Multimedia Event Detection for Large Scale Video

Benjamin Elizalde
Outline

• Motivation
• TrecVID task
• Related work
• Our approach (System, TF/IDF)
• Results & Processing time
• Conclusion & Future work
• Agenda
Motivation

• YouTube alone claims 72 hours uploaded per minute, with 3 billion viewers a day.

• Video need to be searched, sorted, retrieved based on content descriptions that are “higher-level”.

• High demand in industry, even higher demand in intelligence community.
Hrishikesh Aradhye, "Finding cats playing pianos: Discovering the next viral hit on YouTube".
TrecVID Multimedia Event Detection

• 17 teams.

• SRI AURORA
  – SRI International (SRI) (Sarnoff)
  – International Computer Science Institute, University of California, Berkeley (ICSI)
  – Cycorp
  – University of Central Florida (UCFL)
  – UMass Amherst

• ICSI contribution: **Acoustic**, visual, and multimodal methods for Video Event Detection.
Task: Multimedia Event Detection

• Given:
  • An event kit which consists of an event name, definition, explication, video example.

• Wanted:
  • A system that can search multimedia recordings for user-defined events.
What is Video Event Detection?

• An event:
  – is a complex activity occurring at a specific place and time;
  • involves people interacting with other people and/or objects;
  • consists of a number of human actions, processes, and activities that are loosely or tightly organized and that have significant temporal and semantic relationships to the overarching activity;
  • is directly observable.
Sample Video 1: “Board Tricks”
Sample Video 2: “Board Tricks”
Test Video: “Board Tricks”
Related Work: TrecVid MED 2010

- Making a cake
- Batting a run in
- Assembling a shelter
## Related Work: TrecVid MED 2010

<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</td>
<td>Cakes (close-up view)</td>
<td>Original audio</td>
</tr>
<tr>
<td>with hands (hands visible)</td>
<td></td>
<td>Dubbed audio</td>
</tr>
<tr>
<td>Person batting baseball</td>
<td></td>
<td>Speech comprehensible</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Music</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cheering</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Clapping</td>
</tr>
</tbody>
</table>

General Observations

<table>
<thead>
<tr>
<th>Year</th>
<th>Number of classes</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>3</td>
</tr>
<tr>
<td>2011</td>
<td>15</td>
</tr>
<tr>
<td>2012</td>
<td>30</td>
</tr>
</tbody>
</table>

- E010  Grooming an animal
- E011  Making a sandwich
- E014  Repairing an appliance
- E022  Cleaning an appliance
- E025  Marriage proposal
- E029  Winning a race without a vehicle
General Observations

• Single-model approach is problematic:
  • Too noisy

• Classifier ensembles problematic:
  • Which classifiers to build?
  • Training data?
  • Annotation?
  • Idea doesn’t scale!
General Observations Audio

• Speech Recognition (incl. keyword spotting) mostly infeasible (50% of videos contain no speech, speech is arbitrary languages, quality varies.

• Acoustic event detection has same issues as visual object recognition.

• Music indicative of events but not with high confidence.
Theoretical Framework

• **Concepts** without **percepts** are empty; **percepts** without **concepts** are blind. (Kant)

• **Percepts** are an impression of an object obtained by use of the senses.

• Concepts = Events
• Percepts = Observations

Elizalde Benjamin, Gerald Friedland et al. There is No Data Like Less Data: Percepts for Video Concept Detection on Consumer-Produced Media. ACM Multimedia 2012 Workshop.
Contribution

• Extract audio percepts.

• Determine which percepts are uncommon across concepts but common to the same concept.

• Detect video concepts by detecting common percepts.
Conceptual System Overview

Framework:

Multimedia Document → Percepts Extraction → Percepts Selection → Classification

Concept (train) → Concept (test)

Realization:

Audio Track → Diarization & K-Means → TFIDF → SVM

Concept (train) → Concept (test)
Diarization or Percept Extraction

• Based on ICSI Speaker Diarization System...
  – Who spoke when?

• ...but:
  – Speech/Speaker specific components removed.
  – Tuned for generic event diarization.

Diarization or Percept Extraction
Diarization or Percept Extraction

• Input:
  – Features: MFCC (19) + D + DD.
  – HTK format.
  – Window size 25ms every 10ms.
  – High number of initial segments.
  – Assumed 54 clusters/sounds.
  – Gaussian Mixture Models.
Diarization or Percept Extraction

• Output:
  – Segmentation file.
  – A GMM file with one GMM per segment.

SPEAKER [label] [channel] [begin time] [length] < NA > < NA > [speaker ID] < NA > [word-based transcription]

• Weight
• Variance
• Mean
Conceptual System Overview

Framework:
Multimedia Document → Percepts Extraction → Percepts Selection → Classification

Realization:
Audio Track → Diarization & K-Means → TFIDF → SVM

Concept (train) → Concept (test)
Percepts Dictionary

- Percepts extraction works on video-by-video basis.
- Use clustering to unify percepts across videos in one concept and build prototype percepts.
The GMMs’ weight, mean and variance are combined to create simplified super vectors.

- Apply K-means to cluster the super vectors.

- Represent videos by (K) super vectors of prototype percepts = “words.”

- 300 clusters
- Euclidean distance
- 10 iterations
Distribution of “Words”

Histogram of top-300 “words”.

Near Zipfian Distribution!
Properties of “Words”

• Sometimes same “word” describes more percepts (homonym).

• Sometimes same percept is described by the different “words” (synonym).

• => Problem?
  – (ant, aunt)
  – (smart, bright)
Conceptual System Overview

Framework:

Multimedia Document → Percepts Extraction → Percepts Selection → Classification

Realization:

Audio Track → Diarization & K-Means → TFIDF → SVM

Concept (train) → Concept (test)

Concept (train) → Concept (test)
Term Frequency / Inverse Document Frequency

- Reflects how important a word is to a document in a collection.
- Will the most common words in English help us finding a specific document?
  - Search for “the good fat grey kitty”
- Variations are often used by search engines for scoring and ranking documents.
Term Frequency / Inverse Document Frequency

- **TF(c_i, D_k)** is the frequency of “word” c_i in concept D_k
  - \( P(c_i = c_j | c_j \in D_k) \) is the probability that “word” c_i equals c_j in concept D_k

- **IDF(c_i)** and tells you whether a word is common or rare across the documents
  - \(|D|\) is the total number of concepts
  - \( P(c_i \in D_k) \) is the probability of “word” c_i in concept D_k

\[
TF(c_i, D_k) = \frac{\sum_j n_j P(c_i = c_j | c_j \in D_k)}{\sum_j}
\]

\[
IDF(c_i) = \log \frac{|D|}{\sum_k P(c_i \in D_k)}
\]
Conceptual System Overview

**Framework:**

- Multimedia Document
  - Percepts Extraction
  - Percepts Selection
  - Classification
    - Concept (train)
    - Concept (test)

**Realization:**

- Audio Track
  - Diarization & K-Means
  - TFIDF
  - SVM
    - Concept (train)
    - Concept (test)
Support Vector Machine Classifier

• Input
  – A histogram per clip with TF/IDF weighted values.
  – Multiclass SVM
  – Intersection Kernel

• Output
  – Score for each of the possible classes.
Audio-Only Detection on MED-DEV11

Error at FA=6%: Miss = 58%

Surpassed Year-1 ALADDIN goal audio-only! (goal: 75% miss at 6% FA)
Database includes 150k files with 3 min duration average.

Diarization Time: \(150,000 \times \frac{3}{60} \div 24 \div 50 = 7 \text{ days}\)
Conclusions

• 150k videos = no more looking at dataset...

• Teach computers to think, but not necessarily like a human.

• Event/Concept detection is still blooming.
What would you do?
Future Work

• Many knobs to tune.
• Reduce ambiguities in percepts extraction.
• Exploit temporal dimension better: (“sentences”, “paragraphs”?)
• Diarization using CUDA parallelization.
Thank You! Questions?

• Email: benmael@icsi.berkeley.edu

• Work together with: Gerald Friedland, Robert Mertens, Luke Gottlieb, and others.