**Driving Scientific Discovery with Social Media Images and Videos**

Groovin’. Snowball dances to the beat. 
Irena Schulz, Bird Lovers Only Rescue Inc

Dr. Gerald Friedland
fractor@icsi.berkeley.edu

---

**Consumer-Produced Videos are Growing in the Internet**

- YouTube claims 65k-100k video uploads per day or 48-72 hours every minute
- Youku (Chinese YouTube) claims 80k video uploads per day
- Facebook claims 415k video uploads per day!

---

**Why do we care?**

Consumer-Produced Multimedia allows empirical studies at never-before seen scale.

Spontaneous motor entrainment to music in multiple vocal mimicking species
A Schachner, TF Brady, IM Pepperberg, MD Hauser - Current Biology, 2009
Challenges I

User-provided tags are:
- sparse
- any language
- imply random context

Solution: Use the actual audio and video content for search.

Challenges II

Research to search the actual audio and video information is hindered by:
- YouTube videos not legally downloadable
- No reliable annotation
- Search in YouTube doesn’t work (see Challenges I…)

The Multimedia Commons

100M videos and images, and a growing pool of tools for research with easy access through Cloud Computing

Collaboration Between Academia and Industry:

Benchmarks & Grand Challenges:

Supported in part by NSF Grant 1251276
"BIGDATA: Small: DCM: DA: Collaborative Research: SMASH:
Scalable Multimedia content AnalysiS in a High-level language"
Work on Multimedia Content Retrieval

- Computer Vision: Focus on solving the AI problem, e.g. through object labeling
- Video Retrieval:
  - Computer Vision techniques
  - Motion
  - Audio
  - Metadata

Our Approaches to Content-based Video Search

- Focus on events (time and location)
- Combine text and image/video similarity searches and event search
- Try to ‘translate’ multimedia data into text
Events: Multimodal Location Estimation

http://mmle.icse.berkeley.edu

Intuition for the Approach

Node: Geolocation of video

Edge: Correlated locations (e.g. common tag, visual, acoustic feature)

$p(x_i | \{t^k_i\})$

Edge Potential: Strength of an edge, (e.g. posterior distribution of locations given common tags)

$p(x_j | \{t^k_j\})$

$p(x_i, x_j | \{t^k_i\} \cap \{t^k_j\})$


An Experiment

Listen!

• Which city was this recorded in?

• Solution: Tokyo, highest confidence score!
Evento360: Search with Combined Textual, Visual, and Acoustic Features

‘Translate Multimedia’: Scenario
Empirical Study: How do Children learn to catch a ball?

Example Video
https://www.youtube.com/watch?v=o6QXcP3Xvus

Properties of Consumer-Produced Videos of Multimedia Commons
- Visuels: No constraints in angle, number of cameras, cutting, editing
- Audio: 70% heavy noise, 50% speech, any language, 40% dubbed, 3% professional content
- Metadata: geotags correlated with technology adaptation, tags in high part of Zipf distribution
Analyzing the Audio Track

Cameron learns to catch (http://www.youtube.com/watch?v=o6QXcP3Xvus)

Approach

- Extract “audible units” aka percepts.
- Determine which percepts are common across a set of videos we are looking for but uncommon to others.
- Similar to text document search.

Conceptual System Overview

Percepts Extraction

- High number of initial segments
- Features: MFCC19+D+DD+MSG
- Minimum segment length: 30ms
- Train Model(A,B) from Segments A,B belonging to Model(A) and Model(B) and compare using BIC:

  \[ \log p(X|\Theta) - \frac{1}{2} \lambda K \log N \]

- Derived from Speaker Diarization
Percepts Dictionary

- Percepts extraction works on a per-video basis
- Use k-means to unify percepts across videos in the same set and build "prototype percepts"
- Represent video sets by supervectors of prototype percepts = "words"

Questions...

- How many unique "words" define a particular concept?
- What’s the occurrence frequency of the "words" per set of video?
- What’s the cross-class ambiguity of the "words"?
- How indicative are the highest frequent "words" of a set of videos?

Properties of “Words”

- Sometimes same “word” describes more percepts (homonym)
- Sometimes same percepts are described by the different “words” (synonym)
- Sometimes multiply “words” needed to describe one percepts
  => Problem?

Distribution of “Words”

Long-Tailed Distribution (~ Zipf)
**Zipf?**

**TF/IDF**

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

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

- \(TF(c_i, D_k)\) 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 \(D_k\).
- \(|D_k|\) is the total number of concepts.
- \(P(c_i \in D_k)\) is the probability of "word" \(c_i\) in concept \(D_k\).

**Classify the Words**

- Have: New input video and set of representative videos
- Need: Does this belong to the same set
- Classifier takes 300 tuples of ("words", TF-IDF values) as input
- Use SVM with Intersection Kernel (IKSVM) / Deep Learning

**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)
Audio-Only Detection in TRECVID MED 2011

Visualization of Zipfian Percepts

- Top-1 percepts very representative of concept.

Thank You! Questions?