Deriving Knowledge from Audio and Multimedia Data

?

Dr. Gerald Friedland

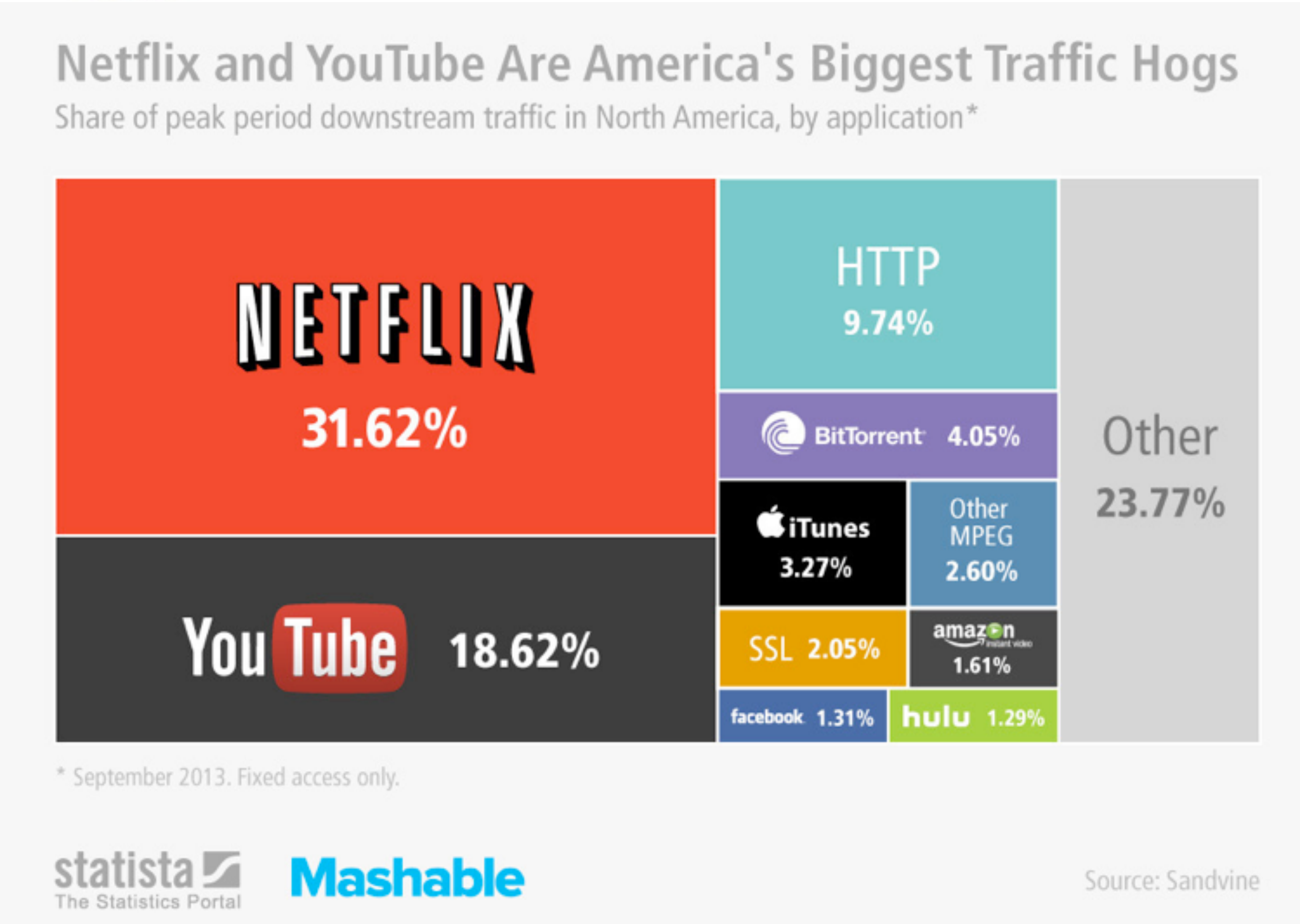
Director Audio and Multimedia Lab

International Computer Science Institute

Berkeley, CA

fractor@icsi.berkeley.edu

Multimedia in the Internet is Growing



Multimedia People at ICSI

Research Staff

- Jaeyoung Choi
- Adam Janin

Research Assistants

- Julia Bernd
- Bryan Morgan

Graduate Students

- Khalid Ashraf
- (T.J. Tsai)

Current Visitors

- Liping Jing

Affiliated Researchers

- Dan Garcia, Kurt Keutzer (UCB)
- Howard Lei (Cal State Hayward)
- Karl Ni (Lawrence Livermore Lab)

Undergraduates

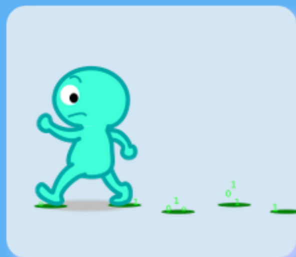
- Itzel Martinez, Jessica Larson, Marissa Pita, Florin Langer, Justin Kim, Regina Ongawarsito, Megan Carey

What are we interested in?

Three main themes:

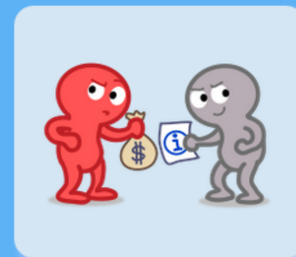
- Audio Analytics
- Video Retrieval
- Privacy (Education)

Ten Principles for Online Privacy

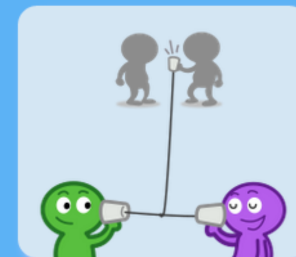


You're Leaving Footprints

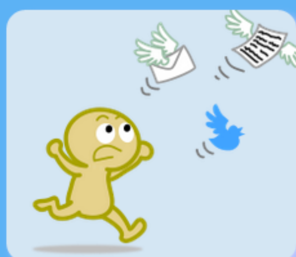
There's No Anonymity



Information Is Valuable



Someone Could Listen



Sharing Releases Control

Search Is Improving

Online Is Real

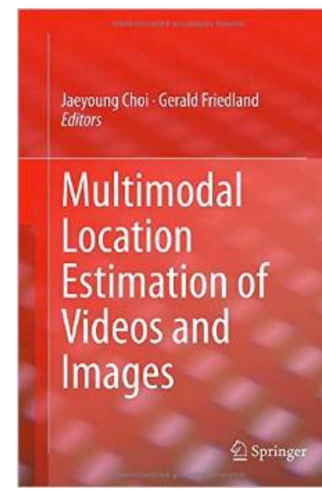
Identity Isn't Guaranteed

You Can't Escape

Privacy Requires Work

http://teachingprivacy.org

Multimodal Location Estimation

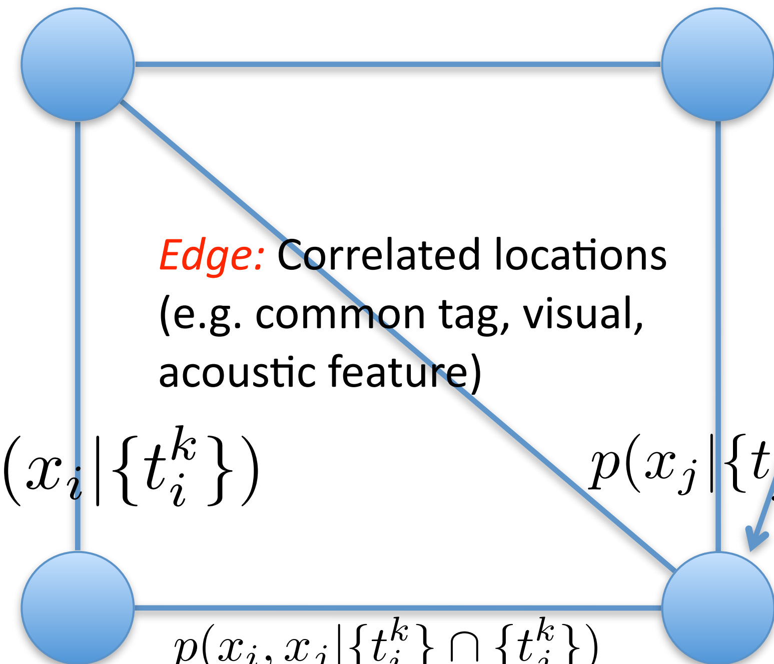


http://mmle.icsi.berkeley.edu

Intuition for the Approach

{berkeley, sathergate, campanile}

{berkeley, haas}



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

Node: Geolocation of video

$p(x_i|\{t_i^k\})$

$p(x_j|\{t_j^k\})$

$p(x_i, x_j|\{t_i^k\} \cap \{t_j^k\})$

{campanile}

{campanile, haas}

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

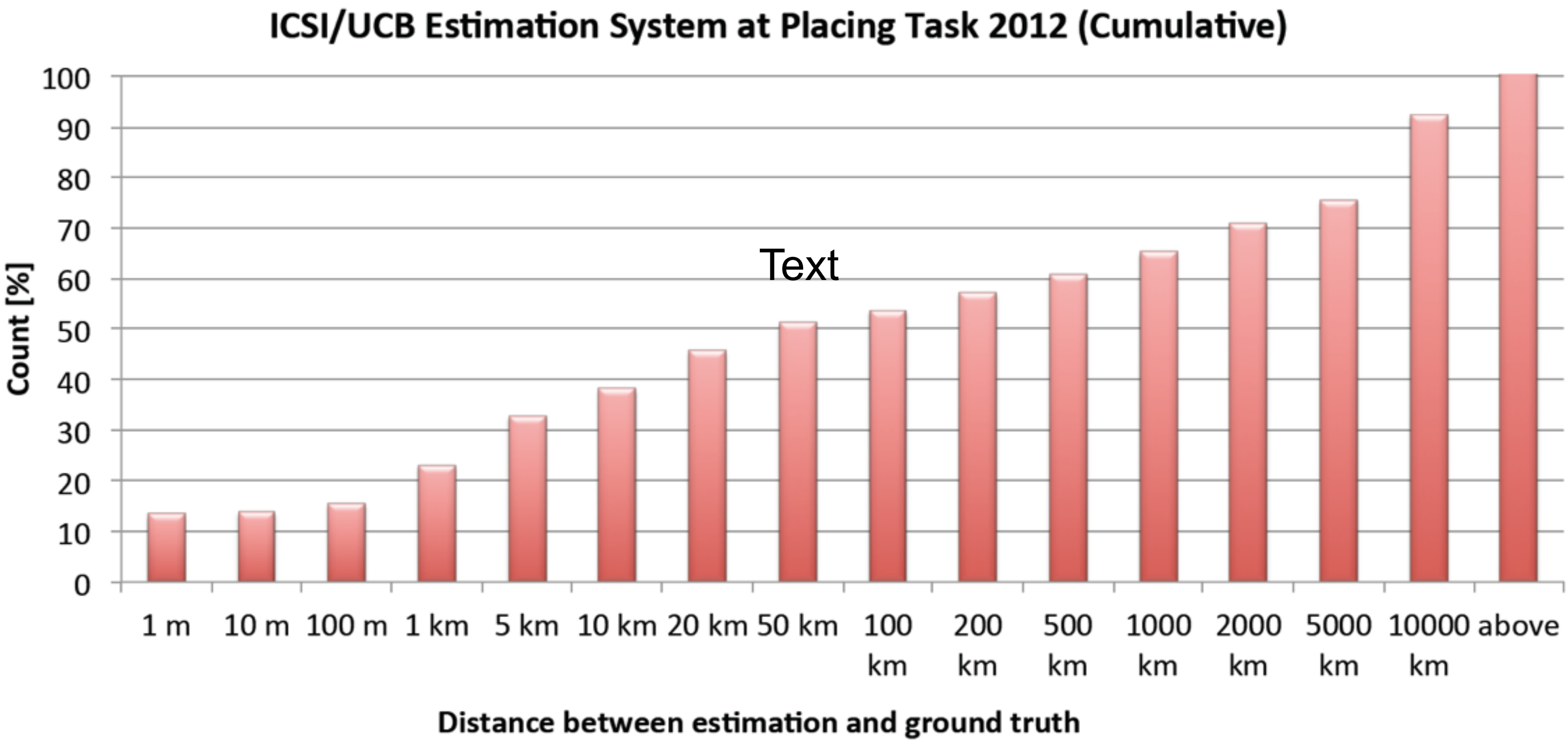
ICSI

INTERNATIONAL COMPUTER SCIENCE INSTITUTE

MediaEval Benchmark

MediaEval Benchmarking Initiative for Multimedia Evaluation

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



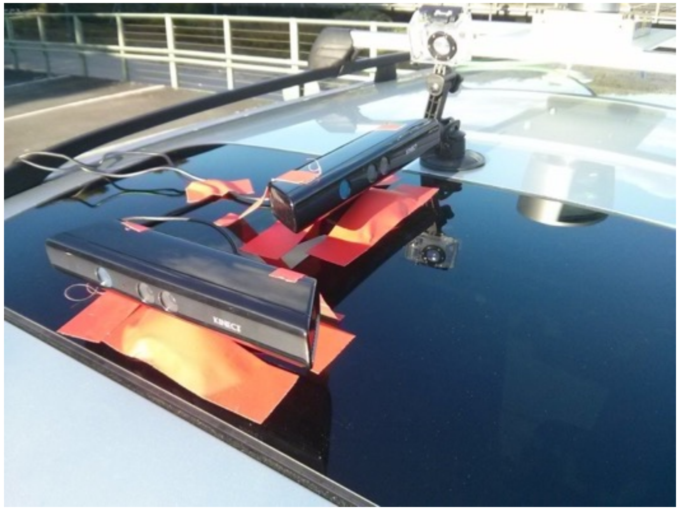
J. Choi, G. Friedland, V. Ekambaram, K. Ramchandran: "Multimodal Location Estimation of Consumer Media: Dealing with Sparse Training Data," in Proceedings of IEEE ICME 2012, Melbourne, Australia, July 2012.

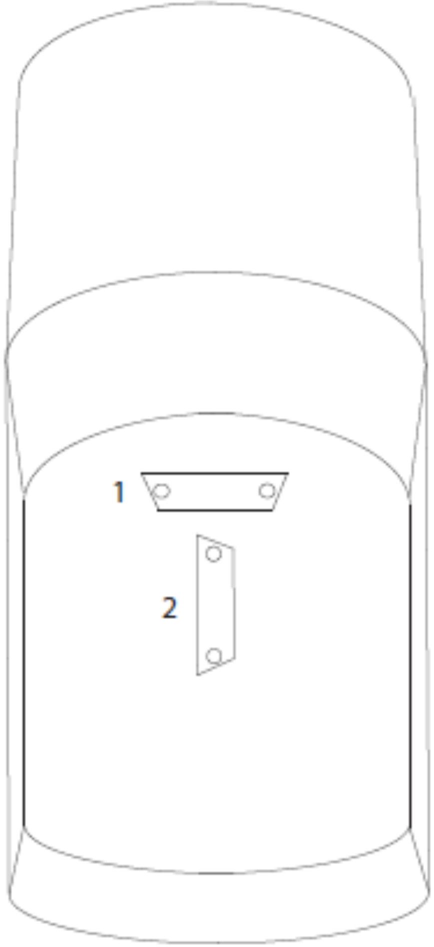
An Experiment

Listen!

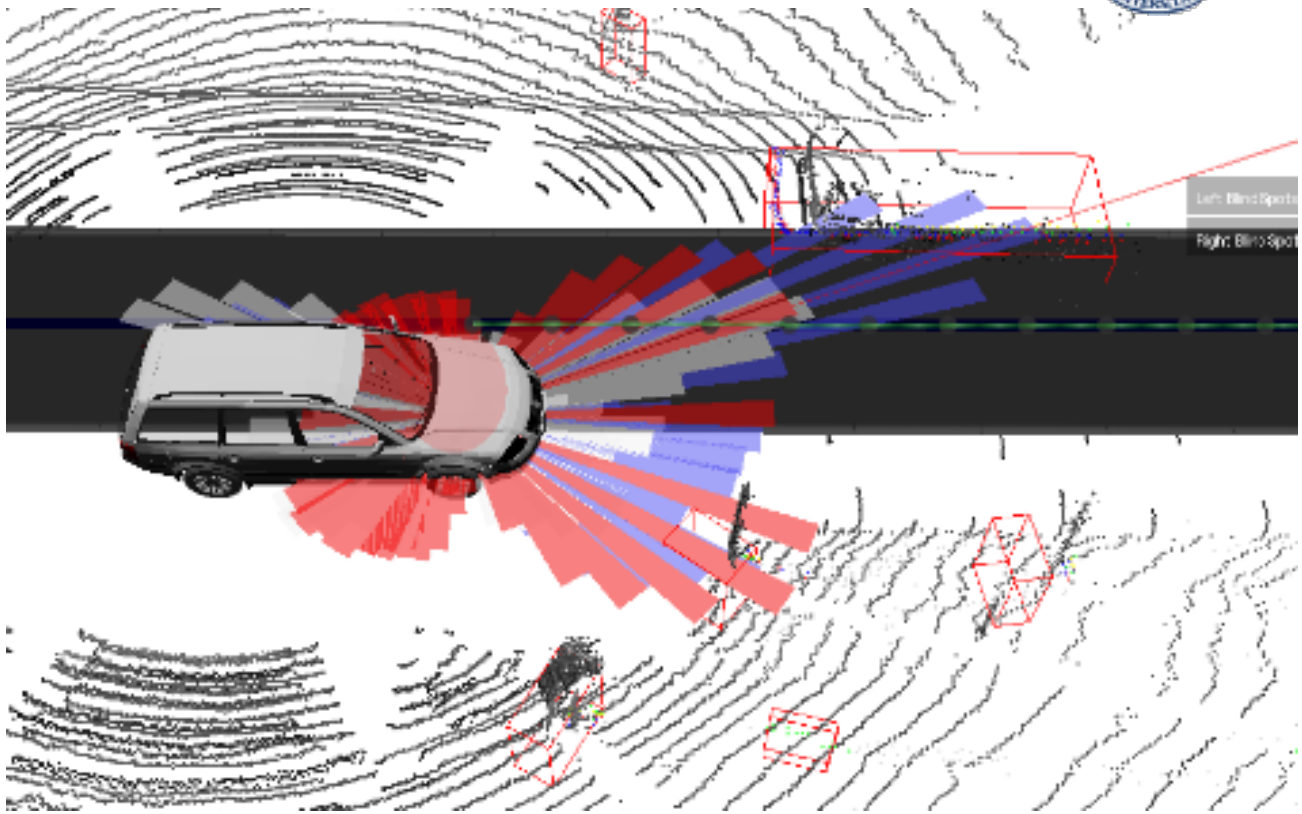
- Which city was this recorded in?
  Pick one of: Amsterdam, Bangkok, Barcelona, Beijing, Berlin, Cairo, CapeTown, Chicago, Dallas, Denver, Duesseldorf, Fukuoka, Houston, London, Los Angeles, Lower Hutt, Melbourne, Moscow, New Delhi, New York, Orlando, Paris, Phoenix, Prague, Puerto Rico, Rio de Janeiro, Rome, San Francisco, Seattle, Seoul, Siem Reap, Sydney, Taipei, Tel Aviv, Tokyo, Washington DC, Zuerich
- Solution: Tokyo, highest confidence score!

Autonomous Vehicles





Result



- Blue histogram shows combined likelihoods, example – sound source vehicle in red box
- Most likely direction shown as a red line

INTERNATIONAL COMPUTER SCIENCE INSTITUTE

Sound Recognition

- Car honk
- Glass break
- Fire alarm
- Person yelling
- etc…

AUDEME

Multimedia Retrieval

Consumer-Produced Videos are Growing

- 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.

Google "giving directions to a location"

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

3 results (0.13 seconds)

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"**

Challenge

User-provided tags are:

- sparse
- any language
- imply random context

Solution: Use the actual audio and video content for search

INTERNATIONAL COMPUTER SCIENCE INSTITUTE

The Multimedia Commons Project

A research community around the YFCC100M dataset and the YLI corpus

- 100M images, 1M videos
- Hosted on Amazon
- CFT with SEJITs-based content analysis tools
- Annotations: YLI corpus

http://multimediacommons.org/

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 (to appear).

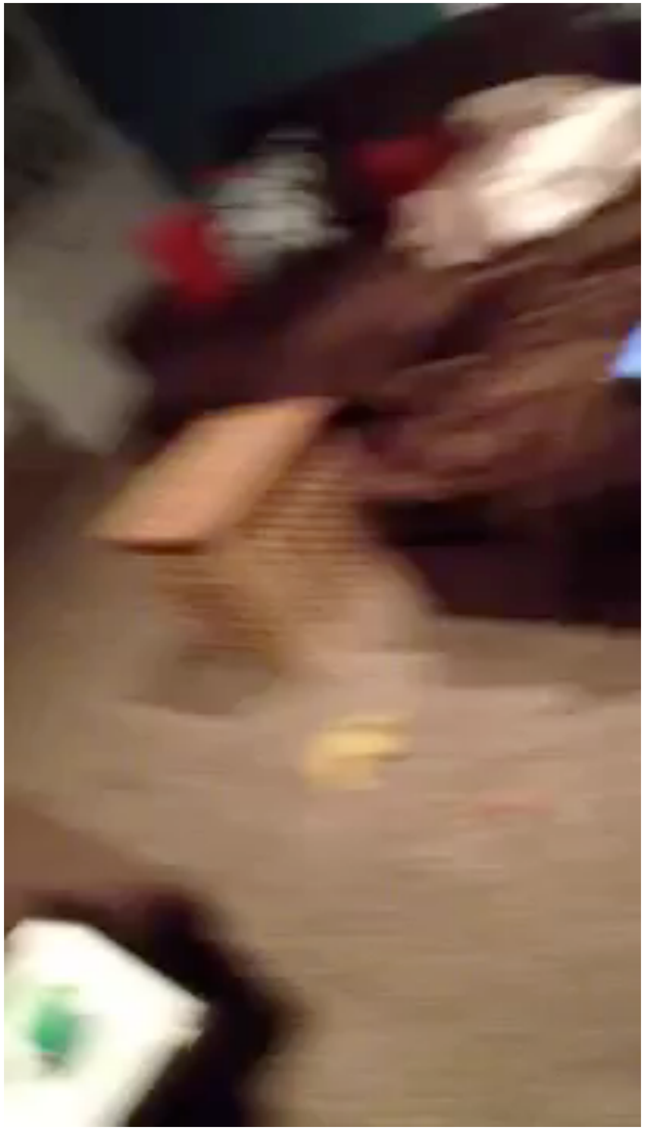
Restricting Ourselves to Audio Content (for now)

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

Properties of Consumer-Produced Videos

- No constraints in angle, number of cameras, cutting
- 70% heavy noise
- 50% speech, any language
- 40% dubbed
- 3% professional content

Example Video



Challenges

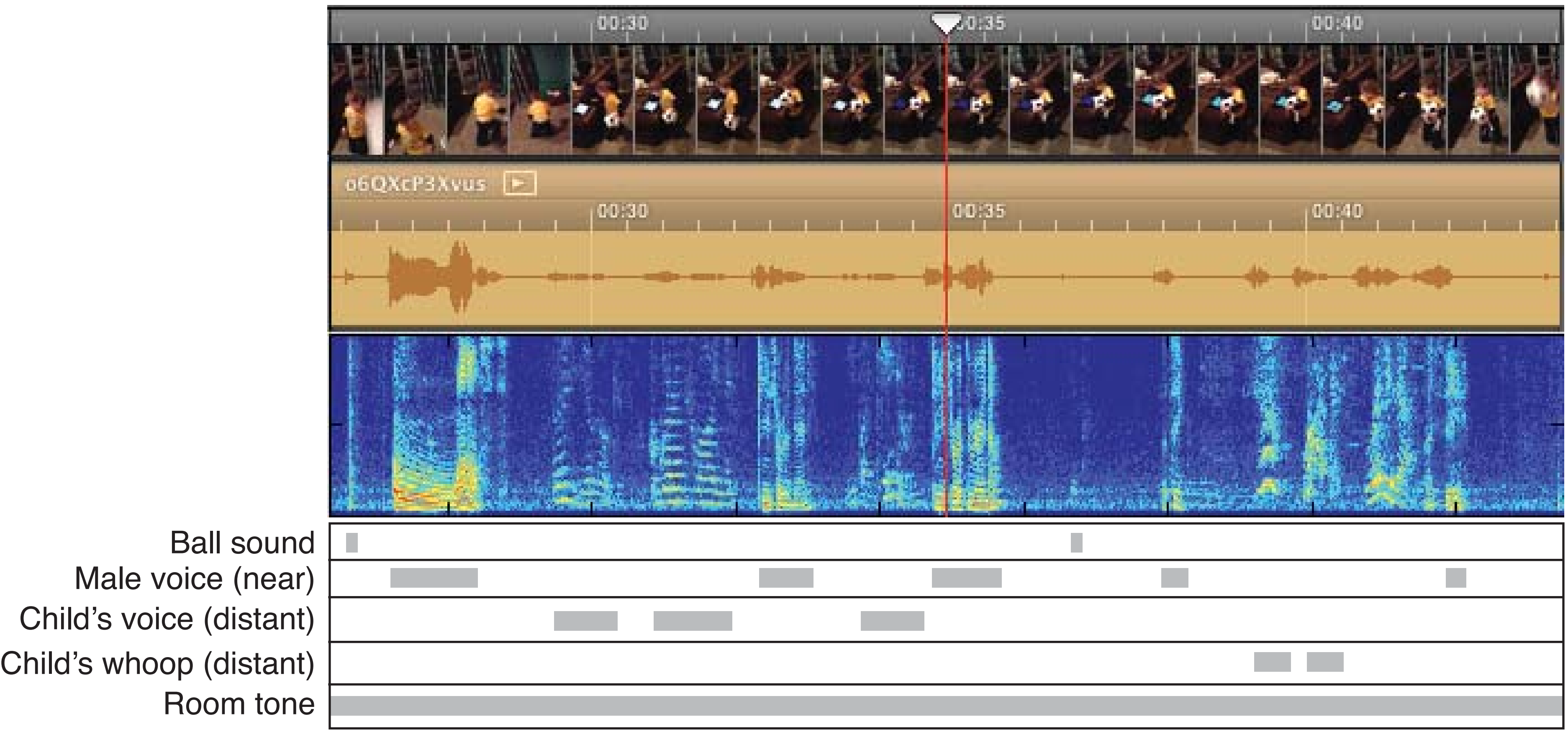
Audio signal is composed of the

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

INTERNATIONAL COMPUTER SCIENCE INSTITUTE

Analyzing the Audio Track

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



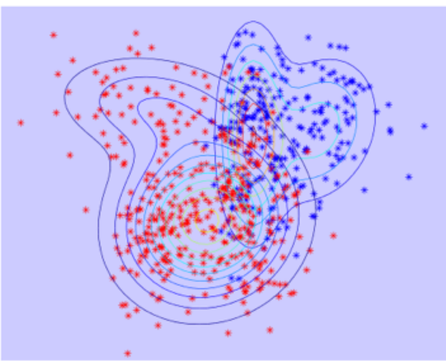
Three High-Level Approaches

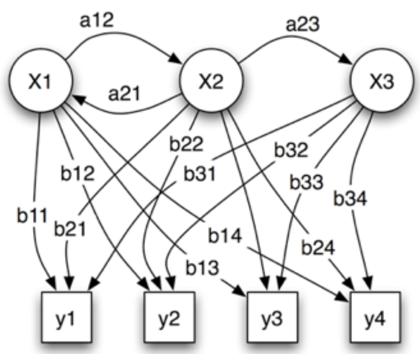
- Get into signal processing
- Ignore the issue and just have the machine figure it out
- Do both.

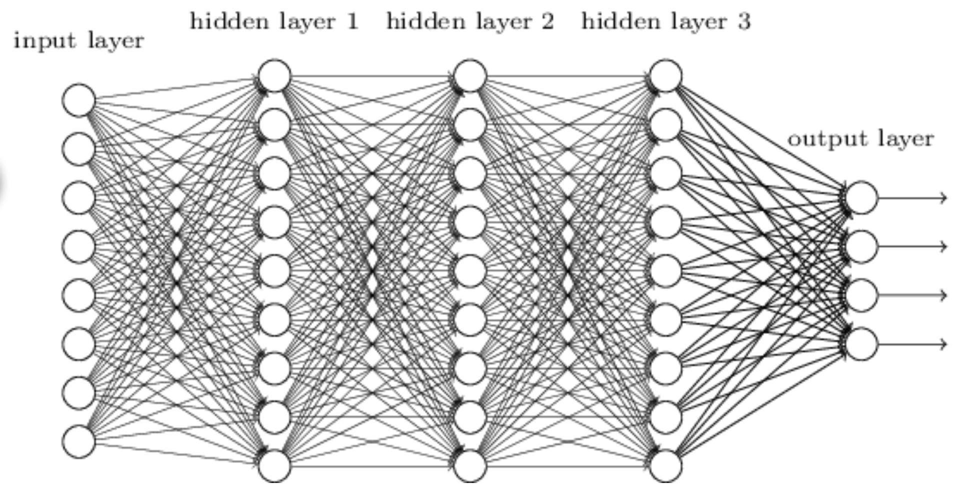
Ignore the Signal Properties, build a Classifier

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

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 Neur Networks EUSIPCO 2014. (Lisbon, Portugal)

Benjamin Elizalde, Mirco Ravanelli, Karl Ni, Damian Borth, Gerald Friedland. “Audio-Concept Features and Hidden Marko Models for Multimedia Event Detection” Interspeech Workshop on Speech, Language and Audio in Multimedia SLAM 201 (Penang, Malaysia)

General Observations

Classifier problems:

- Too much noise
- If it works: Why does it work?
- Idea doesn't scale to text search

Other Work: TRECVID MED 2010

Making a cake

Batting a run in

Assembling a shelter

TrecVid MED 2010: Classifier Ensembles

Human Action Concepts	*Scene Concepts*	*Audio Concepts*
▪ Person walking	▪ Indoor kitchen	▪ Outdoor rural
▪ Person running	▪ Outdoor with grass/trees visible	▪ Outdoor urban
▪ Person squatting	▪ Baseball field	▪ Indoor quiet
▪ Person standing up	▪ Crowd (a group of 3+ people)	▪ Indoor noisy
▪ Person making/assembling stuffs with hands (hands visible)	▪ Cakes (close-up view)	▪ Original audio
		▪ Dubbed audio
▪ Person batting baseball		▪ Speech comprehensible
		▪ Music
		▪ Cheering
		▪ Clapping

Yu-Gang Jiang, Xiaohong Zeng, Guangnan Ye, Subhabrata Bhattacharya, Dan Ellis, Mubarak Shah, Shih-Fu Chang: ***Columbia-UCF TRECVID2010 Multimedia Event Detection: Combining Multiple Modalities, Contextual Concepts, and Temporal Matching***, Proceedings of TrecVid 2010, Gaithersburg, MD, December 2010.

General Observations

- Classifier Ensembles problematic:
  - Which classifiers to build?
  - Training data?
  - Annotation?
  - Idea doesn't scale... or does it?

Alexander Hauptmann, Rong Yan, and Wei-Hao Lin: **"How many high-level concepts will fill the semantic gap in news video retrieval?",** in Proceedings of the 6th ACM international conference on Image and Video retrieval, CIVR '07, pages 627–634, New York, NY, USA, 2007. ACM.

Percepts

Definition: *an impression of an object obtained by use of the senses.*

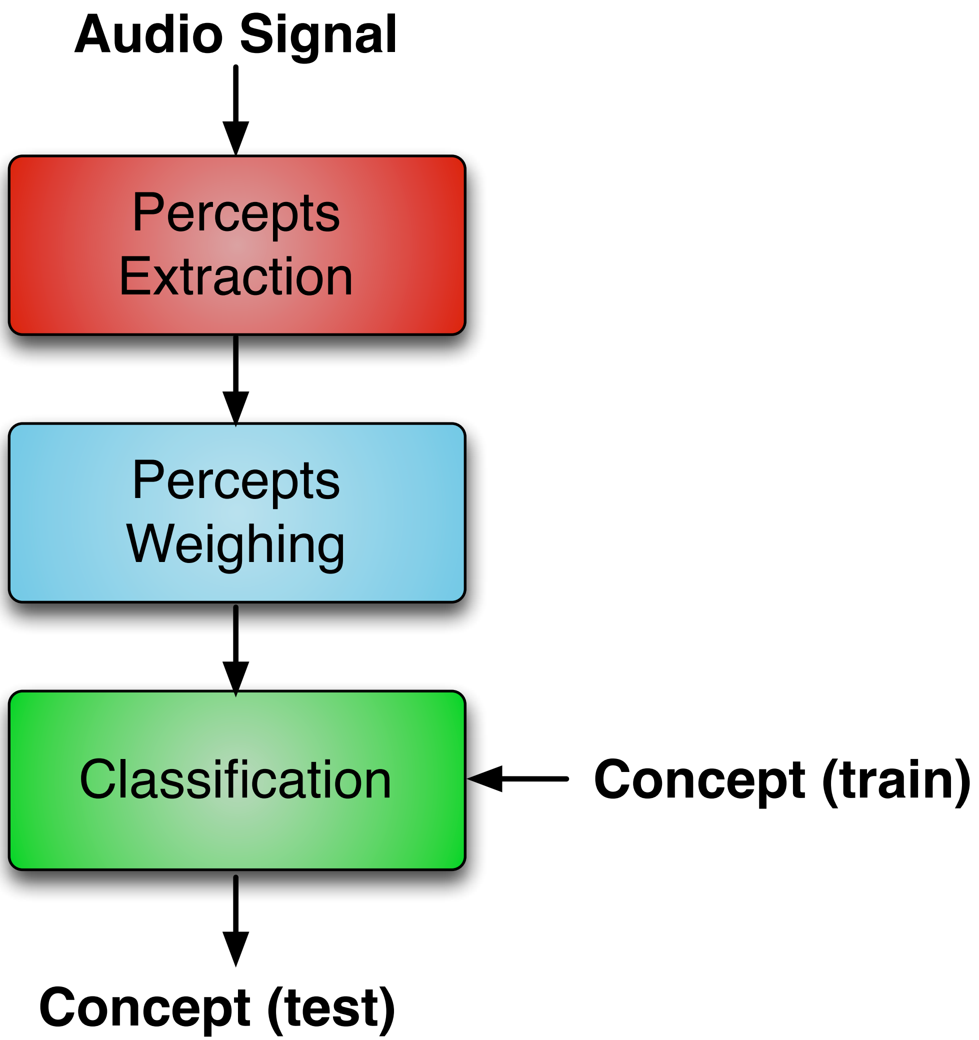
(Merriam Webster's)

- Well re-discovered in robotics btw...

My 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



Finding Perceptual Similar Units

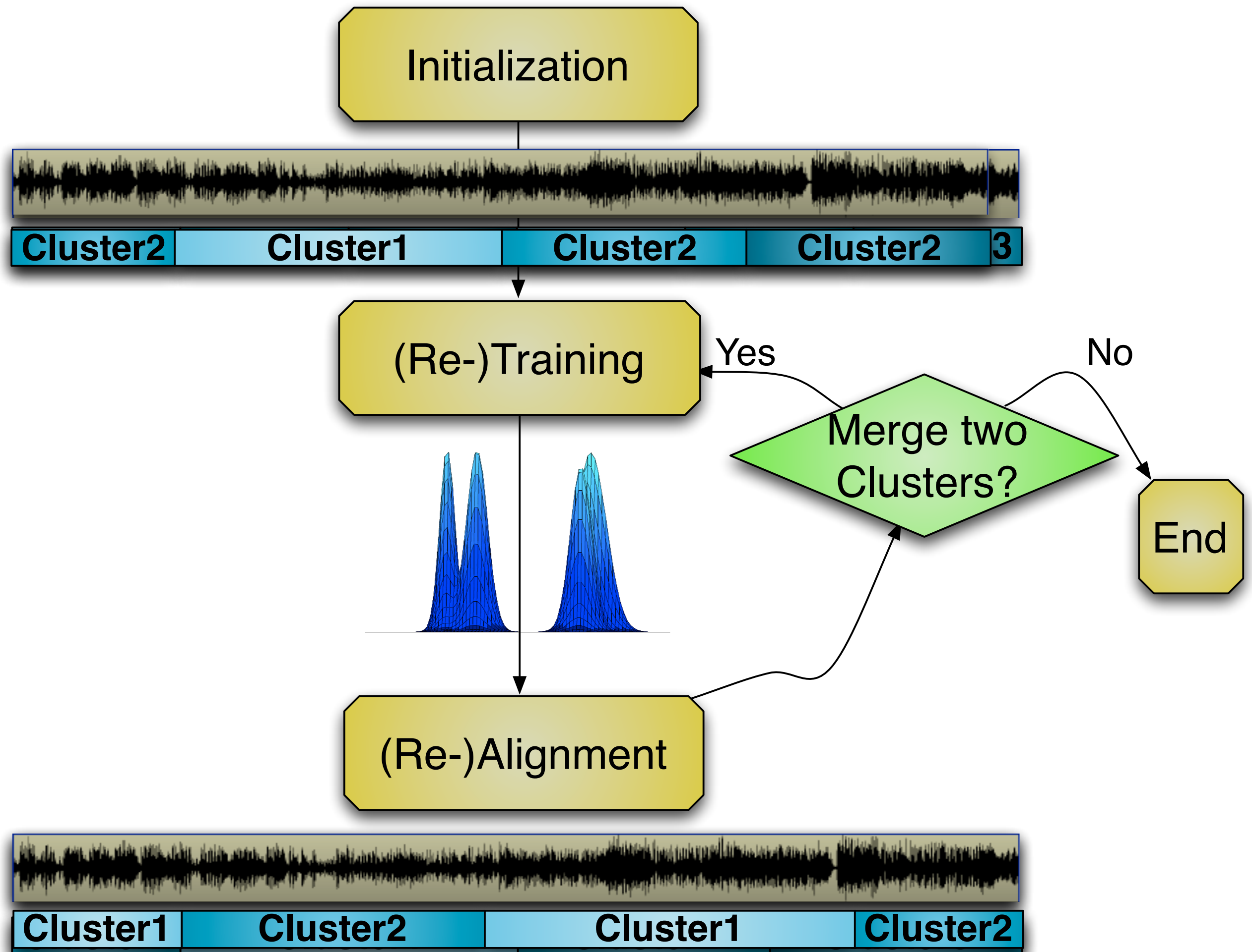
- „Edge detection“ like in Image Processing doesn’t work
- Building a classifier for similar audio requires too many parameters
- What’s a similarity metric?

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:

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

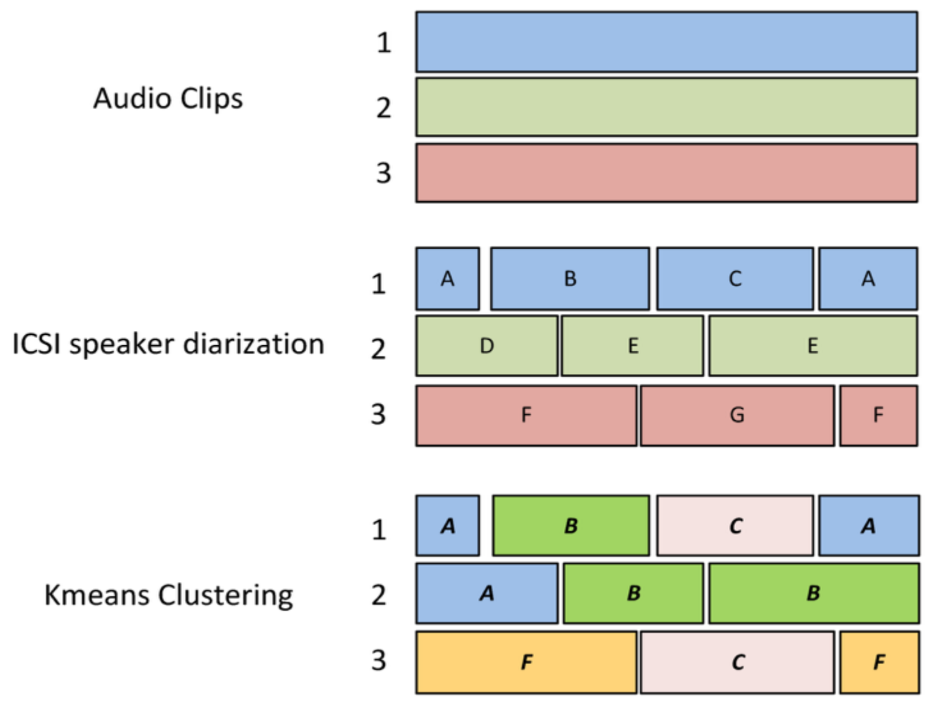
Percepts Extraction



- Start with too many clusters (initialized randomly)
- Purify clusters by comparing and merging similar clusters
- Resegment and repeat until no more merging needed

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...

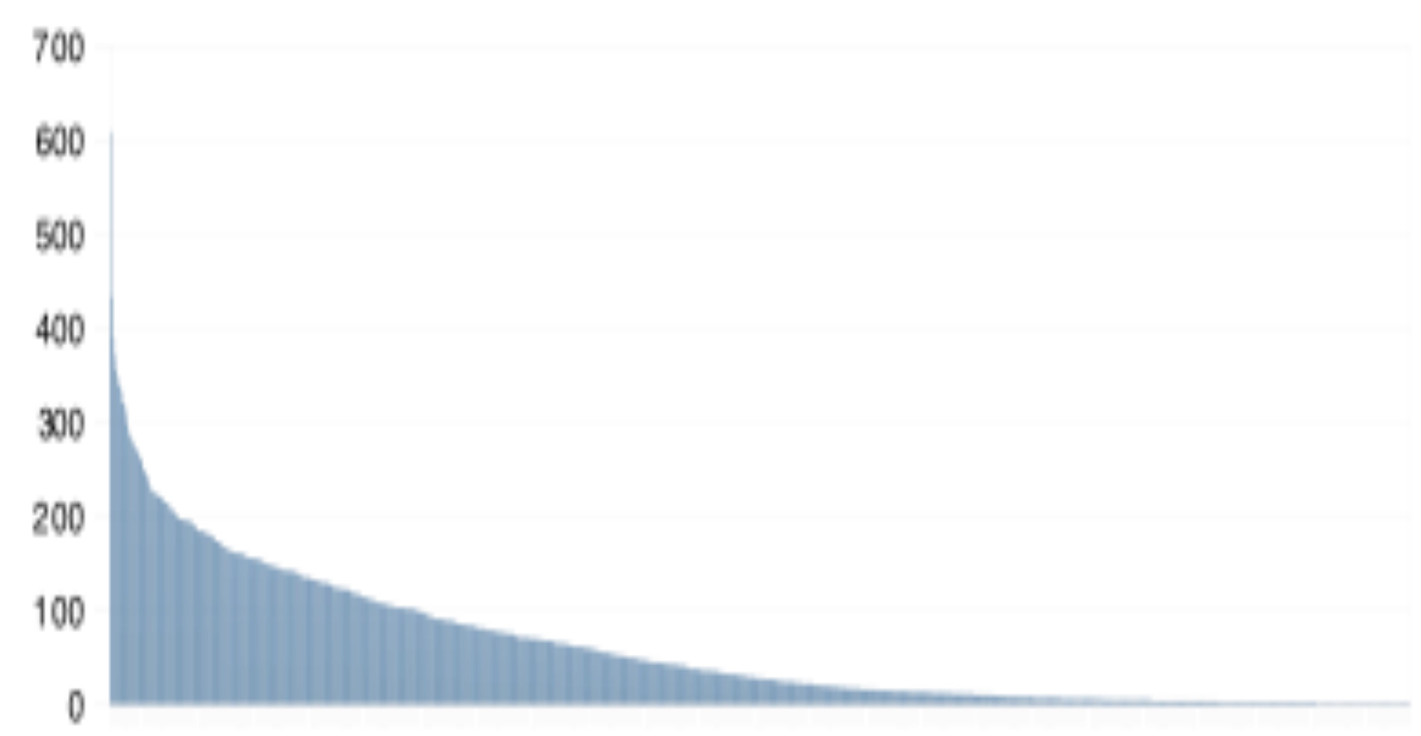
- How many unique “words“ define a particular concept?
- What’s the occurrence frequency of the „words“ per set of video?
- What’s the cross-class ambiguity of the „words“?
- How indicative are the highest frequent „words“ of a set of videos?

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)

TF/IDF on Supervectors

- Zipf distribution already observed by other researchers as well (Bhiksha Raj, Alex Hauptman, Sad Ali, etc)
- Zipf distribution allows to treat supervector representation of percepts as “words” in a document.
- Use TF/IDF for assigning weights

Recap: TF/IDF

$$TF(c_i, D_k) = \frac{\Sigma_j n_j P(c_i = c_j \mid c_j \in D_k)}{\Sigma_j} \qquad IDF(c_i) = \log \frac{|D|}{\Sigma_k P(c_i \in D_k)}$$

- $TF(c_{i,}\ D_k)$ 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

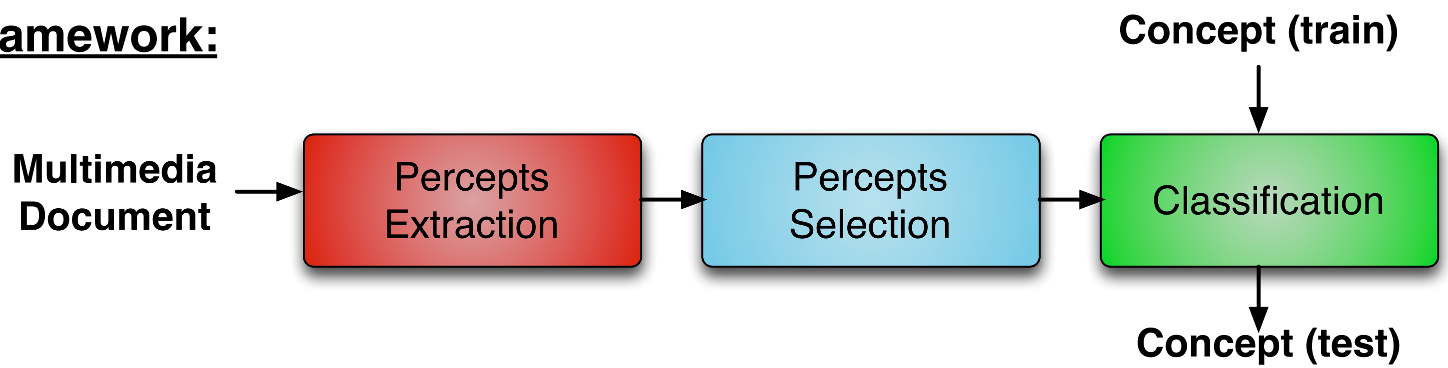
Classify the Words

- 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

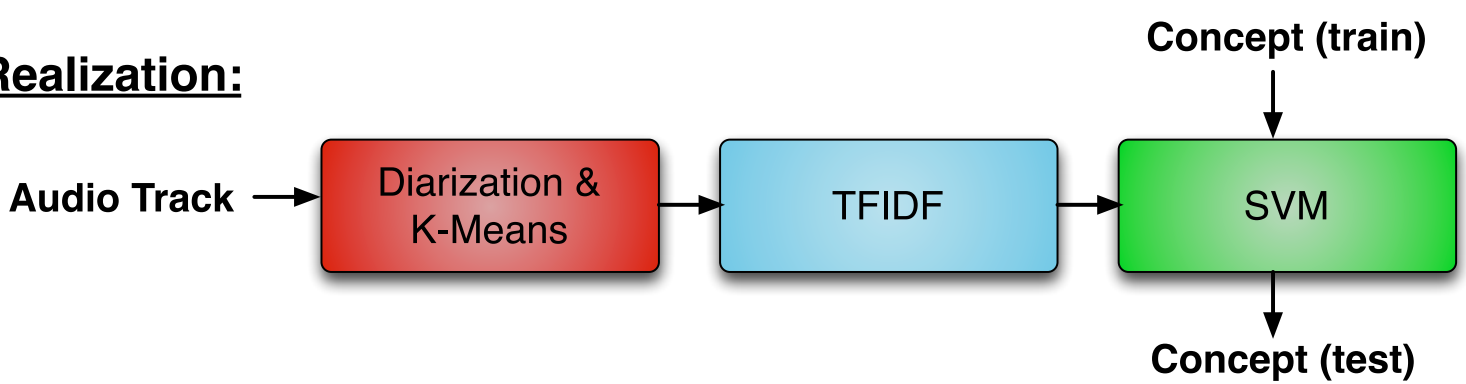
INTERNATIONAL COMPUTER SCIENCE INSTITUTE

System Overview

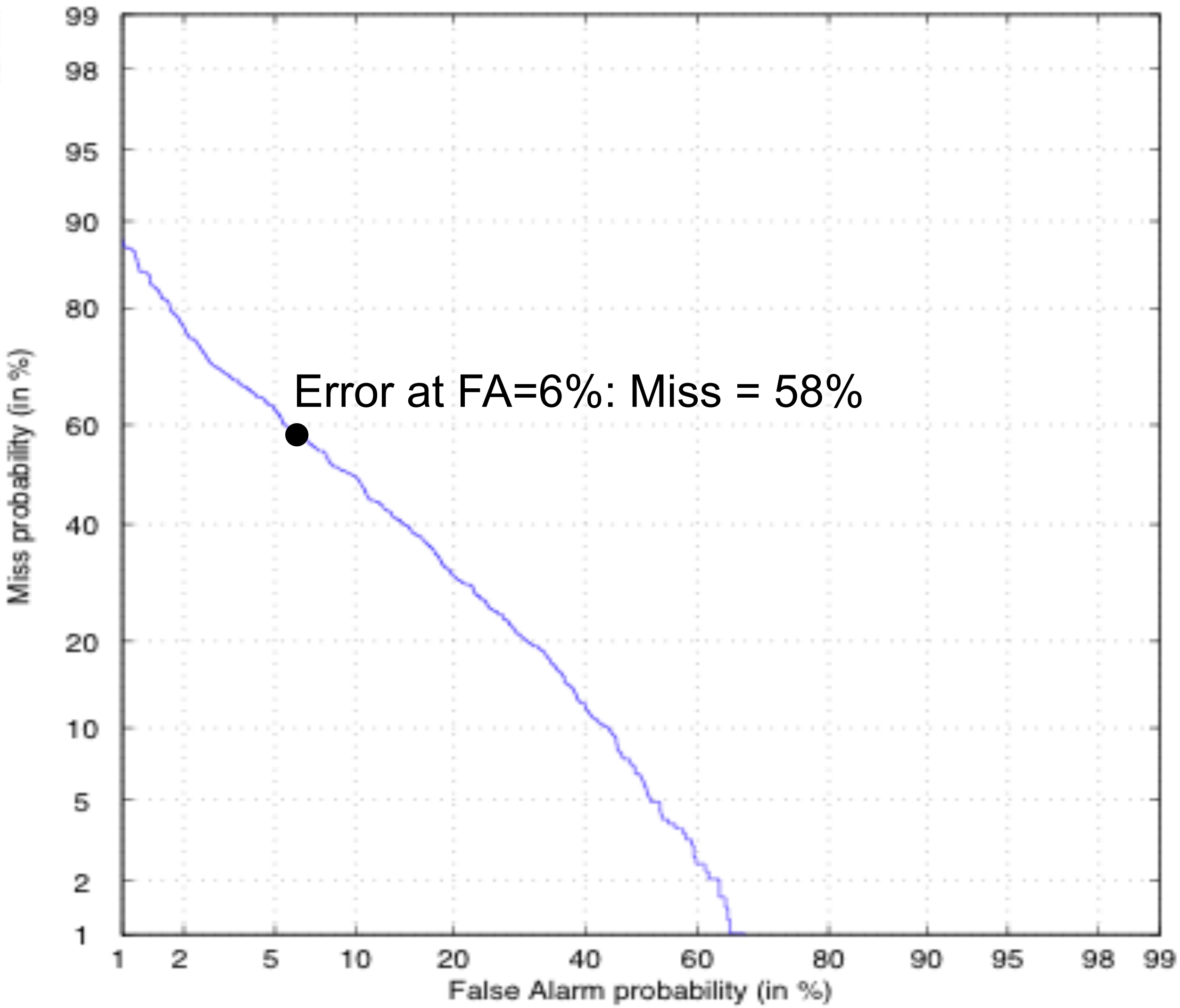
Framework:



Realization:



Audio-Only Detection on MED-DEV11



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 %

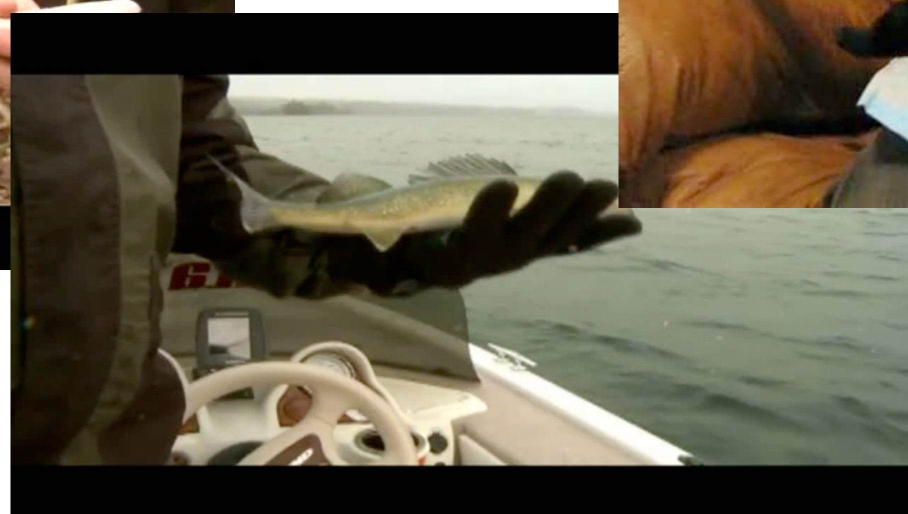
Properties of: Zipfian “Properties”

- 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 %

Visualization of Zipfian Percepts

- Top-1 percepts very representative of concept.



Demo/Development Interface

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

Open Questions

- Exploit multimodality early
- Reduce ambiguities in percepts extraction
- What's the optimal percept? How can we tune?
- Exploit temporal dimension better: ("sentences", "paragraphs"?)
- Is there are universal set of percepts?

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

Thank You!
Questions?