Audio and (Multimedia) Methods for Video Analysis

Dr. Gerald Friedland,
fractor@icsi.berkeley.edu
Current Situation

Multimedia Computing

Computer Vision

Speech Processing

Computer Listening

Music Processing

CASA

Area being worked on

Area not being worked on
What is this Seminar About?

This Seminar

- Computer Vision
- Multimedia Computing
- Speech Processing
- Computer Listening
- Music Processing
- CASA

Area being worked on

Area not being worked on
What is this Seminar about?

- Introduction to audio processing from a CS point of view (application driven)
- Audio content (including speech) analysis with a focus on videos
- Multimodal Integration of visual and acoustic data
Why you should you be interested in the Topic

- Acoustic processing of (large scale, consumer produced) videos (outside speech and music domain) is a brand-new emerging field driven by:
  - Massive government funding (e.g. IARPA Aladdin, IARPA Finder, DARPA BOLT)
  - Industry’s needs to make consumer videos retrievable and searchable
  - Interesting academic questions
Btw....

- ISCA has founded a new Special Interest Group on Audio and Speech Retrieval
- IEEE CS just founded a new Special Interest Group on Acoustic and Multimedia Processing for Multimedia (SIGASAP)
• What are the best methods for generic audio analysis (rather than specialized music or speech processing)?
• How does the performance compare to Computer Vision?
• When is the vision better, when is listening better, when is the combination better?
• How can we benefit from the visual modality in the acoustic domain and vice versa?
• Does audio contain more content than video (per byte)?
• Are we targeting a human baseline? (Can a computer detect a dog whistle?)
• What is the computer actually detecting?
• What are the tasks that acoustic processing can solve?
• What interesting applications come out of that?
Some Examples of Research at ICSI

- Video Navigation
- Video Concept Detection
- Forensic User Matching
- Copyright Detection
- Location Estimation
- Fast Approaches in Parlab
Video Navigation

Navigation Capabilities: 1985

- Record
- Store
- Play
- Random Seek
- Annotate Manually
Video Navigation

Navigation Capabilities: 2010

- Record
- Store
- Stream
- Play
- Random Seek
- Annotate Manually
Motivation

Yahoo! Video Segmentation Task of ACM Multimedia Grand Challenge 2009:

“Create navigation based on narrative elements”
Joke-O-Mat: Browsing Sitcoms Punchline by Punchline

Gerald Friedland, Luke Gottlieb, and Adam Janin

[friedland|gottlieb|janin]@icsi.berkeley.edu
Video Concept Detection

- TrecVid MED 2011
- Goal: Find videos with certain concepts “feeding an animal”, “changing a tire” from a large collection of unconstrained videos
- Data: 5k test videos, 15 concepts in 496 training recordings
"Videos in the wild" aka "Found data"

Fig. 2. Several frames from the MediaEval 2010 test set as described in Section 3.3.

In order to get an impression of the dataset, we manually watched randomly chosen videos from the training set. Only a fraction of them were recorded in a controlled environment such as inside a studio at a radio station. The other were home video style with ambient noises and unstable camera settings. About a fraction of the videos were edited to contain changed scenes or fast or slow replay. The relatively short lengths of each video should be noted as the maximum length of Flickr videos is limited to a few seconds. Moreover, about a fraction of videos in our data set have less than a few seconds playtime. Figure 2 shows several sample frames from the MediaEval test set.

However, metadata provided by the user often provides direct and sensible clues for the task. A fraction of videos in the training set were annotated by their uploaders with at least a titled tags or description, often including location information. For a human, it is a fairly straightforward task to determine from the metadata which keyword or keywords combination indicates the smallest and most accurate geographical entity. However, for a machine, extracting a list of toponym candidate keywords and further choosing a correct single keyword or combination of keywords is a challenging task. Misspelled or compound words concatenated without spaces are commonly found in user-annotated metadata and these add more difficulty to the task. For example, "my trip to fishermanswharf san francisco" should resolve to the "Fisherman's Wharf" in "San Francisco."
Approach

- Segment the video using “noise diarization”
- Cluster the models of the segments into the 300 most important “words”
- Use the “words” to describe the concepts
- Use a TF/IDF approach for retrieval! (back to TREC!)
Distribution of Words in a Concept

Almost Zipfian Distribution
Results

Error at FA=6%: Miss = 58%
Location Estimation

• Question: Can audio help find the recording location (GPS coordinates) of a random Flickr video?

• Answer: Yes, but we don’t really know why!
Location Estimation

- Goal: Match city locations of a set of videos to within 50km radius
- Data: 1080 training videos, 36 cities (across globe)
- Test set: 285 matching trials, 4845 non-matching trials
- Anecdotal: Human baseline = random!
An Experiment

Listen!

• Which city was this recorded in?

**Pick one of:** Amsterdam, Bangkok, Barcelona, Beijing, Berlin, Cairo, Cape Town, Chicago, Dallas, Denver, Duesseldorf, Fukuoka, Houston, London, Los Angeles, Lower Hutt, Melbourne, Moscow, New Delhi, New York, Orlando, Paris, Phoenix, Prague, Puerto Rico, Rio de Janeiro, Rome, San Francisco, Seattle, Seoul, Siem Reap, Sydney, Taipei, Tel Aviv, Tokyo, Washington DC, Zuerich

• **Solution:** Tokyo, 4th highest confidence score!
CityID Approach

**Training**

City-independent audio → Extract MFCC features → Train city-independent GMM model → City-independent GMM model

City-dependent audio → Extract MFCC features → Train city-dependent GMM model → City-dependent GMM model

**Testing**

Audio from unknown city → Extract MFCC features

City-dependent GMM model → Compute likelihood ratio → Score

City-independent GMM model → Score
Results

DET curve for common-user city identification with 540 videos and 9,700 trials

EER = 22%
Why does this work?

• Possible explanation for non–random results:
  • Sounds that are particular to each city
  • Bird noises? Traffic noise?
  • Community effects (e.g. camera models)
  • Your idea...
Conclusion

• Non-Speech/Non-Music processing shows (surprising) results

• Useful for many tasks, especially in unconstrained situations (social media videos, robot)

• Still vastly unexplored territory (current audio research mainly based on artificial corpora)
Conclusion

• Non-Speech/Non-Music processing can potentially improve:
  • context for visual algorithms
  • video retrieval, navigation, indexing
  • robotics
  • localization
  • forensics

• Audio processing is usually more efficient than visual processing (less data!)
How is the Seminar Organized?

- 4–5 lectures fundamental introduction (from me)
- Guest lectures
- Presentations from students (important and current research papers)
Lecture Material


• (Constantly changing) draft available http://www.mm-creole.org
How do you receive Credit?

• 2 Credits (graded or ungraded) based on:
  – presentation of research paper(s)
    or
  – a project
List of Papers

See website
Next Week

Remote Lecture (Skype) from PISA (MediaEval 2011 Workshop)