The Multimedia Commons: Driving Scientific Discovery with Social Media Images and Videos

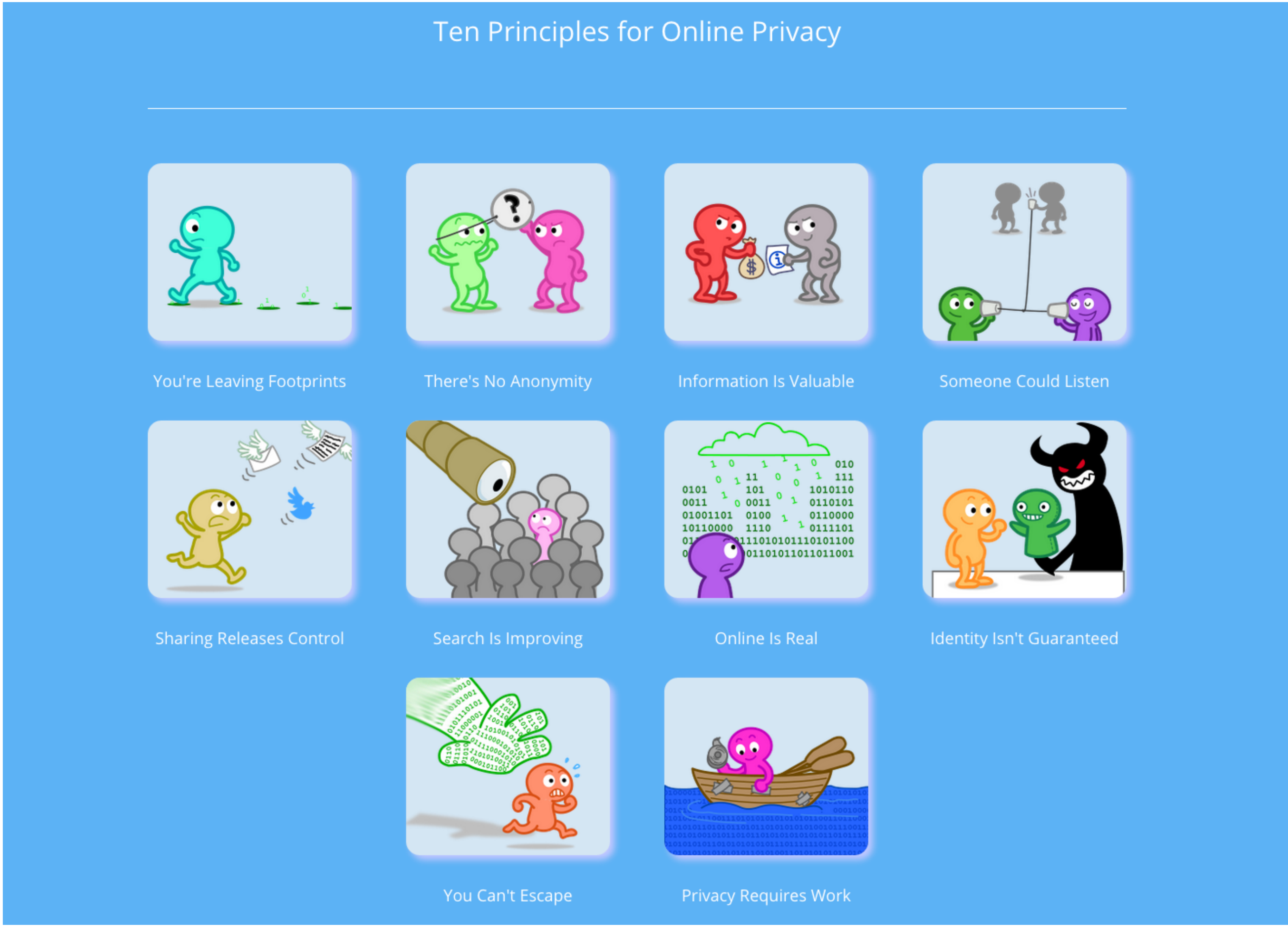
Groovin'. Snowball dances to the beat. Irena Schulz, Bird Lovers Only Rescue Inc

Dr. Gerald Friedland

Director Audio and Multimedia Lab

International Computer Science Institute

fractor@icsi.berkeley.edu



http://teachingprivacy.org

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

Google "giving directions to a location"

Web Images Maps **Videos** News More Search tools

3 results (0.13 seconds)

Ad related to **"giving directions to a location"**

Maps & Driving Directions
driving-directions.easymaps.co/
Enter Address Or Location. Free Maps & Directions w/Toolbar!

Key considerations for all maps from the Course Creating a Map with Illu...

www.lynda.com › ... › Creating a Map with Illustrator
Are you **giving directions to a location**, or general information about the area. How much of the area should ...

Community Helpers - SlideServe

www.slideserve.com/kagami/community-helpers
Aug 3, 2012
This will help with **giving directions to a location**. Materials: Our maps A step by step direction route on chart ...

PPT – A Study on Wearable Computing PowerPoint presentation | free to download - PowerShow.com

www.powershow.com/.../A_Study_on_Wearable_Computing_powerpoi...
Technology which allows for the human and ... The concept of wearable computers attempts to bridge the 'interaction gap' ... Sprout. Spot. 17 /18. Conclusion .

Stay up to date on these results:

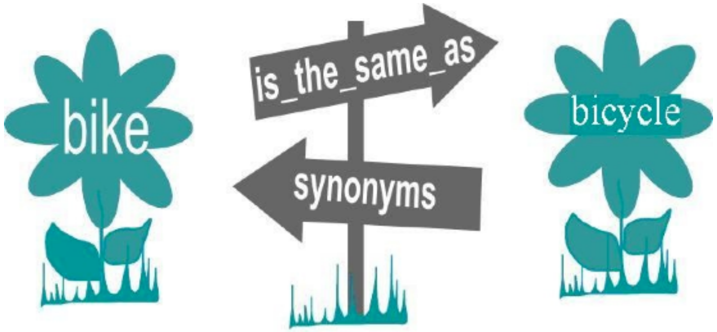
- Create an email alert for **"giving directions to a location"**

Challenges I

User-provided tags are:

- sparse
- any language
- imply random context

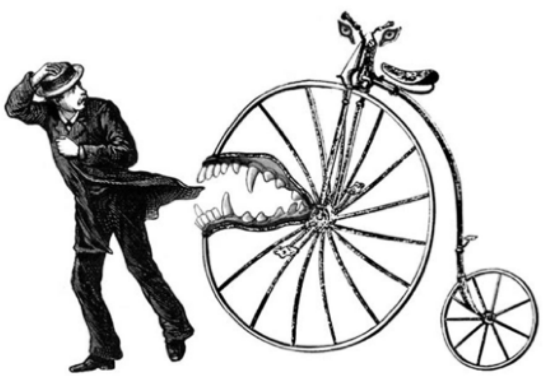
Solution: Use the actual audio and video content for search.



Challenges II

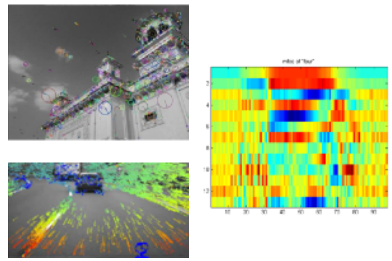
Research to search the actual audio and video information is hindered by:

- YouTube videos not legally downloadable
- No reliable annotation
- Search in YouTube doesn't work (see Challenges I…)


A VICIOUS CYCLE

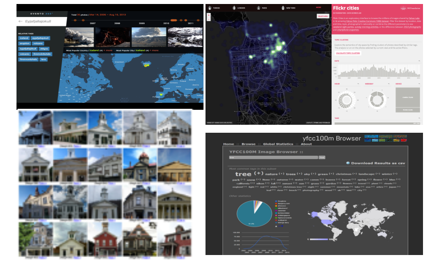
The Multimedia Commons

100.2M Photos
800K Videos



Features for Machine Learning
(Visual, Audio, Motion, etc.)

User-Supplied Metadata
and New Annotations



Tools for Searching,
Processing, and Visualizing

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

Collaboration Between Academia and Industry:

ICSI INTERNATIONAL COMPUTER SCIENCE INSTITUTE

LAWRENCE LIVERMORE NATIONAL LABORATORY

YAHOO! LABS

amazon web services

Berkeley UNIVERSITY OF CALIFORNIA

Benchmarks & Grand Challenges:

DIGITAL VIDEO RETRIEVAL at NIST

ACM mm

2016 ICMR

MediaEval Benchmark

Creative Commons or
Public Domain

NSF

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

YFCC100M

(Yahoo Flickr Creative Commons 100M)

100.2M Photos
800K Videos

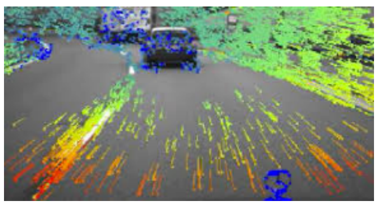
User-Supplied Metadata

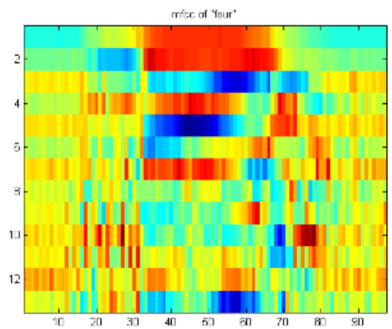
YAHOO! LABS

B. Thomee, D. A. Shamma, B. Elizalde, G. Friedland, K. Ni, D. Poland, D. Borth, L. Li: *The New Data in Multimedia Research*, Communications of the ACM, Feb 2016.

YLI Corpus

(Yahoo Livermore ICSI Corpus)





DIGITAL VIDEO RETRIEVAL at NIST

ACM mm

MediaEval Benchmark

Features for Machine Learning (Visual, Audio, Motion, etc.)

Ground Truth Annotations

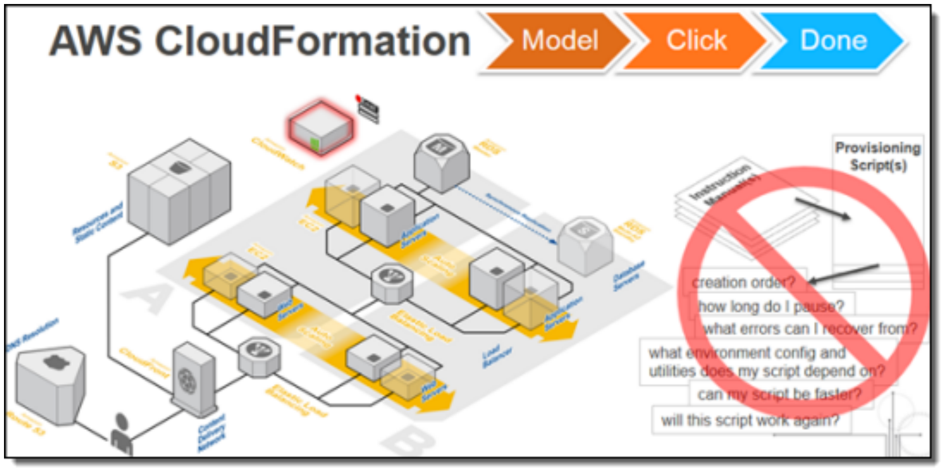
LAWRENCE LIVERMORE NATIONAL LABORATORY

ICSI INTERNATIONAL COMPUTER SCIENCE INSTITUTE

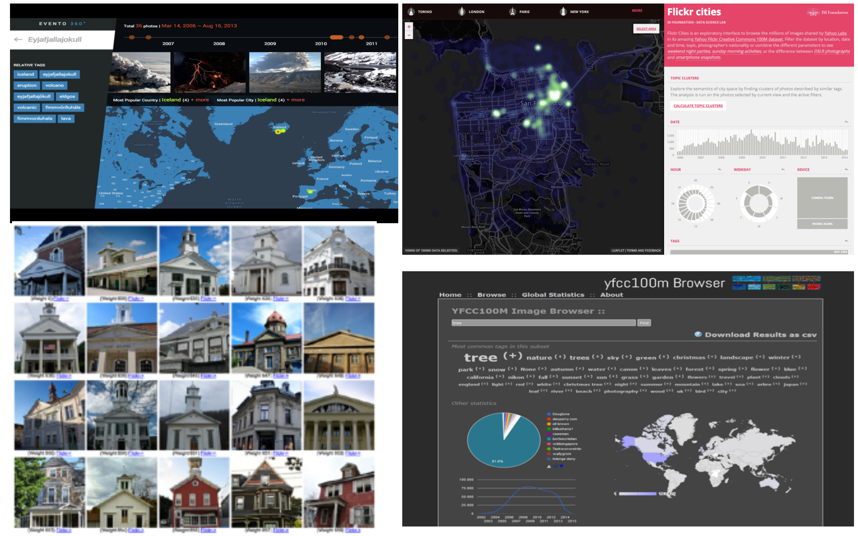
Julia Bernd, Damian Borth, Benjamin Elizalde, Gerald Friedland, Heather Gallagher, Luke Gottlieb, Adam Janin, Sara Karabashlieva, Jocelyn Takahashi, Jennifer Won: *The YLI-MED Corpus: Characteristics, Procedures, and Plans*, ICSI Technical Report TR-15-001, arXiv:1503.04250.

SMASH

(Scalable Multimedia content AnalysiS in a High-level Language)



Storage and
Cloud Formation Template



Tools for Searching,
Processing, and Visualizing

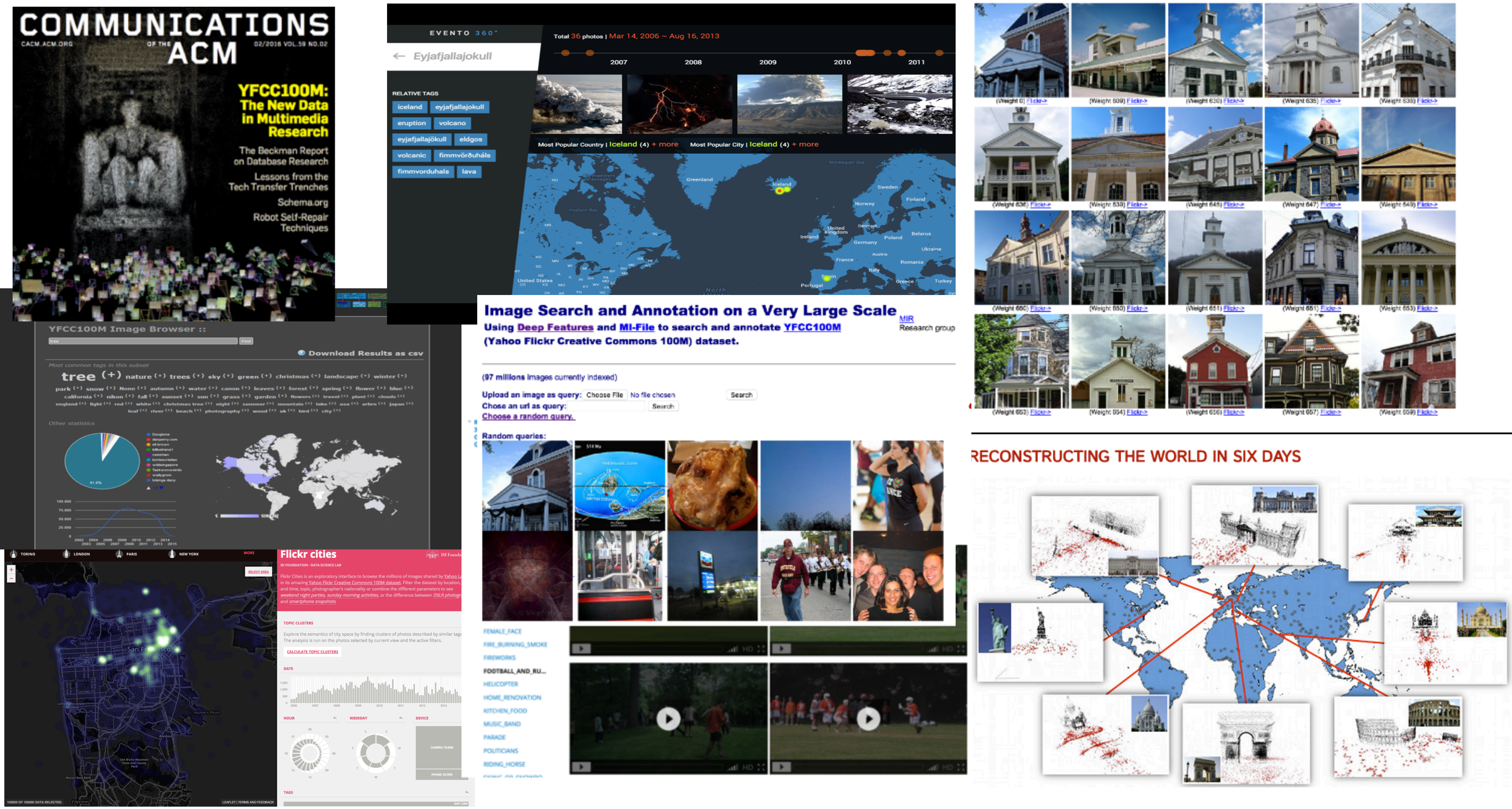
amazon web services™

INTERNATIONAL COMPUTER SCIENCE INSTITUTE

Berkeley
UNIVERSITY OF CALIFORNIA

http://multimedia.icsi.berkeley.edu/scalable-big-data-analysis/smash/

The Multimedia Commons: An Open Infrastructure for Large-Scale Multimedia Research



http://mmcommons.org

Challenges II

Challenges I

Work on Multimedia Content Retrieval

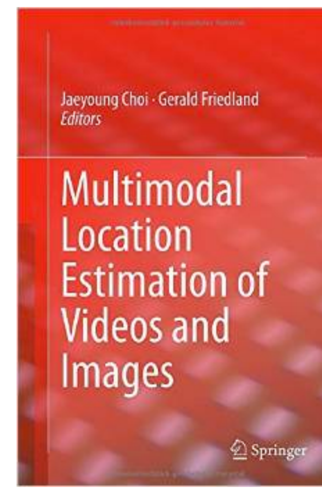
- Computer Vision: Focus on solving the AI problem, e.g. through object labeling

- 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

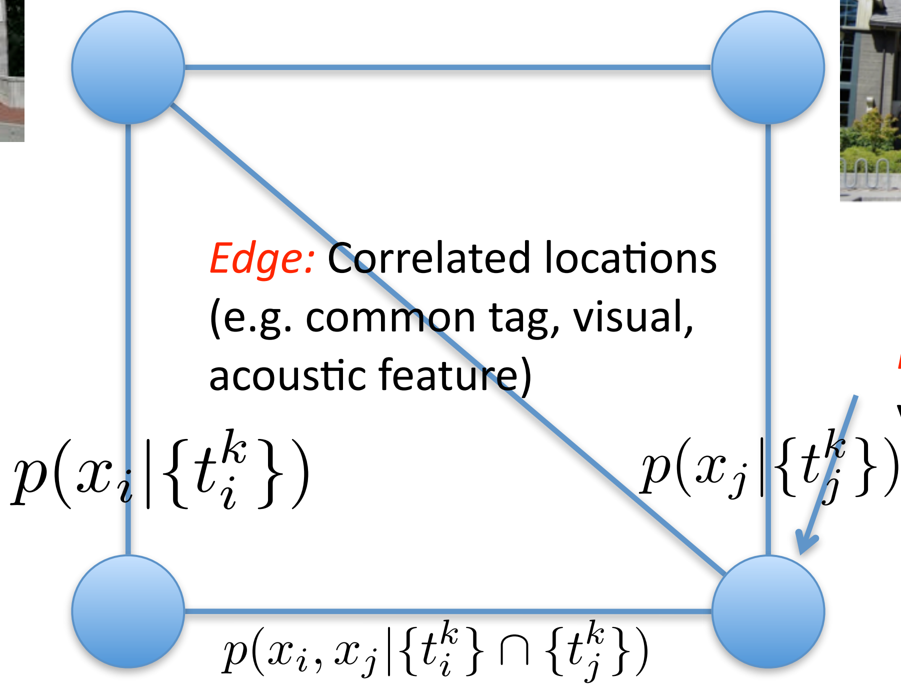


http://mmle.icsi.berkeley.edu

Intuition for the Approach

{berkeley, sathergate, campanile}

{berkeley, haas}

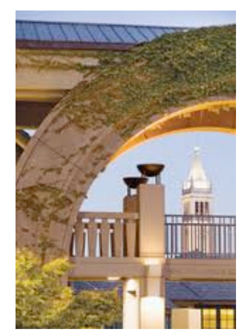


Edge: Correlated locations (e.g. common tag, visual, acoustic feature)

Node: Geolocation of video

{campanile}

{campanile, haas}

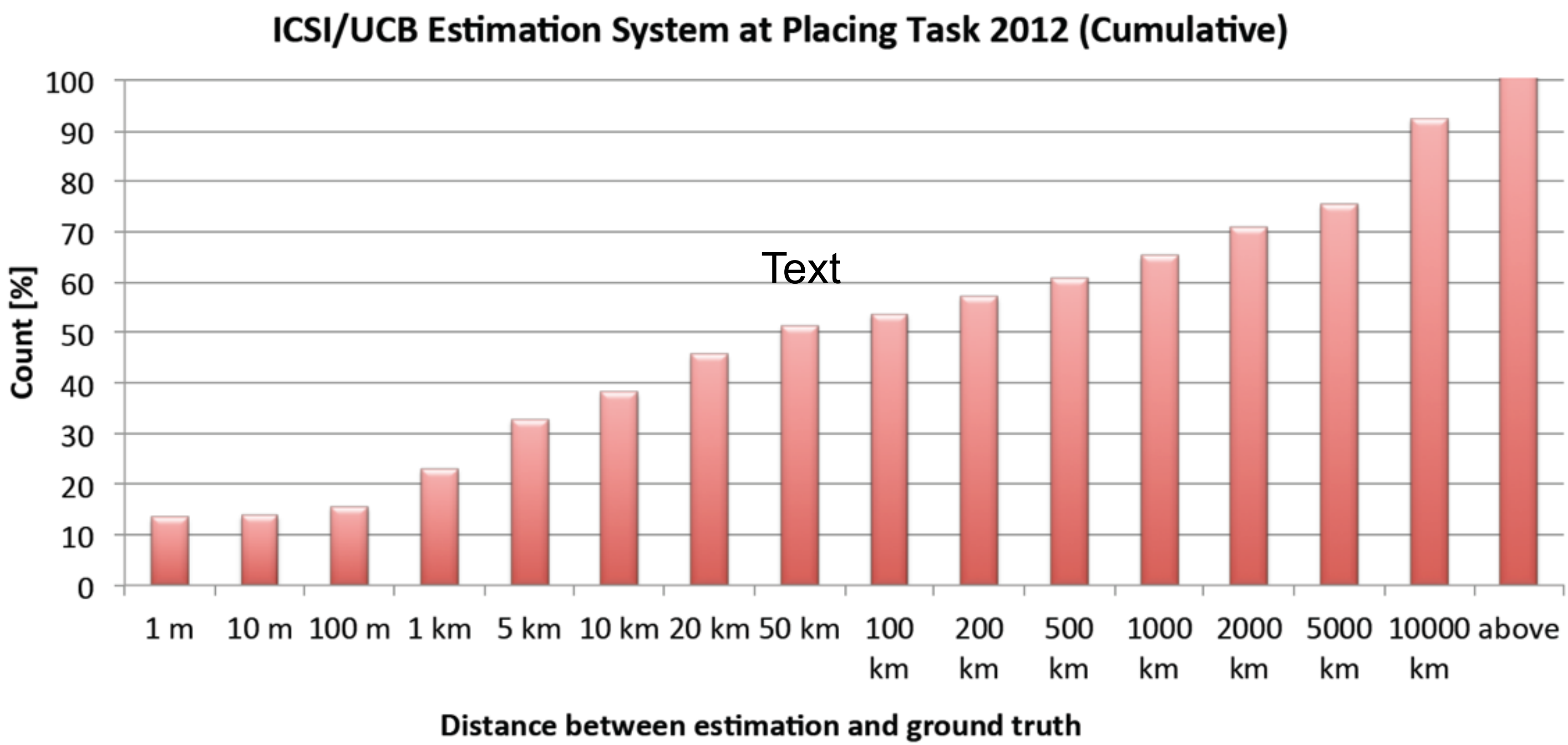


Edge Potential: Strength of an edge, (e.g. posterior distribution of locations given common tags)

MediaEval Benchmark

MediaEval Benchmarking Initiative for Multimedia Evaluation

The "multi" in multimedia: speech, audio, visual content, tags, users, context



J. Choi, G. Friedland, V. Ekambaram, K. Ramchandran: "Multimodal Location Estimation of Consumer Media: Dealing with Sparse Training Data," in Proceedings of IEEE ICME 2012, Melbourne, Australia, July 2012.

An Experiment

Listen!

- Which city was this recorded in?
  Pick one of: Amsterdam, Bangkok, Barcelona, Beijing, Berlin, Cairo, CapeTown, 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, highest confidence score!

Evento360: Search with Combined Textual, Visual, and Acoustic Features

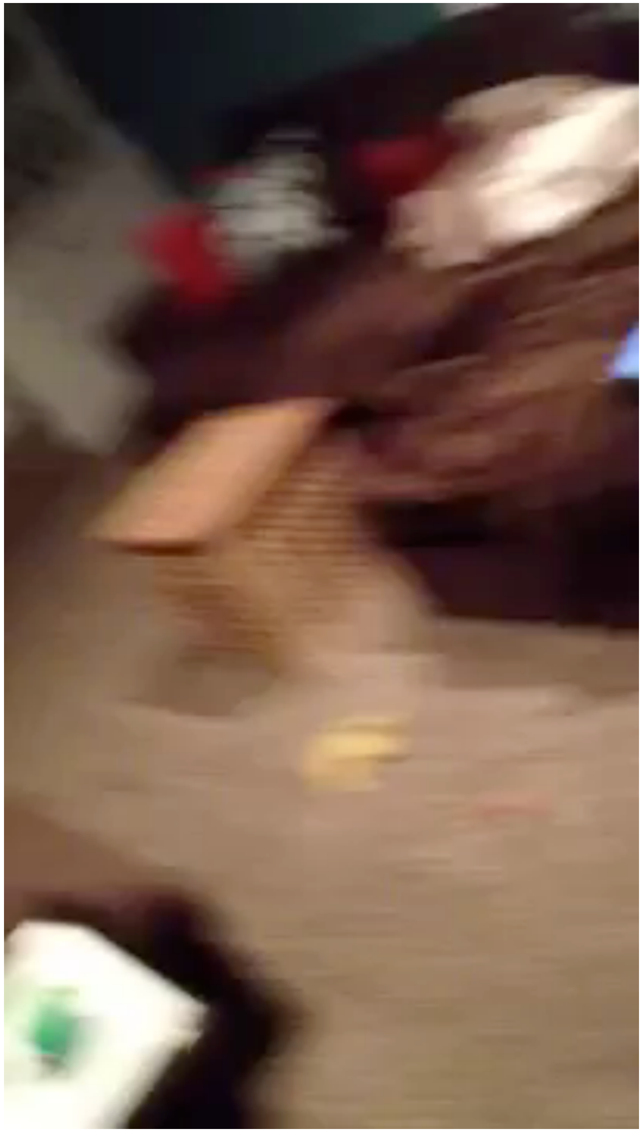
‘Translate Multimedia’: Scenario

Empirical Study: How do Children learn to catch a ball?

Properties of Consumer-Produced Videos of Multimedia Commons

- Visuals: 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

Example Video



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

Restricting Myself to Audio Content (for now)

- Where I have experience
- Lower dimensionality
- Underexplored Area
- Useful data source for other audio tasks

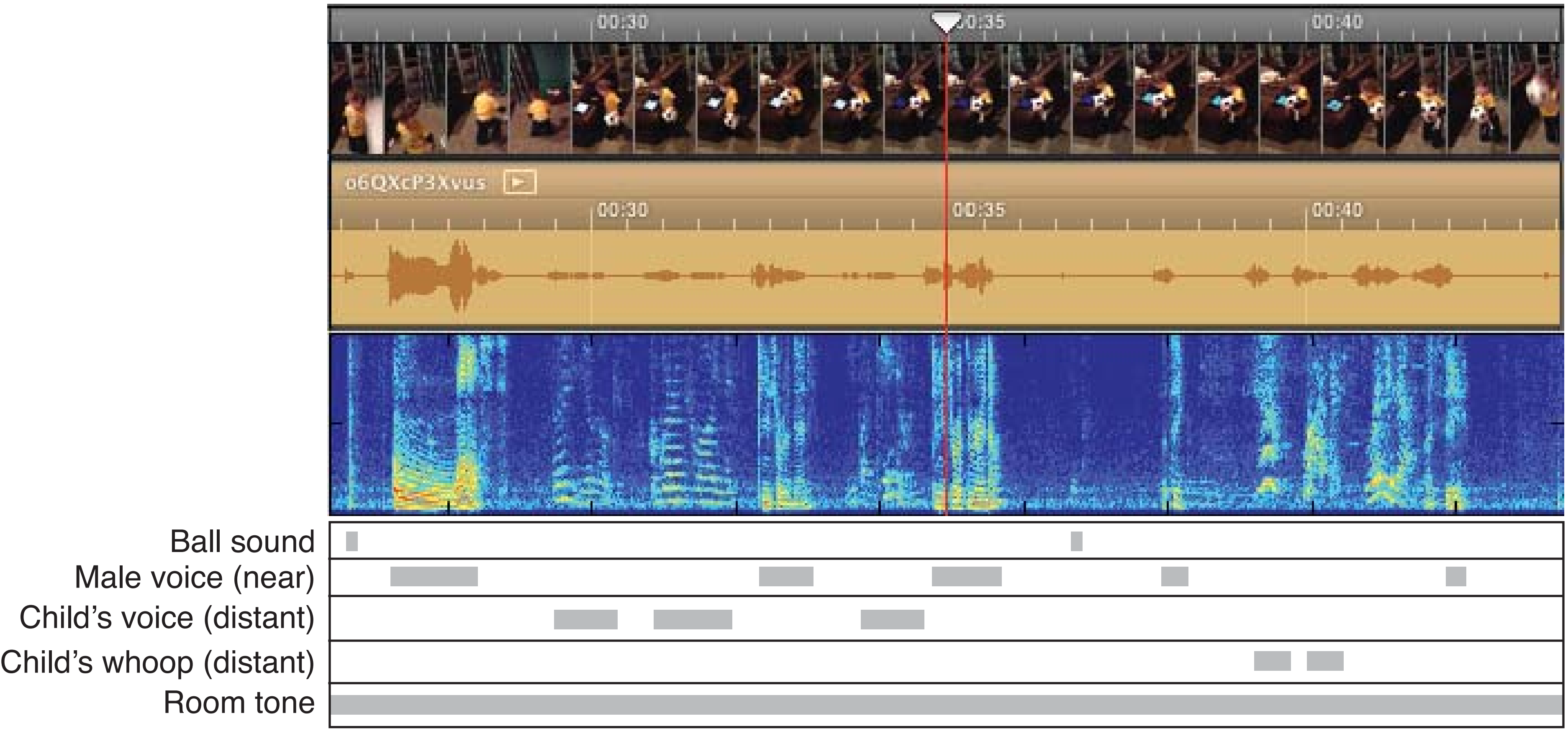
Challenges

Audio signal is composed of the

- actual signal,
- the microphone,
- the environment,
- noise,
- other audio
- compression,
- etc…

Analyzing the Audio Track

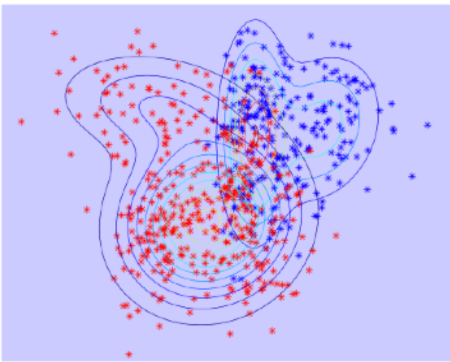
Cameron learns to catch (http://www.youtube.com/watch?v=o6QXcP3Xvus)

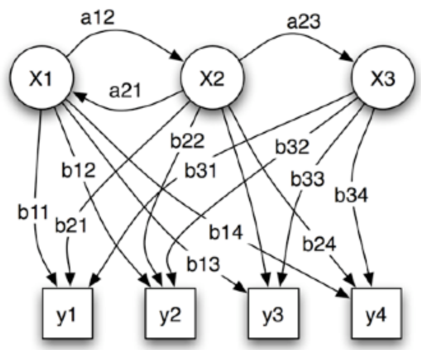


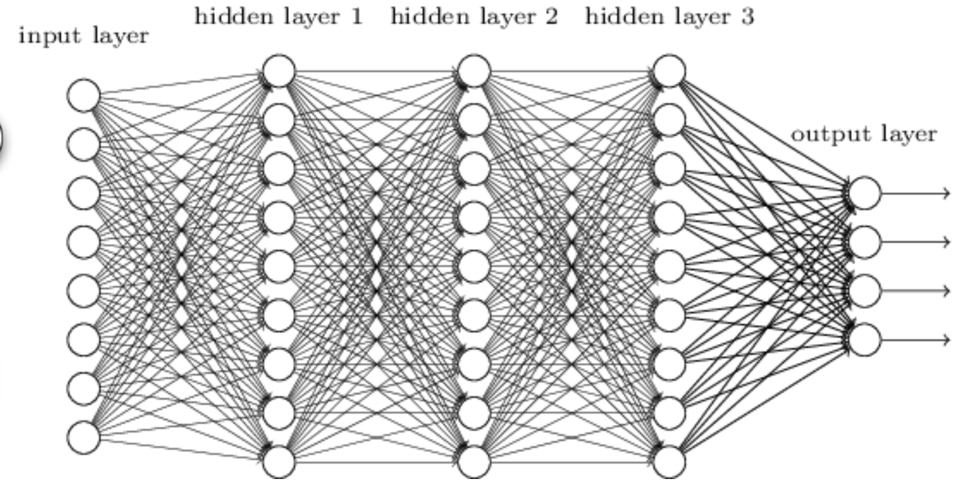
Three High-Level Approaches

- Get into signal processing
- Ignore the issue and just have the machine figure it out
- Do both.

Build a Classifier…







Benjamin Elizalde, Howard Lei, Gerald Friedland, "An i-vector Representation of Acoustic Environments for Audio-based Video Event Detection on User Generated Content" IEEE International Symposium on Multimedia ISM2013. (Anaheim, CA, USA)

Mirco Ravanelli, Benjamin Elizalde, Karl Ni, Gerald Friedland, "Audio Concept Classification with Hierarchical Deep Neural Networks EUSIPCO 2014. (Lisbon, Portugal)

Benjamin Elizalde, Mirco Ravanelli, Karl Ni, Damian Borth, Gerald Friedland. "Audio-Concept Features and Hidden Markov Models for Multimedia Event Detection" Interspeech Workshop on Speech, Language and Audio in Multimedia SLAM 2014. (Penang, Malaysia)

L. Jing, B Liu, A. Janin, J. Choi, J. Bernd. M. Mahoney, G. Friedland: A Discriminative and Compact Audio Representation for Event Detection, ACM Multimedia 2016 (to appear).

Ignore the Signal Properties, build a Classifier

Event	Category	Train	DevTest
E001	Board Tricks	160	111
E002	Feeding Animal	160	111
E003	Landing a Fish	122	86
E004	Wedding	128	88
E005	Woodworking	142	100
E006	Birthday Party	173	0
E007	Changing Tire	110	0
E008	Flash Mob	173	0
E009	Vehicle Unstuck	131	0
E010	Grooming animal	136	0
E011	Make a Sandwich	124	0
E012	Parade	134	0
E013	Parkour	108	0
E014	Repairing Appliance	123	0
E015	Sewing	116	0
Other	Random other	N/A	3755

DIGITAL VIDEO RETRIEVAL at NIST

General Observations

Classifier problems:

- Too much noise
- If it works: Why does it work?
- Idea doesn’t scale to text search: How many classes?
- Annotations: Boundaries not clear

Idea: Go unsupervised!

Zipf?

Percepts

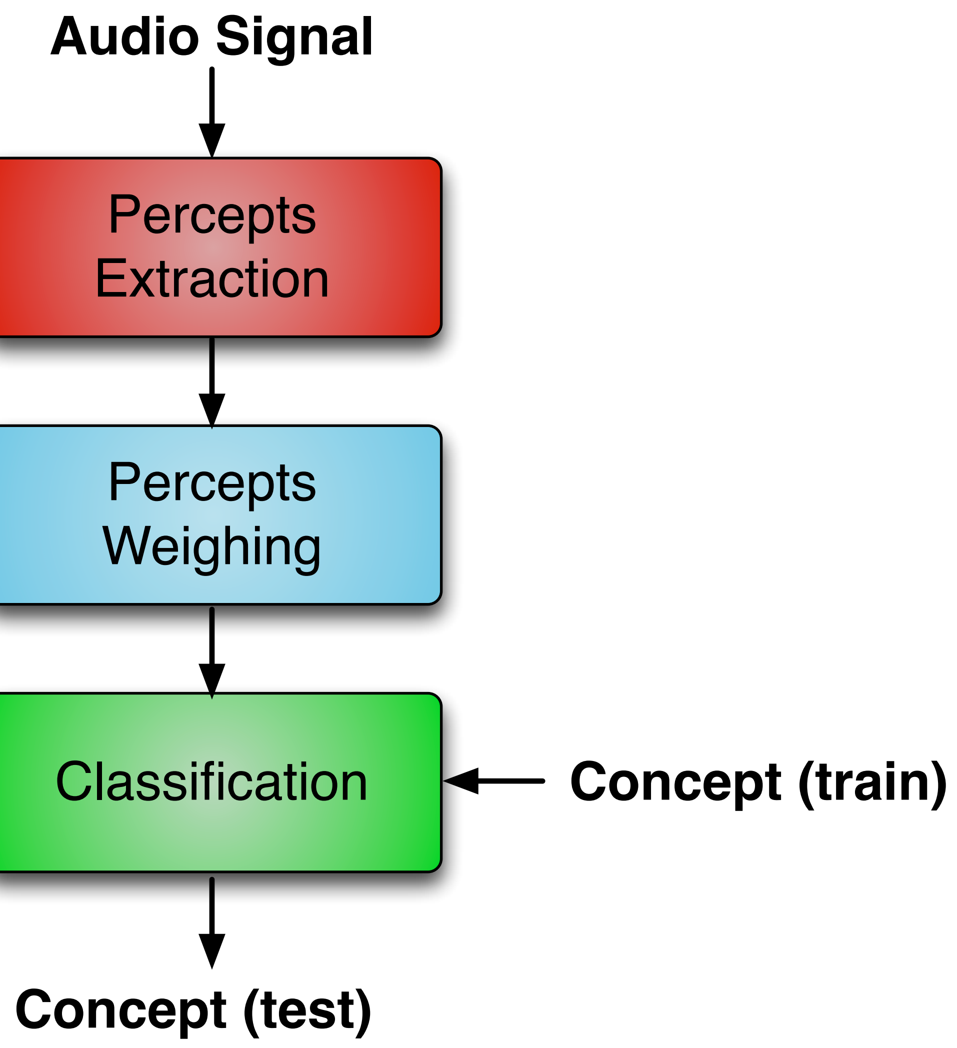
Definition: *an impression of an object obtained by use of the senses.* (Merriam Webster's)

- Well re-discovered in robotics btw...

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

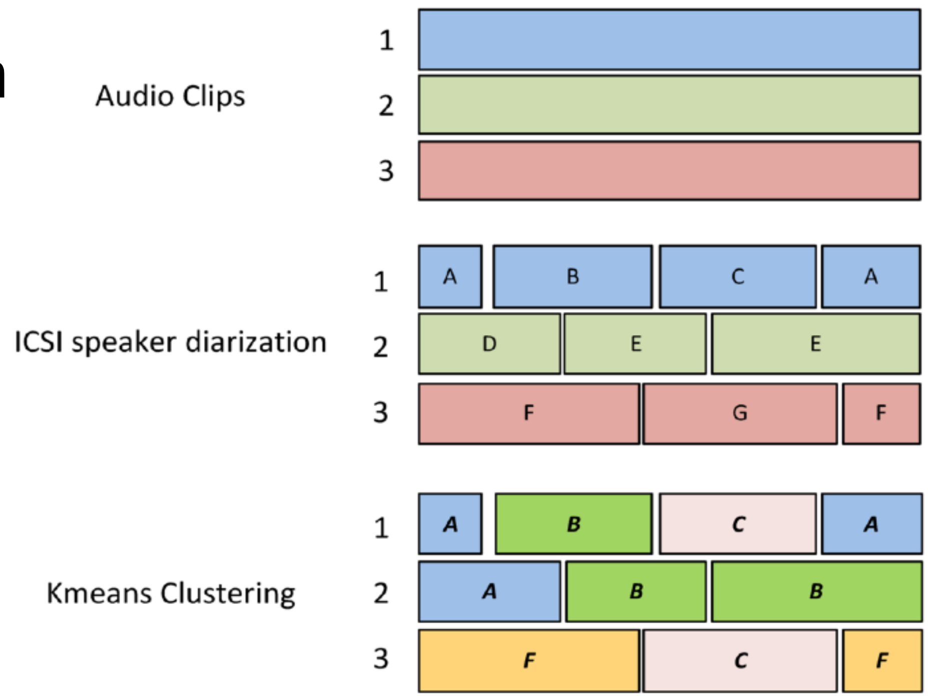
- High number of initial segments
- Features: MFCC19+D+DD+MSG
- Minimum segment length: 30ms
- Train Model(A,B) from Segments A,B belonging to Model(A) and Model(B) and compare using BIC:

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

- 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...

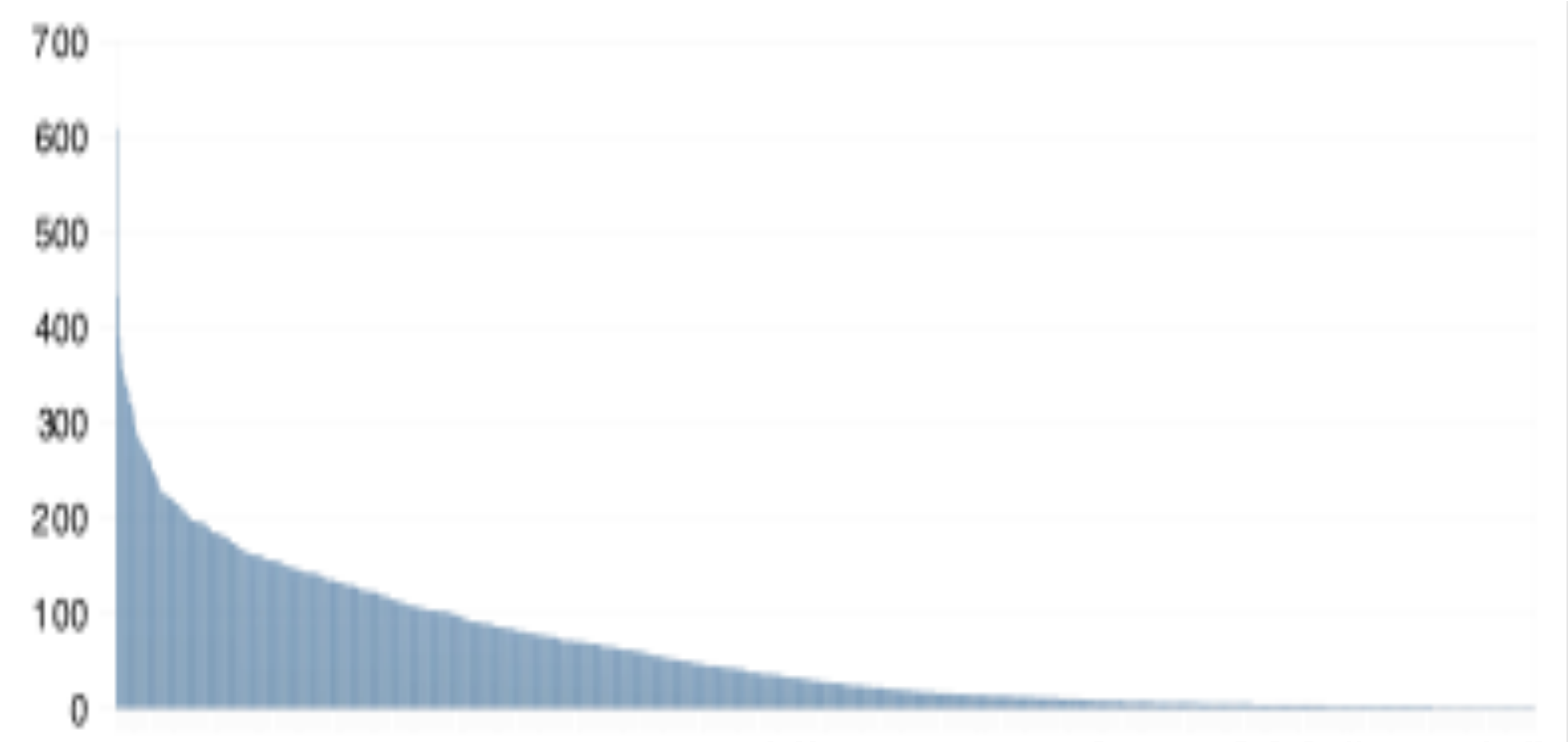
- How many unique “words“ define a particular concept?
- What’s the occurrence frequency of the „words“ per set of video?
- What’s the cross-class ambiguity of the „words“?
- How indicative are the highest frequent „words“ of a set of videos?

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”



Histogram of top-300 “words”.

Long-Tailed Distribution (~ Zipf)

Recap: TF/IDF

$$TF(c_i, D_k) = \frac{\sum_j n_j P(c_i = c_j \mid c_j \in D_k)}{\sum_j}$$

$$IDF(c_i) = \log \frac{|D|}{\sum_k P(c_i \in D_k)}$$

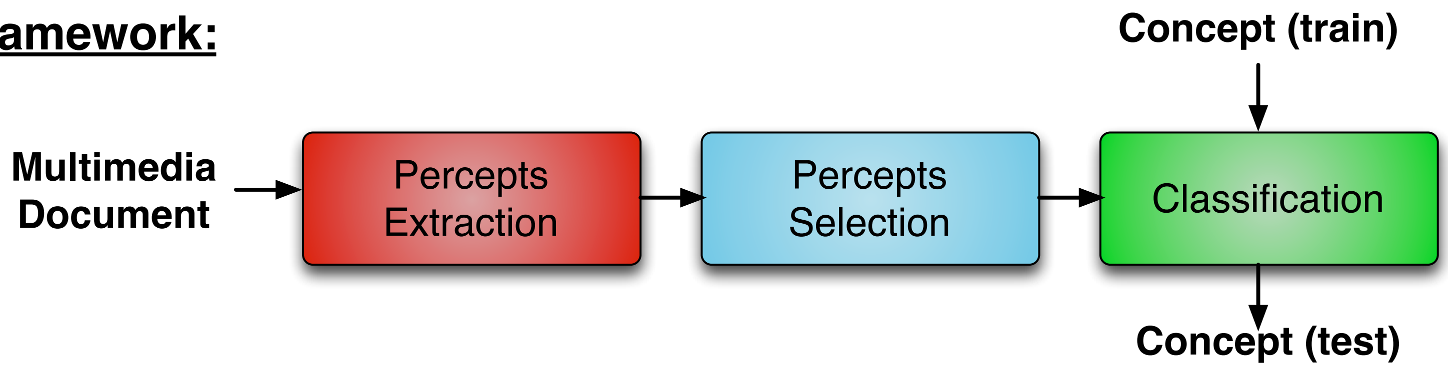
- $TF(c_{i,} D_k)$ is the frequency of "word" c_i in concept D_k.
- $P(c_i = c_j | c_j \in D_k)$ is the probability that "word" c_i equals c_j in concept D_k
- $|D|$ is the total number of concepts
- $P(c_i \in D_k)$ is the probability of "word" c_i in concept D_k

Classify the Words

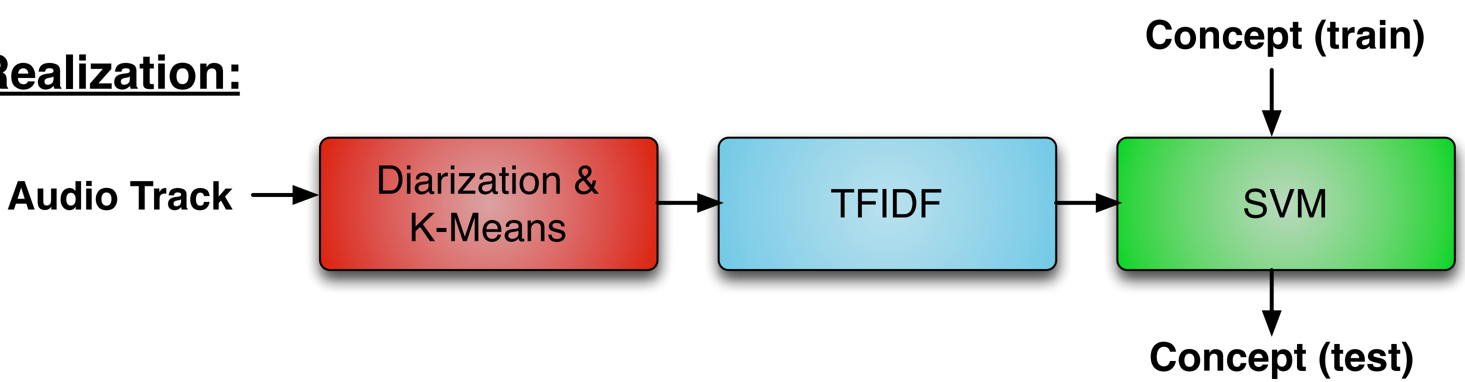
- Have: New input video and set of representative videos
- Need: Does this belong to the same set
- Classifier takes 300 tuples of ("words",TF-IDF values) as input
- Use SVM with Intersection Kernel (IKSVM) / Deep Learning

System Overview

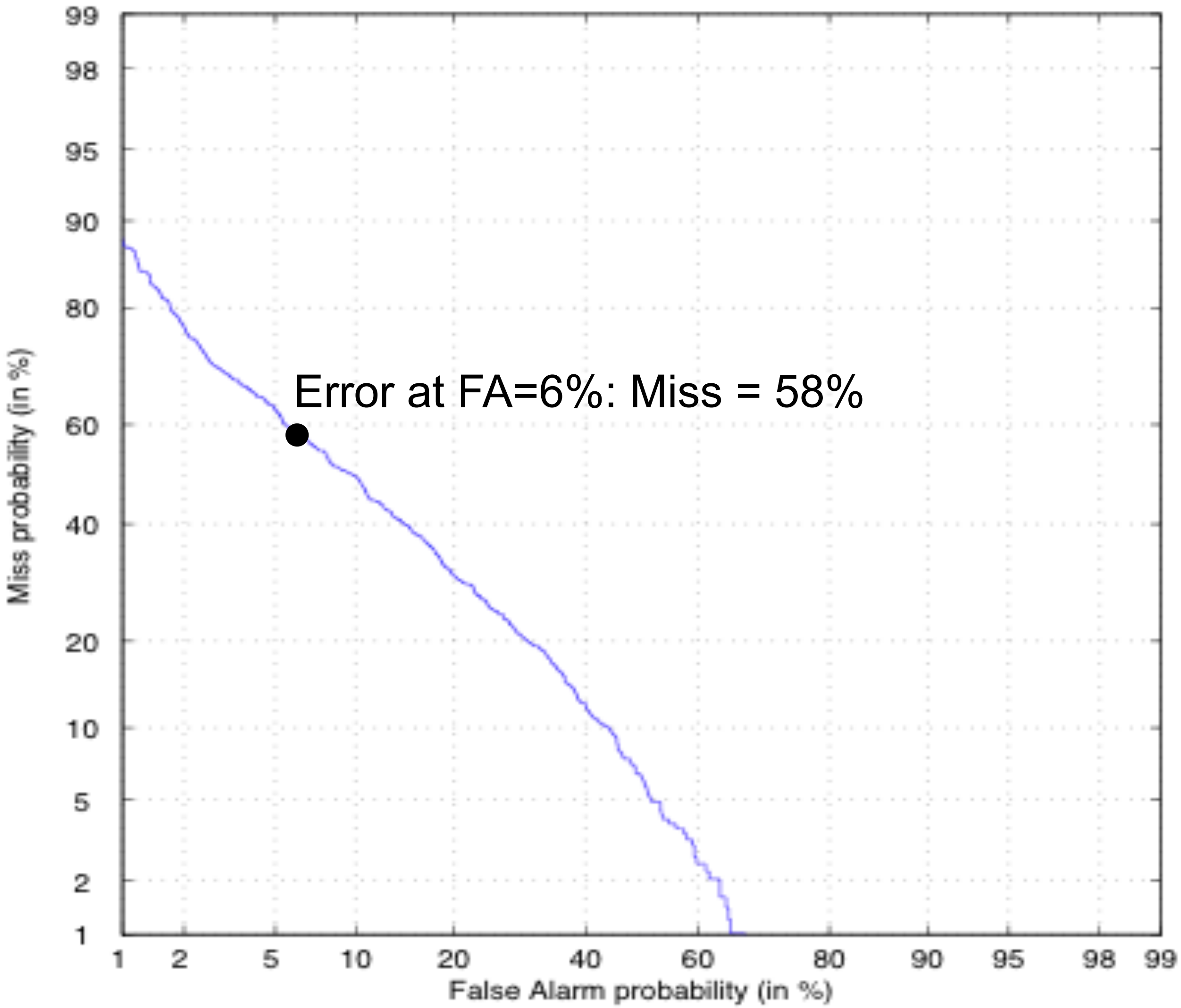
Framework:



Realization:



Audio-Only Detection in TRECVID MED 2011



Visualization of Zipfian Percepts

- Top-1 percepts very representative of concept.

Future Work

- Can Big Data beat signal processing?
- Explore audio analysis methods for computing
- Create multimedia content analysis algorithms that are universal, i.e. work with any data
- Enable scientific discovery by leveraging consumer-produced images and videos.

Thank You!

Questions?

Let there be Zipf

- Let's assume the distributions of Percepts per Concept follows a ranking function: $f(k, s, N) = \frac{1/k^s}{\sum_{n=1}^{N}(1/n^s)}$

  with *k* rank (sorted by highest to lowest frequency), $s=1$, *N* number of Percepts.

Observations

- It follows the CDF is:

$$CDF(k, s, N) = \frac{H_{k,s}}{H_{N,s}}$$

with k rank (sorted by highest to lowest frequency), s=*1*, N number of Percepts and $H_{n,m} = \sum_{k=1}^{n} \frac{1}{k^m}$

Properties of Zipfian “Percepts”

- Distribution allows to distinguish key-percepts from noise: A lot less data is better for training!

Error	Baseline	Top 20	Low 20
False Alarm	6 %	6 %	6 %
Miss	72 %	66 %	79 %
EER	31 %	31 %	35 %

Properties of: Zipfian “Properties”

- Distribution allows prediction of “completeness” of training data

Top N	Actual Hits	Predicted Hits	*Error*	Ambiguity
1	17 %	16 %	1 %	0 %
3	35 %	30 %	5 %	0 %
5	46 %	36 %	10 %	20 %
10	56 %	46 %	10 %	24 %
20	84 %	57 %	27 %	27 %
40	99 %	68 %	31 %	31 %