tried.
  – User feedback utilization methods need to be incorporated. E.g. interactive learning.
Multiple Detections Visualization

• Objective:
  – To find visualization methods that enable multiple events representation.

• Solution:
  – Visualize the events in a graph-based layout: each node is an individual event and the edge between them representing the temporal relation.
Outline

• Retrospective Event Detection
  – System Overview
  – Temporal Modeling for Event Detection
  – Performance Evaluation

• Interactive Event Detection
  – Interactive Visualization
  – Risk Ranking
  – Performance Evaluation