Recognizing Complex Events in Internet Videos with Audio-Visual Features

Yu-Gang Jiang

yjiang@ee.columbia.edu

In collaboration with Xiaohong Zeng¹, Guangnan Ye¹, Subh Bhattacharya², Dan Ellis¹, Mubarak Shah², Shih-Fu Chang¹, Alexander C. Loui³

Columbia University¹ University of Central Florida² Kodak Research Labs³
We take photos/videos everyday/everywhere...

Outline

• A System for Recognizing Events in Internet Videos
  – Best performance in TRECVID 2010 Multimedia Event Detection Task
  – Features, Kernels, Context, etc.

• Internet Consumer Video Analysis
  – A Benchmark Database
  – An Evaluation of Human & Machine Performance
Outline

• A System for Recognizing Events in Internet Videos
  – Best performance in TRECVID 2010 Multimedia Event Detection Task
  – Features, Kernels, Context, etc.

• Internet Consumer Video Analysis
  – A Benchmark Database
  – An Evaluation of Human & Machine Performance
The TRECVID Multimedia Event Detection Task

- **Target:** Find videos containing an event of interest
- **Data:** unconstrained Internet videos
  - 1700+ training videos (~50 positive each event); 1700+ test videos

- **Making a cake**
- **Assembling a shelter**
- **Batting a run in**
The system: 3 major components

Feature extraction
- SIFT
- Spatial-temporal interest point
- MFCC audio feature

Classifiers
- $\chi^2$ SVM
- EMD-SVM

21 scene, action, audio concepts

Semantic Diffusion with Contextual Detectors

Best performance in TRECVID2010

*Multimedia event detection (MED) task*

Mean Minimal Normalized Cost

- Run1: Run2 + “Batter” Reranking
- Run2: Run3 + Scene/Audio/Action Context
- Run3: Run6 + EMD Temporal Matching
- Run4: Run6 + Scene/Audio/Action Context
- Run5: Run6 + Scene/Audio Context
- Run6: Baseline Classification with 3 features
Per-event performance

Batting a run in (MNC)

Assembling a shelter (MNC)

Making a cake (MNC)

Run1: Run2 + “Batter” Reranking
Run2: Run3 + Scene/Audio/Action Context
Run3: Run6 + EMD Temporal Matching
Run4: Run6 + Scene/Audio/Action Context
Run5: Run6 + Scene/Audio Context
Run6: Baseline Classification with 3 features
Roadmap > audio-visual features

Feature extraction:
- SIFT
- Spatial-temporal interest point
- MFCC audio feature

Classifiers:
- $\chi^2$ SVM
- EMD-SVM

21 scene, action, audio concepts

Semantic Diffusion with Contextual Detectors

Examples:
- Making a cake
- Assembling a shelter
- Batting a run in
Three audio-visual features...

- **SIFT (visual)**
  - D. Lowe, IJCV 04.

- **STIP (visual)**
  - I. Laptev, IJCV 05.

- **MFCC (audio)**
  - 16ms 16ms
Bag-of-\(X\) representation

- \(X = \text{SIFT} / \text{STIP} / \text{MFCC}\)
- **Soft weighting** (Jiang, Ngo and Yang, ACM CIVR 2007)
Results of audio-visual features

- Measured by Average Precision (AP)

<table>
<thead>
<tr>
<th></th>
<th>Assembling a shelter</th>
<th>Batting a run in</th>
<th>Making a cake</th>
<th>Mean AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Visual STIP</td>
<td>0.468</td>
<td>0.719</td>
<td>0.476</td>
<td>0.554</td>
</tr>
<tr>
<td>Visual SIFT</td>
<td>0.353</td>
<td>0.787</td>
<td>0.396</td>
<td>0.512</td>
</tr>
<tr>
<td>Audio MFCC</td>
<td>0.249</td>
<td>0.692</td>
<td>0.270</td>
<td>0.404</td>
</tr>
<tr>
<td>STIP+SIFT</td>
<td>0.508</td>
<td>0.796</td>
<td>0.476</td>
<td>0.593</td>
</tr>
<tr>
<td>STIP+SIFT+MFCC</td>
<td><strong>0.533</strong></td>
<td><strong>0.873</strong></td>
<td><strong>0.493</strong></td>
<td><strong>0.633</strong></td>
</tr>
</tbody>
</table>

- STIP works the best for event detection
- The 3 features are highly complementary!
Roadmap > **temporal matching**

**Feature extraction**
- SIFT
- Spatial-temporal interest point
- MFCC audio feature

**Classifiers**
- $\chi^2$ SVM
- EMD-SVM

**Semantic Diffusion with Contextual Detectors**
- 21 scene, action, audio concepts
Temporal matching with EMD kernel

- Earth Mover’s Distance (EMD)

Given two clip sets \( P = \{(p_1, w_{p_1}), \ldots, (p_m, w_{p_m})\} \) and \( Q = \{(q_1, w_{q_1}), \ldots, (q_n, w_{q_n})\} \), the EMD is computed as

\[
EMD(P, Q) = \frac{\sum_i \sum_j f_{ij} d_{ij}}{\sum_i \sum_j f_{ij}}
\]

\( d_{ij} \) is the \( \chi^2 \) visual feature distance of video clips \( p_i \) and \( q_j \). \( f_{ij} \) (weight transferred from \( p_i \) and \( q_j \)) is optimized by minimizing the overall transportation workload \( \sum_i \sum_j f_{ij} d_{ij} \).

- EMD Kernel: \( K(P, Q) = \exp^{-\rho EMD(P, Q)} \)

Temporal matching results

- EMD is helpful for two events
  - results measured by minimal normalized cost (lower is better)

```
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8
r6-baseline
r3-base+EMD
```

5% gain
Roadmap > contextual diffusion

Feature extraction:
- SIFT
- Spatial-temporal interest point
- MFCC audio feature

Classifiers:
- $\chi^2$ SVM
- EMD-SVM

Semantic Diffusion with Contextual Detectors

21 scene, action, audio concepts

Actions:
- Batting a run in
- Assembling a shelter
- Making a cake

Images:
- Making a cake
- Assembling a shelter
- Batting a run in
Event context

- Events generally occur under particular scene settings with certain audio sounds!
  - Understanding contexts may be helpful for event detection

Scene Concepts: grass, Baseball field, sky

Action Concepts: running, walking, Speech comprehensible, Cheering/Clapping

Audio Concepts:
Contextual concepts

• 21 concepts are defined and annotated over TRECVID MED development set.

<table>
<thead>
<tr>
<th>Human Action Concepts</th>
<th>Scene Concepts</th>
<th>Audio Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person walking</td>
<td>Indoor kitchen</td>
<td>Outdoor rural</td>
</tr>
<tr>
<td>Person running</td>
<td>Outdoor with grass/trees visible</td>
<td>Outdoor urban</td>
</tr>
<tr>
<td>Person squatting</td>
<td>Baseball field</td>
<td>Indoor quiet</td>
</tr>
<tr>
<td>Person standing up</td>
<td>Crowd (a group of 3+ people)</td>
<td>Indoor noisy</td>
</tr>
<tr>
<td>Person making/assembling stuffs with hands (hands visible)</td>
<td>Cakes (close-up view)</td>
<td>Original audio</td>
</tr>
<tr>
<td>Person batting baseball</td>
<td></td>
<td>Dubbed audio</td>
</tr>
</tbody>
</table>

• SVM classifier for concept detection
  – STIP for action concepts, SIFT for scene concepts, and MFCC for audio concepts
Concept detection: example results

- Baseball field
- Cakes (close-up view)
- Crowd (3+ people)
- Grass/trees
- Indoor kitchen
Contextual diffusion model

• Semantic diffusion
  [Y.-G. Jiang, J. Wang, S.F. Chang & C.W. Ngo, ICCV 2009]
  – Semantic graph
    • Nodes are concepts/events
    • Edges represent concept/event correlation
  – Graph diffusion
    • Smooth detection scores w.r.t. the correlation

Project page and source code:
Contextual diffusion results

• Context is *slightly* helpful for two events
  – results measured by minimal normalized cost (lower is better)
Outline

• A System for Recognizing Events in Internet Videos
  – Best performance in TRECVID 2010 Multimedia Event Detection Task
  – Features, Kernels, Context, etc.

• Internet Consumer Video Analysis
  – A Benchmark Database
  – An Evaluation of Human & Machine Performance

What are Consumer Videos?

• **Original unedited** videos captured by ordinary consumers
  - Interesting and very diverse contents
  - Very weakly indexed
    - On average, 3 tags per consumer video on YouTube vs. 9 tags each YouTube video has
  - Original audio tracks are preserved; good for audio-visual joint analysis

![Social Media Logos]
Columbia Consumer Video (CCV) Database

Basketball  Skiing  Dog  Wedding Reception  Non-music Performance

Baseball  Swimming  Bird  Wedding Ceremony  Parade

Soccer  Biking  Graduation  Wedding Dance  Beach

Ice Skating  Cat  Birthday Celebration  Music Performance  Playground
CCV Snapshot

• # videos: 9,317
  – (210 hrs in total)
• video genre
  – unedited consumer videos
• video source
  – YouTube.com
• average length
  – 80 seconds
• # defined categories
  – 20
• annotation method
  – Amazon Mechanical Turk

The trick of digging out consumer videos from YouTube:
Use default filename prefix of many digital cameras: “MVI and parade”.

wedding ceremony
wedding reception
biking
graduation
baseball
birthday
soccer
playground
bird
wedding dance
basketball
beach
ice skating
cat
parade
skiing
swimming
dog
non-music perf.
music perf.
### Existing Database?

<table>
<thead>
<tr>
<th>Existing Database</th>
<th>CCV Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Action Recognition</td>
<td>Unconstrained YouTube videos</td>
</tr>
<tr>
<td>– KTH &amp; Weizmann</td>
<td>Higher-level complex events</td>
</tr>
<tr>
<td>• (constrained environment) 2004-05</td>
<td></td>
</tr>
<tr>
<td>– Hollywood Database</td>
<td>More videos &amp; better defined categories</td>
</tr>
<tr>
<td>• (12 categories, movies) 2008</td>
<td></td>
</tr>
<tr>
<td>– UCF Database</td>
<td>More videos &amp; larger content variations</td>
</tr>
<tr>
<td>• (50 categories, YouTube Videos) 2010</td>
<td></td>
</tr>
<tr>
<td>Kodak Consumer Video</td>
<td>More videos &amp; categories</td>
</tr>
<tr>
<td>• (25 classes, 1300+ videos) 2007</td>
<td></td>
</tr>
<tr>
<td>LabelMe Video</td>
<td></td>
</tr>
<tr>
<td>• (many classes, 1300+ videos) 2009</td>
<td></td>
</tr>
<tr>
<td>TRECVID MED 2010</td>
<td></td>
</tr>
<tr>
<td>• (3 classes, 3400+ videos) 2010</td>
<td></td>
</tr>
</tbody>
</table>
Crowdsourcing: Amazon Mechanical Turk

- A web services API that allows developers to easily integrate human intelligence directly into their processing.

What can I do for you?

Internet-scale workforce

Task

Is this a “parade” video?
- Yes
- No

$???

financial rewards
Mark all the categories that appear in any part of the video.

Instructions:

- Watch the entire video as more categories may appear over time.
- Mark all the categories that appear in any part of the video.
- Make sure audio is on.
- If no matching category is found, mark the box in front of "None of the categories matches".
- For categories that appears to be relevant but you’re not completely sure, please still mark it.
- Please mouse-over or click on the category names to read detailed definitions.

Reliability of Labels: each video was assigned to four MTurk workers
Human Recognition Performance

• How to measure human (MTurk workers) recognition accuracy?
  – We manually and carefully labeled 896 videos
    • Golden ground truth!

• Consolidation of the 4 sets of labels

![Graph showing precision and recall for different vote counts (1-vote, 2-votes, 3-votes, 4-votes).]

Plus additional manual filtering of 6 positive sample sets: 94% final precision
Human Recognition Performance (cont.)

workers (sorted by # of submitted HITs)

<table>
<thead>
<tr>
<th></th>
<th>precision</th>
<th>recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>17</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>27</td>
<td>27</td>
<td>27</td>
</tr>
<tr>
<td>36</td>
<td>36</td>
<td>36</td>
</tr>
<tr>
<td>77</td>
<td>77</td>
<td>77</td>
</tr>
<tr>
<td>248</td>
<td>248</td>
<td>248</td>
</tr>
<tr>
<td>255</td>
<td>255</td>
<td>255</td>
</tr>
<tr>
<td>446</td>
<td>446</td>
<td>446</td>
</tr>
<tr>
<td>694</td>
<td>694</td>
<td>694</td>
</tr>
<tr>
<td>770</td>
<td>770</td>
<td>770</td>
</tr>
</tbody>
</table>
Machine Recognition System

Feature extraction
- SIFT
- Spatial-temporal interest points
- MFCC audio feature

Classifier
- $\chi^2$ kernel SVM

Average Late Fusion

Machine Recognition Accuracy

- Measured by average precision
  - SIFT works the best for event detection
  - The 3 features are highly complementary!
Human vs. Machine

- Human has much **better recall**, and is much **better for non-rigid objects**
- Machine is close to human on top-list precision
## Human vs. Machine: Result Examples

<table>
<thead>
<tr>
<th></th>
<th>true positives</th>
<th>false positives</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>found by human &amp; machine</td>
<td>found by human only</td>
</tr>
<tr>
<td>wedding dance</td>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
</tr>
<tr>
<td>soccer</td>
<td><img src="image6" alt="Image" /></td>
<td><img src="image7" alt="Image" /></td>
</tr>
<tr>
<td>cat</td>
<td><img src="image11" alt="Image" /></td>
<td><img src="image12" alt="Image" /></td>
</tr>
</tbody>
</table>
Summary

• The combination of the three audio-visual features is key for good video event recognition performance

• Temporal matching is useful for some complex events

• Current automatic event recognition methods are not that bad

• A new dataset (CCV) for consumer video analysis
Dataset download

- Unique YouTube Video IDs,
- Labels,
- Training/Test Partition,
- Three Audio/Visual Features

http://www.ee.columbia.edu/dvmm/CCV/

Fill out this ...
THANK YOU!

e-mail: yjiang@ee.columbia.edu