High-level Event Recognition in Internet Videos

Yu-Gang Jiang
School of Computer Science
Fudan University, Shanghai, China
ygj@fudan.edu.cn

Joint work with Guangnan Ye\textsuperscript{1}, Subh Bhattacharya\textsuperscript{2},
Dan Ellis\textsuperscript{1}, Mubarak Shah\textsuperscript{2}, Shih-Fu Chang\textsuperscript{1}, Alexander C. Loui\textsuperscript{3}

Columbia University\textsuperscript{1} University of Central Florida\textsuperscript{2} Kodak Research Labs\textsuperscript{3}
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.

• Columbia Consumer Video Database
  – A Benchmark and An Evaluation of Human & Machine Performance

• Speeded Up Event Recognition
  – “Real-time” event recognition
Outline

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

• Columbia Consumer Video Database
  – A Benchmark and An Evaluation of Human & Machine Performance

• Speeded Up Event Recognition
  – “Real-time” event recognition
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

Best performance in TRECVID2010

*Multimedia event detection (MED) task*

![Bar chart showing mean minimal normalized cost for different runs.]

- 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

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

21 scene, action, audio concepts

Semantic Diffusion with Contextual Detectors

Making a cake
Assembling a shelter
Batting a run in
Temporal matching with EMD kernel

• Earth Mover’s Distance (EMD)

Given two clip sets $P = \{(p_1, w_{p1}), \ldots, (p_m, w_{pm})\}$ and $Q = \{(q_1, w_{q1}), \ldots, (q_n, w_{qn})\}$, the EMD is computed as

$$\text{EMD}(P, Q) = \frac{\Sigma_i \Sigma_j f_{ij} d_{ij}}{\Sigma_i \Sigma_j f_{ij}}$$

$d_{ij}$ 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 $\Sigma_i \Sigma_j f_{ij} d_{ij}$.

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

Temporal matching results

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

![Minimal Normalized Cost](chart.png)

- 5% gain
Roadmap > contextual diffusion

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

Making a cake
Assembling a shelter
Batting a run
Event context

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

![Minimal Normalized Cost](image)

- 3% gain
Outline

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

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

• Speeded Up Event Recognition
  – “Real-time” event recognition

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
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
  – Internet: 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.”
Crowdsourcing: Amazon Mechanical Turk

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

Internet-scale workforce

What can I do for you?

Task

Is this a “parade” video?
- Yes
- No

$???

financial rewards
MTurk: Annotation Interface

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.

Sports
- Basketball
- Baseball
- Soccer
- Ice Skating
- Skating
- Swimming
- Biking

Animal
- Cat
- Dog
- Bird

Celebration
- Graduation
- Birthday
- Wedding Reception
- Wedding Ceremony
- Wedding Dance

Others
- Music Performance
- Non-music Performance
- Parade
- Beach
- Playground

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

Current Time: 10 sec

Submit

Replay | Continue Playing
Original URL: http://www.youtube.com/watch?v=-0n50a7seNl

$ 0.02
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

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

workers (sorted by # of submitted HITs)
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>true positives</th>
<th>found by human&amp;machine</th>
<th>found by human only</th>
<th>found by machine only</th>
</tr>
</thead>
<tbody>
<tr>
<td>wedding dance</td>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
</tr>
<tr>
<td>soccer</td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
<td>n/a</td>
</tr>
<tr>
<td>cat</td>
<td><img src="image6.png" alt="Image" /></td>
<td><img src="image7.png" alt="Image" /></td>
<td>n/a</td>
</tr>
</tbody>
</table>
Outline

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

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

• Speeded Up Event Recognition
  – “Real-time” event recognition

The Baseline Framework (Recall)

- 4 Factors on speed: Feature, Classifier, Fusion, Frame Sampling

- Feature efficiency is measured in seconds needed for processing an 80-second video sequence (for SIFT: 0.5fps).
- Classification time is measured by classifying a video using classifiers of all the 20 categories

Total: 1003 seconds / video
Feature Options

- (Sparse) SIFT
- STIP
- MFCC
- Dense SIFT (DIFT)
- Dense SURF (DURF)
- Self-Similarities (SSIM)
- Color Moments (CM)
- GIST
- LBP
- TINY

Suggested feature combinations:

Classifier Kernels

- Chi Square Kernel
- Histogram Intersection Kernel (HI)
- Fast HI Kernel (fastHI)

Multi-modality Fusion

- Early Fusion
- Kernel Fusion
- Late Fusion

![Graphs showing mAP for different methods: (a) MFCC, DURF, SSIM, CM, GIST, LBP and (b) MFCC, DURF.](image)
Frame Sampling


- MFCC

Sampling audio frames is always harmful.
Frame Sampling

- DURF

Uniformly sampling 16 frames per video seems sufficient.
Speeded Up Event Recognition (SUPER)

- The most favorite components:

<table>
<thead>
<tr>
<th></th>
<th>Baseline</th>
<th>SUPER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vis. frame sampling</td>
<td>–</td>
<td>Max 16</td>
</tr>
<tr>
<td>Aud. frame sampling</td>
<td>–</td>
<td></td>
</tr>
<tr>
<td>Features</td>
<td>SIFT, STIP, MFCC</td>
<td>DURF, MFCC</td>
</tr>
<tr>
<td>Classifier</td>
<td>$\chi^2$ SVM</td>
<td>fastHI SVM</td>
</tr>
<tr>
<td>Fusion</td>
<td>Late</td>
<td>Early</td>
</tr>
</tbody>
</table>

| mAP                          | 0.595    | 0.557       |
| Time*                        | 1003s    | 4.56s       |

*Classifying a video of 80 s duration, using models of 220 classes.

220-fold speed-up!
Summary

• The combination of audio-visual features is a key to 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

• Speed efficiency can be improved significantly with minor performance loss
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!

email: ygj@fudan.edu.cn