IBM-Northwestern@TRECVID 2014: Surveillance Event Detection (SED)

Yu Cheng *, Jingjing Liu †, Lisa Brown †, Quanfu Fan †, Rogerio Feris †, Alok Choudhary *, Sharath Pankanti †

† IBM Research
* Northwestern University
Outline

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

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

Testing Sequence

```
Testing Sequence
[y_i]

```

Segmentation + Classification

```
Testing Sequence Segment

[y_{i+1}]
```

Detection Result

Training Sequence 1

```
Training Sequence 1
```

Sliding Window

```
Extract MoSIFT features
```

```
Fisher Vector coding
```

```
Multiclass SVM training
```

```
Model
```

```
Temporal Modeling n=2
```

```
Temporal Sequence Modeling n>2
```

```
Temporal Prior
```

```
Training
```

```
SED 2013
```

```
SED 2014
```

```
Testing
```
Sequence Temporal Modeling

• Emphasises:
  – Long distance temporal relationship Vs. Short range temporal contexts.
  – Modeling on visual words level Vs. Modeling on event level.

<table>
<thead>
<tr>
<th>Primary Runs Results</th>
<th>IBM 2014</th>
<th>IBM2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>ActDCR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CellToEar</td>
<td>0.9914</td>
<td>1.0007</td>
</tr>
<tr>
<td>Embrace</td>
<td>0.7456</td>
<td>0.8</td>
</tr>
<tr>
<td>ObjectPut</td>
<td>1.0046</td>
<td>1.004</td>
</tr>
<tr>
<td>PeopleMeet</td>
<td>0.8160</td>
<td>1.0361</td>
</tr>
<tr>
<td>PeopleSplitUp</td>
<td>0.8278</td>
<td>0.8433</td>
</tr>
<tr>
<td>PersonRuns</td>
<td>0.8111</td>
<td>0.8346</td>
</tr>
<tr>
<td>Pointing</td>
<td>1.0050</td>
<td>1.0175</td>
</tr>
</tbody>
</table>
Motivation

**Speech Recognition**

This is a hard problem to solve.

**Video Event Detection**

This -> is -> a -> hard -> problem -> to -> solve.

PeopleMeet -> Pointing -> Null -> ... -> Splitup....
Our Method – Framework

Input Video

Split into small segments with fixed length

Temporal Sequence Modeling

Joint Segmentation and Classification

Detected Events
Problem Formulation

\[ X = \{x_1, x_2, \ldots, x_n\} : \text{detections of video sequence} \]
\[ Y = \{y_1, y_2, \ldots, y_m\} : \text{event class labels of each detection} \]

Joint event classification and segmentation by maximizing

\[
 f(Y, X, Z) = \sum_{i=1}^{m} \varphi(y_i|x_i) + \mu \sum_{1 \leq k \leq i-1} \phi(z_i|z_{i-k}, \ldots, z_{i-1}) 
\]

\[ Z = \{z_1, z_2, \ldots, z_l\} : \text{visual sequence (visual words or events label)} \]

Classification: multi-class SVM

Solver: dynamic programming (M. Hoai et al, 2011)
Temporal Sequence Modeling

a) Markov Model

\[ P(x_{1:N}) = \prod_{i=1}^{N} P(x_i|x_1, \ldots x_{i-1}) = P(x_1)P(x_2|x_1)P(x_3|x_2)P(x_4|x_3) \ldots \]

b) Non-Markov Model

\[ P(x_{1:N}) = \prod_{i=1}^{N} P(x_i|x_1, \ldots x_{i-1}) = P(x_1)P(x_2|x_1)P(x_3|x_2, x_1)P(x_4|x_3, \ldots x_1) \ldots \]

Statistical counting in Markov model (i.e. \textbf{n}^{th}-order when \text{len}(u)=n)

\[ G_u(s) = \frac{N(us)}{\sum_{s' \in \Sigma} N(us')} \quad \Sigma = x_1, x_2, \ldots, x_T \]

Issues: sparsity, overfitting and scalability
Hiearchical PYP: G[u] is a PYP with a base of the PYP its parent.
Modeling on event vs. on visual words

\[ p(y_i | y_1, \cdots, y_{i-1}) \]

\[ p(z_i | z_{i-k} \cdots z_{i-1}) = \hat{p}(w_{i_1}, \ldots, w_{i_2} | w_{i_{i-k}}, \ldots, w_{i_{i-1}}) \]

\( z_i \) : the \( i \)-th segmentation

\( w_i \) : the \( i \)-th visual word in \( z_i \)

[G. Zipf. Selective studies and the principle of relative frequency in language. 1932.]
Performance Evaluation

<table>
<thead>
<tr>
<th>Primary Runs Results</th>
<th>IBM 2014</th>
<th>Others’ Best 2014</th>
<th>IBM2013</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ranking</td>
<td>ActDCR</td>
<td>ActDCR</td>
</tr>
<tr>
<td>CellToEar</td>
<td>1</td>
<td>0.9914</td>
<td>1.0032</td>
</tr>
<tr>
<td>Embrace</td>
<td>1</td>
<td>0.7456</td>
<td>0.7845</td>
</tr>
<tr>
<td>ObjectPut</td>
<td>2</td>
<td>1.0046</td>
<td>1.0023</td>
</tr>
<tr>
<td>PeopleMeet</td>
<td>1</td>
<td>0.8160</td>
<td>0.9125</td>
</tr>
<tr>
<td>PeopleSplitUp</td>
<td>2</td>
<td>0.8278</td>
<td>0.8134</td>
</tr>
<tr>
<td>PersonRuns</td>
<td>1</td>
<td>0.8111</td>
<td>0.8339</td>
</tr>
<tr>
<td>Pointing</td>
<td>2</td>
<td>1.0050</td>
<td>1.0040</td>
</tr>
</tbody>
</table>

- Compared to our last year’s system (IBM 2013):
  - this year system got improvement over 6/7 events (actual DCR of primary run).

- Compared to this year other teams’ results (Others’ Best 2014):
  - our system leads in 4/7 events (actual DCR of primary run).
Outline

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

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

• Motivation:
  – Instead of looking at a single event alone, how can we represent events with strong temporal patterns?
    • E.g. two detected events “Peoplemeet” and “pointing” may exist successively, if we look at them together, it will be effective and efficient.

  – Given thousands of events, how can we differentiate them and present more informative ones to users?
    • E.g. correct some wrong events will get more credit from DCR score, for example, “embrace” → “peoplemeet” vs. “pointing” → “nonevent”.
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.
Event-specific Detection Visualization
Visualization with Temporal Relation

People Meet

Pointing

SplitUp

People Meet

Embrace

Pointing
Risk Ranking of Detected Events

• Objective:
  – To measure the risk of detections by considering: 1) the margin of top two classification candidates; 2) temporal relation; 3) potential gain of DCR;
  – Ranking data patterns by risk scores;
  – Checking and re-annotating the detections from high risk score to low risk score.
Risk Ranking of Detected Events

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

$$R(S_i) = \frac{1 - (\varphi^k(S_i)p(k) - \varphi^{k'}(S_i)p(k'))}{||S_i||} \cdot \begin{cases} w_m \\ w_f \\ w_m + w_f \end{cases}$$

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

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

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

- More important data pattern
- Less important data pattern

- PersonRun
- Embrace
- Pointing
- PepleMeet
- PeopleSplitUp
- CellToEar
## Performance Evaluation

<table>
<thead>
<tr>
<th>Actual DCR</th>
<th>Evaluation Set (25min * 7)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Retro</td>
</tr>
<tr>
<td>CellToEar</td>
<td>0.9914</td>
</tr>
<tr>
<td>Embrace</td>
<td>0.7456</td>
</tr>
<tr>
<td>ObjectPut</td>
<td>1.0046</td>
</tr>
<tr>
<td>PeopleMeet</td>
<td>0.8160</td>
</tr>
<tr>
<td>PeopleSplitUp</td>
<td>0.8278</td>
</tr>
<tr>
<td>PersonRuns</td>
<td>0.8111</td>
</tr>
<tr>
<td>Pointing</td>
<td>1.0050</td>
</tr>
</tbody>
</table>

- **Retro**: retrospective event detection system output.
- **IBM_Inter-2014**: primary run, risk ranking over all events, and interactive experiments are performed jointly with 175min.
- **IBM-Inter-2013**: performed separately for each event with 25 mins.
- **Others’ Best 2014**: 
Conclusions

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

• Interactive System:
  – Event visualization with strong temporal pattern can benefit the efficient interactive system.
  – Risk-based ranking of detected events with temporal pattern can boost the performance.
Future Works

• **Retrospective System:**
  – Exploiting deep learning for this task.
  – 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.