Consumer Video Understanding

A Benchmark Database +

An Evaluation of Human & Machine Performance

Yu-Gang Jiang, Guangnan Ye, Shih-Fu Chang, Daniel Ellis, Alexander C. Loui

Columbia University  Kodak Research

ACM ICMR 2011, Trento, Italy, April 2011
We (Consumers) take photos/videos everyday/everywhere...

What are Consumer Videos?

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

[YouTube] [vimeo] [Facebook] [Flickr] ...
Part I:
A Database
Columbia Consumer Video (CCV) Database
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”.

The bar chart shows the frequency of different categories:
- 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>Database</th>
<th>Description</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human Action Recognition</td>
<td>- KTH &amp; Weizmann (constrained environment)</td>
<td>2004-05</td>
</tr>
<tr>
<td></td>
<td>- Hollywood Database (12 categories, movies)</td>
<td>2008</td>
</tr>
<tr>
<td></td>
<td>- UCF Database (50 categories, YouTube Videos)</td>
<td>2010</td>
</tr>
<tr>
<td>Kodak Consumer Video</td>
<td>- (25 classes, 1300+ videos)</td>
<td>2007</td>
</tr>
<tr>
<td>LabelMe Video</td>
<td>- (many classes, 1300+ videos)</td>
<td>2009</td>
</tr>
<tr>
<td>TRECVID MED 2010</td>
<td>- (3 classes, 3400+ videos)</td>
<td>2010</td>
</tr>
</tbody>
</table>

**CCV Database**
- Unconstrained YouTube videos
- Higher-level complex events
- More videos & better defined categories
- More videos & larger content variations
- More videos & categories
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 appear 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
- Skiing
- Swimming
- Biking

Animal
- Cat
- Dog
- Bird

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

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

Current Time: 1.0 sec

Submit

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

$ 0.02
Part II:
…not Just A Database

An Evaluation of Human & Machine Performance
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.)

![Bar chart showing precision and recall for workers sorted by average labeling time per HIT. The x-axis represents workers, and the y-axis represents precision and recall values. Time is shown in seconds at the top of the bars.](image-url)
Confusion Matrices

Ground-truth Labels

Human Recognition
Machine Recognition System

Yu-Gang Jiang, Xiaohong Zeng, Guangnan Ye, Subh Bhattacharya, Dan Ellis, Mubarak Shah, Shih-Fu Chang,
Best Performance in TRECVID-2010
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
Three Audio-Visual Features...

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

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

- **MFCC (audio)**
Bag-of-X Representation

- **X = SIFT / STIP / MFCC**
- **Soft weighting** (Jiang, Ngo and Yang, ACM CIVR 2007)

Bag-of-SIFT

Keypoint extraction

SIFT feature space

DoG

Hessian Affine

Vocabulary 1

Vocabulary 2

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: Confusion Matrices

Human Recognition

Machine Recognition
### 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 (93.3% vs. 92.9%)</td>
<td><img src="image1" alt="Images" /></td>
<td><img src="image2" alt="Images" /></td>
</tr>
<tr>
<td>soccer (87.5% vs. 53.8%)</td>
<td><img src="image6" alt="Images" /></td>
<td><img src="image7" alt="Images" /></td>
</tr>
<tr>
<td>cat (93.5% vs. 46.8%)</td>
<td><img src="image11" alt="Images" /></td>
<td><img src="image12" alt="Images" /></td>
</tr>
</tbody>
</table>
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!