IBM Research & Columbia University Multimedia Event Detection System

Speaker: Paul Natsev <natsev@us.ibm.com>
IBM T. J. Watson Research Center

On Behalf Of:
IBM Research: Matthew Hill, Gang Hua, John R. Smith, Lexing Xie
IBM Interns: Bert Huang, Michele Merler, Hua Ouyang, Mingyuan Zhou
Columbia Univ.: Shih-Fu Chang, Dan Ellis, Yu-Gang Jiang

TRECVID-2010 Workshop
Gaithersburg, MD
Nov. 15-17, 2010
Multimedia Event Detection (MED) Task Overview

- Judge Y/N for each target event given a YouTube-style video
- Challenging dataset
  - 1700+ diverse videos
  - A few shots vs long and varied
  - Only 50 examples/event

<table>
<thead>
<tr>
<th>Category</th>
<th>#Videos</th>
<th>#Keyframes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembling shelter</td>
<td>48</td>
<td>2,123</td>
</tr>
<tr>
<td>Making cake</td>
<td>48</td>
<td>3,119</td>
</tr>
<tr>
<td>Batting in run</td>
<td>50</td>
<td>347</td>
</tr>
<tr>
<td>Random</td>
<td>1,577</td>
<td>49,247</td>
</tr>
</tbody>
</table>
Key Questions

- Do cross-domain concept classifiers help for complex event detection?
  - Answer: YES! Our best performing feature…

- How do static features/models compare to dynamic ones?
  - Answer: Surprisingly similarly…

- Can we move beyond bag-of-X representations to sequence-of-X?
  - Answer: Exploratory temporal motif features show promise, 2nd best feature…
IBM MED System Architecture

Input Video

- Shot Detection
- 5 fps

- Individual Keyframes
- Frame Bags or Sequences
- Video Segments

Static Features

- GLOBAL
  - 14 global features
  - 7 granularities
- GIST
- SIFT BOW
- Semantic Model Vectors

Dynamic Features

- STIP (HOG/HOF) BOW
- Temporal Motifs
- Probabilistic Motifs (HMM)

Late Fusion

- SVM Train
- Predict and aggregate frame scores
- Video-level event models

Aggregate features

- Frame-level event models

SVM Train

RBF SVM
Static and Dynamic Features

- **Static Features:**
  - Break down video into keyframes
  - Extract 98 global image features
  - GIST features
  - Dense SIFT descriptors (BOW, 1K codebook)
  - Semantic model vectors (272 semantic concept classifiers)

- **Dynamic features**
  - Transcode videos to 5 frames per second
  - Extract Space-Time Interest Points [Laptev et al.]
  - Build dynamic visual words from HOG and HOF descriptors (BOW, 1K codebook, 1x2 temporal pyramid)
  - Temporal motifs (co-occurring sequences or bags of features)
  - Probabilistic motifs (Hierarchical HMM-based)
## Breakdown of features and event modeling approaches

<table>
<thead>
<tr>
<th>Features</th>
<th>Static features</th>
<th>Dynamic features</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Event Models</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frame-level models</td>
<td>98 Global features, GIST, SIFT BoW, Semantic Model Vector</td>
<td>--</td>
</tr>
<tr>
<td>Video-level models</td>
<td>Semantic Model Vector, SIFT BoW (Columbia*)</td>
<td>STIP HOF + Temporal Pyramid, Temporal motifs, Probabilistic motifs (HMM-based), STIP HOG + HOF (Columbia*), Audio BoW (Columbia*)</td>
</tr>
</tbody>
</table>

* For details on Columbia features/runs, see Columbia notebook paper and presentation
Single Best Performing Feature – Semantic Model Vector

- 272 semantic classifiers (cross-domain)
  - Scenes/settings
  - Objects
  - People
  - Activities

- Trained on ~600K images from independent training set

- Using ensemble SVMs on global image features

Keywords: grass, food, airplane, meeting, sport
Other Notable Features

- **Bag-of-visual words**
  - IBM: dense SIFT, 1000-D visual word codebook, soft assignment
  - Columbia: SIFT with DoG and Hessian detectors, 500-d codebooks, spatial pyramid (frame + 4 quadrants), 5000-D total feature length

- **Bag-of-audio-words**
  - Columbia: MFCCs for every 32ms, 4000-d audio word codebook

- **Spatio-Temporal Interest Points (STIP) [Laptev et al.]**
  - Histogram of Gradients (HOG) and Histogram of Flow (HOF)
  - IBM: 1000-D codebook + temporal pyramid, HOF only
  - Columbia: 4000-D codebook, concatenated HOG+HOF

- **Temporal motifs**
  - Mine sequential frequent item-sets from training data
  - Use the presence/absence of item-sets as features

- **Probabilistic motifs**
  - Learn a group of HMMs on feature partitions
  - Use the state histogram of HMMs as features
Results – Normalized Detection Cost (NDC) Per Event

- Actual NDC Assembling shelter
- Actual NDC Batting in run
- Actual NDC Making cake
## Results – Mean Average Precision Over All Events

<table>
<thead>
<tr>
<th>Run</th>
<th>Mean AP (submitted)</th>
<th>Mean AP (*with bug fix)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global</td>
<td>0.10</td>
<td>0.29*</td>
</tr>
<tr>
<td>GIST</td>
<td>0.08</td>
<td>0.23*</td>
</tr>
<tr>
<td>SIFT BoW</td>
<td>0.08</td>
<td>0.24*</td>
</tr>
<tr>
<td>Semantic Model Vector</td>
<td>0.32</td>
<td>0.32</td>
</tr>
<tr>
<td>Combo Static Features</td>
<td>0.39</td>
<td>0.44*</td>
</tr>
<tr>
<td>HoF</td>
<td>0.11</td>
<td>0.11</td>
</tr>
<tr>
<td>HoF Temporal Pyramid</td>
<td>0.13</td>
<td>0.13</td>
</tr>
<tr>
<td>Temporal Motifs</td>
<td>0.30</td>
<td>0.30</td>
</tr>
<tr>
<td>Probabilistic Motifs</td>
<td>0.26</td>
<td>0.26</td>
</tr>
<tr>
<td>Combo Dynamic Features</td>
<td>0.47</td>
<td>0.47</td>
</tr>
<tr>
<td>Combo IBM Runs</td>
<td>0.34</td>
<td>0.49*</td>
</tr>
<tr>
<td>Columbia Audio BoW</td>
<td>0.37</td>
<td>0.37</td>
</tr>
<tr>
<td>Columbia STIP BoW</td>
<td>0.45</td>
<td>0.45</td>
</tr>
<tr>
<td>Columbia SIFT BoW</td>
<td>0.47</td>
<td>0.47</td>
</tr>
<tr>
<td>Combo IBM + CU Runs</td>
<td>0.49</td>
<td>0.54*</td>
</tr>
</tbody>
</table>
Per-Event Performance Breakdown of Constituent Runs

- Assembling shelter
- Batting-in
- Making cake

Static Features
- Global
- GIST
- SIFT
- Model Vector
- HOF
- Pyramid HOF
- Prob. Motifs
- Temp. Motifs

Dynamic Features
- Combo Static
- Combo Dynamic
- Combo IBM
- Combo IBM+CU

Fusion Runs
Per-Event Observations

- Assembling shelter & making cake events
  - Not clear they are very temporal in nature
  - Static features perform on par with, or better than, dynamic features
  - Semantic model vectors outperform everything else
  - Fusion runs dramatically improve upon all constituent runs (over 2x better)

- Batting-in event
  - Most homogeneous event, highest performance of the 3 events
  - Sequence features (motifs) outperform other dynamic features
  - Fusion runs modestly improve upon all constituent runs (over 25% better)

- Fusion with Columbia runs brings an extra 10% improvement → 0.54 MAP
Summary

- **Semantic Model Vector is our single best-performing feature**
  - The cross-domain semantic concept classifiers are very useful

- **New temporal motif representation (sequence-of-X) shows promise**
  - Our second-best feature overall

- **Dynamic and static features perform comparably, surprisingly…**
  - Not all complex events are truly dynamic in nature
  - Still, fusion of dynamic and static features performs best (2x gains)

- **Columbia features/runs bring in complementary info (e.g., audio)**
  - Lead to overall MAP of 0.54 with only 50 training examples per event

- **Comments for the task**
  - If no localization required, AP and NDC give similar rankings
  - So can we use the simpler AP metric? How is cost profile motivated?
Acknowledgments: The Team (in alphabetical order)

- **IBM Research**
  - Matthew Hill
  - Gang Hua
  - Paul Natsev
  - John R. Smith
  - Lexing Xie

- **Summer Interns @ IBM**
  - Bert Huang, Columbia U.
  - Michele Merler, Columbia U.
  - Hua Ouyang, Georgia Tech
  - Mingyuan Zhou, Duke U.

- **Columbia University**
  - Shih-Fu Chang, Dan Ellis, Yu-Gang Jiang