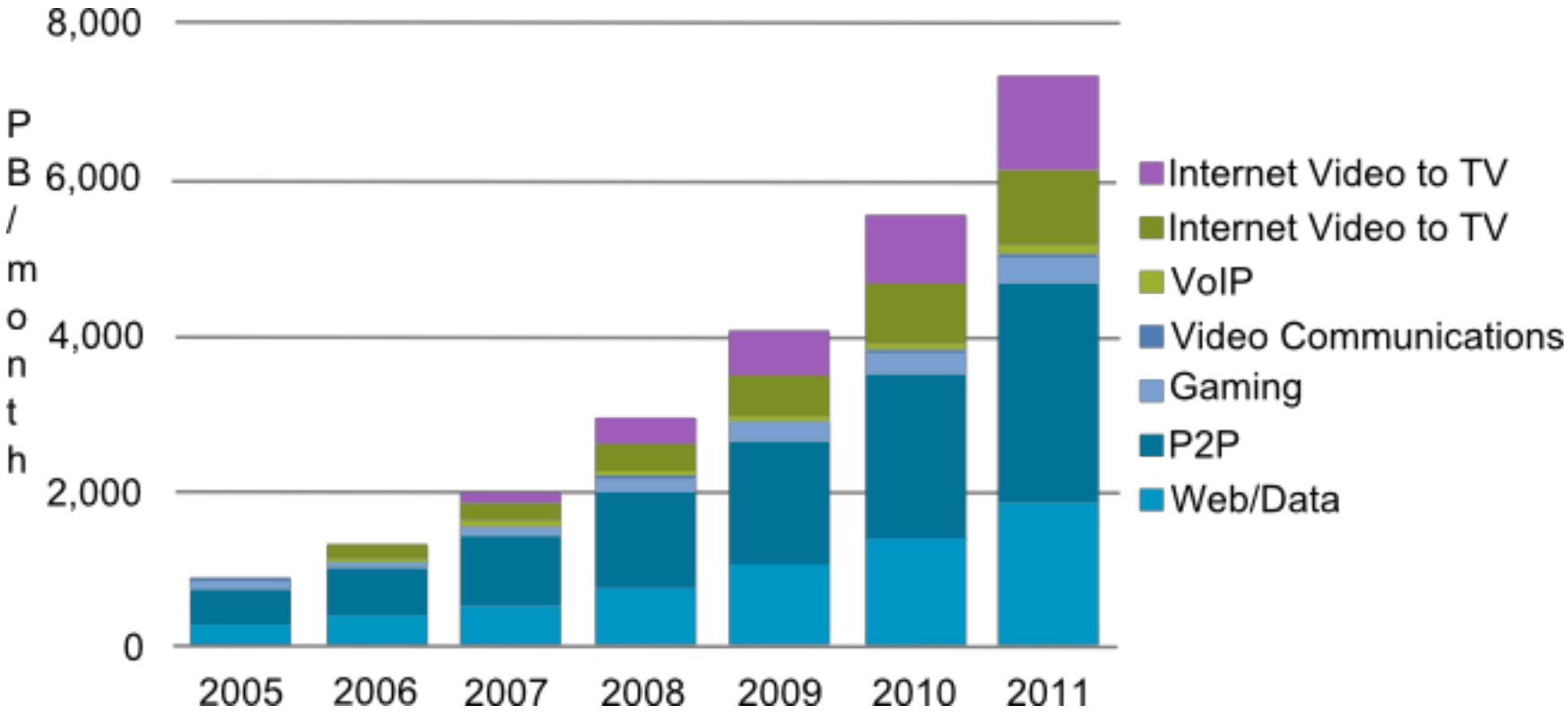
Hands On: Multimedia Methods for Large Scale Video Analysis

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

Multimedia in the Internet is Growing



Raphaël Troncy: *Linked Media: Weaving non-textual content into the Semantic Web*, MozCamp, 03/2009.

Multimedia in the Internet is Growing

- YouTube claims ~~65k~~ 100k video uploads per day or ~~48~~ 72 hours every minute
- Flickr claims 1M images uploads per day
- Twitter: up to 120M messages per day

Why do we care?

- Consumer-Produced Multimedia allows empirical studies at never-before seen scale in <u>various</u> research disciplines such as sociology, medicine, economics, environmental sciences, computer science...

- Recent buzzword: BIGDATA

Problem

How can YOU effectively work on large scale multimedia data (without working at Google)?

What is this class about?

- Introduction to large-scale video processing from a CS point of view (application driven)
- Covers different modalities: Visual, Audio, Tags, Sensor Data
- Covers different side topics, such as Mechanical Turk for annotation

Content of this Class

- Visual methods for video analysis
- Acoustic methods for video analysis
- Meta-data and tag-based methods for video analysis
- Inferring from the social graph and collaborative filtering
- Information fusion and multimodal integration
- Coping with memory and computational issues
- Crowd sourcing for ground truth annotation
- Privacy issues and societal impact of video retrieval

Why you should you be interested in the Topic

- Processing of large scale, consumer produced videos is a brand-new emerging field driven by:
  - Massive government funding (e.g. IARPA Aladdin, IARPA Finder, NSF BIGDATA)
  - Industry's needs to make consumer videos retrievable and searchable
  - Interesting academic questions

Some Examples of Research at ICSI

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

Video2GPS: Multimodal Location Estimation of Flickr Videos

Gerald Friedland
International Computer Science Institute
Berkeley, CA
fractor@icsi.berkeley.edu

Definition

- Location Estimation = estimating the geo-coordinates of the origin of the content recorded in digital media.
- Here: Regression task
  - Where was the media recorded in lattitude, longitude, and altitude?

G. Friedland, O. Vinyals, and T. Darrell: "Multimodal Location Estimation", pp. 1245-1251, ACM Multimedia, Florence, Italy, October 2010.

Motivation 1

Training data comes for free on a never-before seen scale!

Portal	%	Total
YouTube (estimate)	3.0	3M
Flickr	4.5	180M

Allows tackling of tasks of never-before seens difficulty?

Motivation 2

Location-based services are awesome!

Geocoordinates+Time = Unique ID

De-Motivation

Geo-Location enables Cybercasing!

Bits

Business ▪ Innovation ▪ Technology ▪ Society

September 12, 2010, 10:24 AM

Burglars Picked Houses Based on Facebook Updates

By NICK BILTON

News Feed Top News · Most Recent 63

Share: Status Question Photo Link Video

Going to the beach for the weekend! (Someone else will be home though so think again Facebook Bandits!)

Share

Illustration by Nick Bilton/The New York Times

G. Friedland and R. Sommer: "Cybercasing the Joint: On the Privacy Implications of Geotagging", Proceedings of the Fifth USENIX Workshop on Hot Topics in Security (HotSec 10), Washington, D.C, August 2010.

Placing Task

Automatically guess the location of a Flickr video:

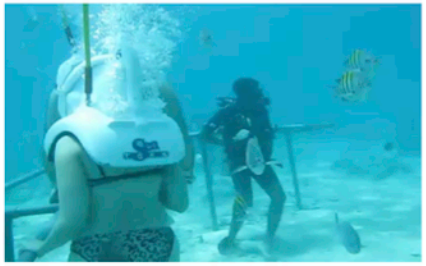
- i.e., assign geo-coordinates (latitude and longitude)
- Using one or more of:
  - Visual/Audio content
  - Metadata (title, tags, description, etc)
  - Social information

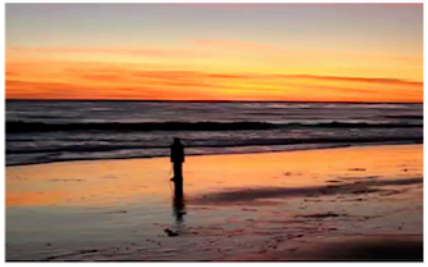
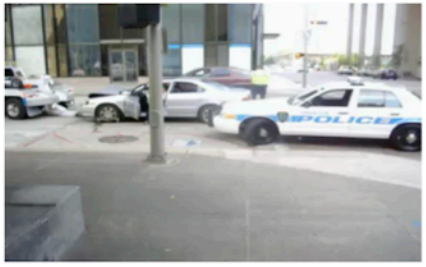
INTERNATIONAL COMPUTER SCIENCE INSTITUTE

Consumer-Produced, Unfiltered Videos...








Data Description (2011)

- Training Data
  - 10k flickr videos/metadata/ visual keyframes (+features)/ geo-tags
  - 6M flickr photos/metadata/ visual features
- Test Data
  - 5k flickr video/metadata/visual keyframes (+features)
  - no geotags
- Test/Training split: by UserID

Your Best Guess???

Our Approach

Tag-based Approach
+ Visual Approach
+ Acoustic Approach

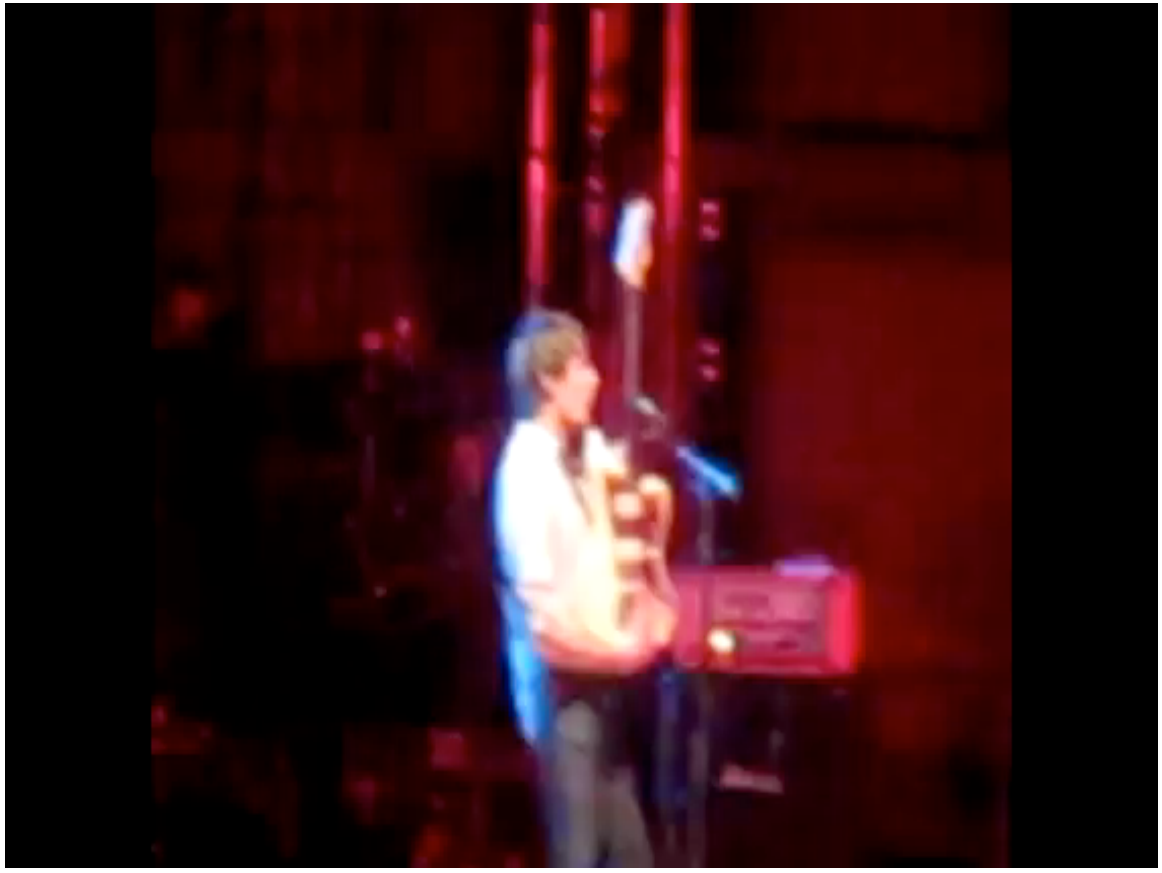
Multimodal Approach

Intuition for the Approach

Investigate `Spatial Variance' of a feature:

- Spatial variance is small: Features is likely location-indicative
- Spatial variance is high: Feature is likely not indicative.

Example Revisited



Tag: ['pavement', 'ucberkeley', 'berkeley', 'greek', 'greektheatre', 'spitonastranger', 'live', 'video‘]

Tag-based Approach

Tag	Matches in Training set	Spatial Variance
pavement	2	5.739
ucberkeley	4	0.132
berkeley	14	68.138
greek	0	N/A
greektheatre	0	N/A
spitonastranger	0	N/A
live	91	6453.109
video	2967	6735.844

- 'greektheatre' would be the correct answer: however, it was not seen in the development dataset
- 'ucberkeley' is the 2nd-best answer: estimation error : 0.332 km

Visual Approach

- Assumption: similar images → similar locations
- Based off related work: Reduce location estimation to an image retrieval problem
- Use median frame of video as keyframe

Visual Features

- Color Histograms – L*a*b* (4, 14, 14 bins, respectively)
  - Chi-square distance
- GIST – 5x5 spatial resolution, 6 orientations and 4 scales
  - Euclidean distance

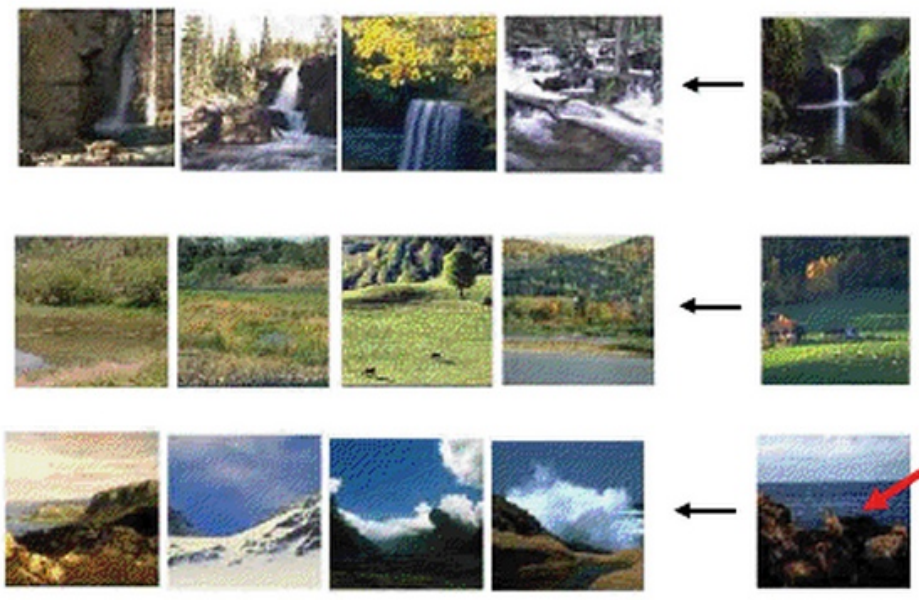
See: A. Oliva, and A. Torralba: Building the Gist of a Scene: The Role of Global Image Features in Recognition, Progress in Brain Research Volume: 155, Issue: 1, Publisher: Elsevier, Pages: 23-36, 2006

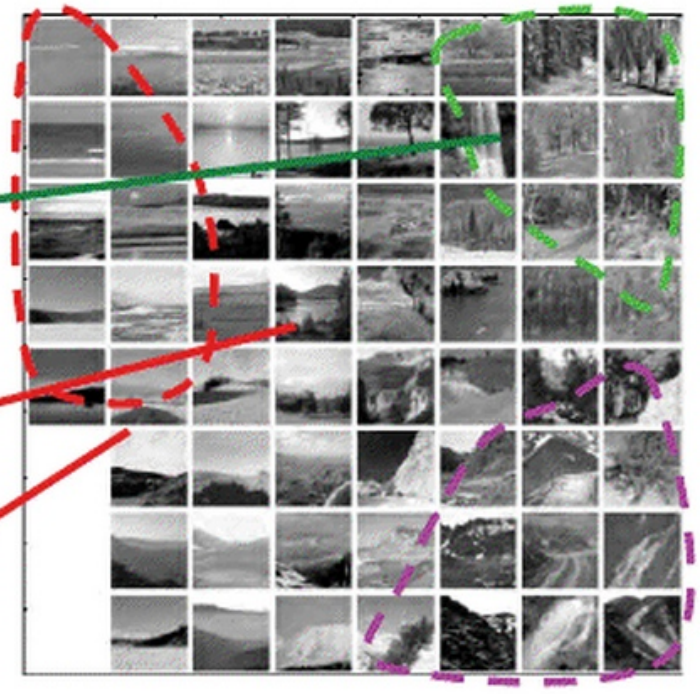
Categorization of Natural Scenes

Confusion Matrix (in % using Layout template) :
Classification of prototypical scenes (400 / category)

	Coast	Countryside	Forest	Mountain
Coast	88.6	8.9	1.2	1.3
Countryside	9.8	85.2	3.7	1.3
Forest	0.4	3.6	91.5	4.5
Mountain	0.4	4.6	3.8	91.2

Local organization: correct for 92 % images
(4 similar images on 7 K-NN)





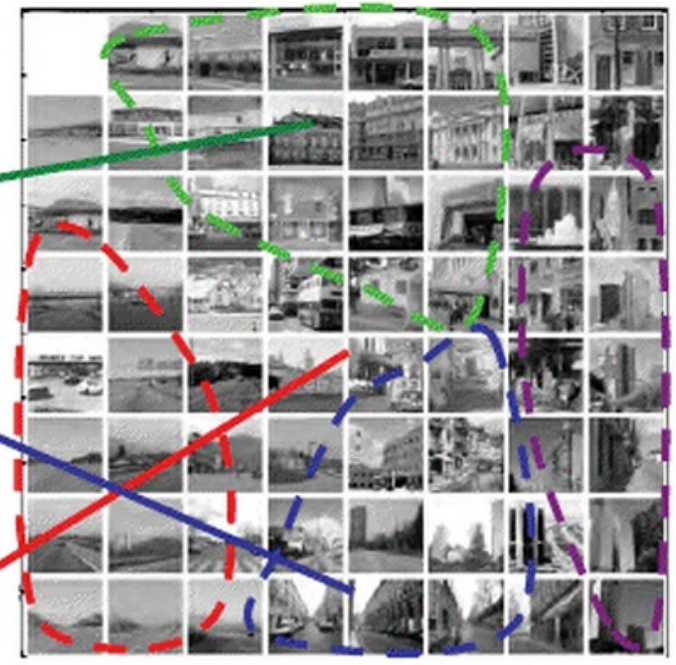
Slide Credit: Olivia

Categorization of Manmade Scenes

Confusion Matrix (in % using Layout template) :
Classification of prototypical scenes (400 / category)

	Highway	Street	City centre	tall building
Highway	91.6	4.8	2.7	0.9
Street	4.7	89.6	1.8	3.4
Centre	2.5	2.3	87.8	7.4
Tall Building	0.1	3.4	8.5	88

Local organization: correct for 86 % images
(4 similar images on 7 K-NN)



Slide Credit: Olivia

Audio Approach

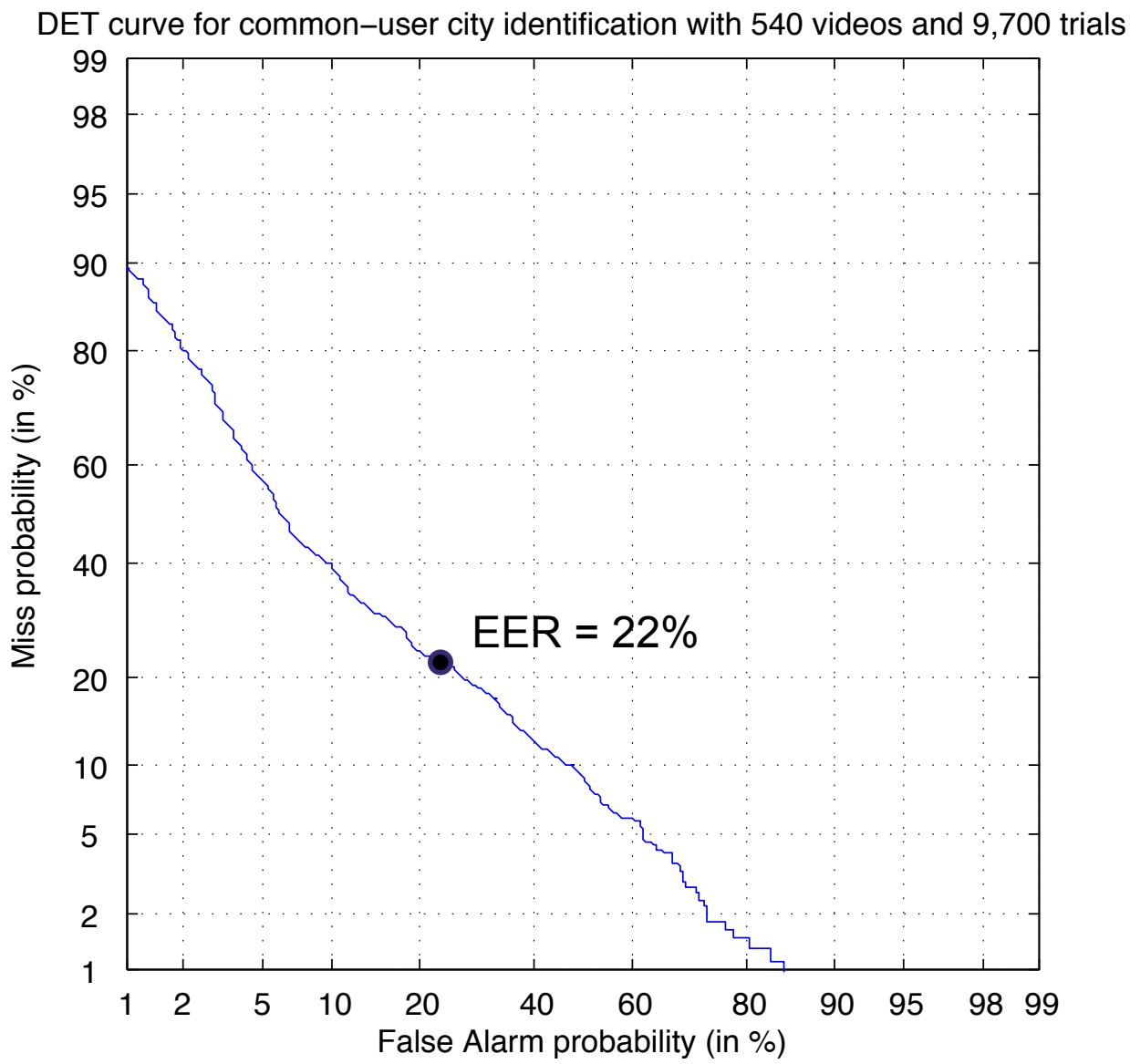
- Idea: Different places have different acoustic “signatures”
- Due to data sparsity: Initially only used for major cities
- Classification system similar to GMM/UBM Speaker ID system using MFCC19 features.

An Experiment

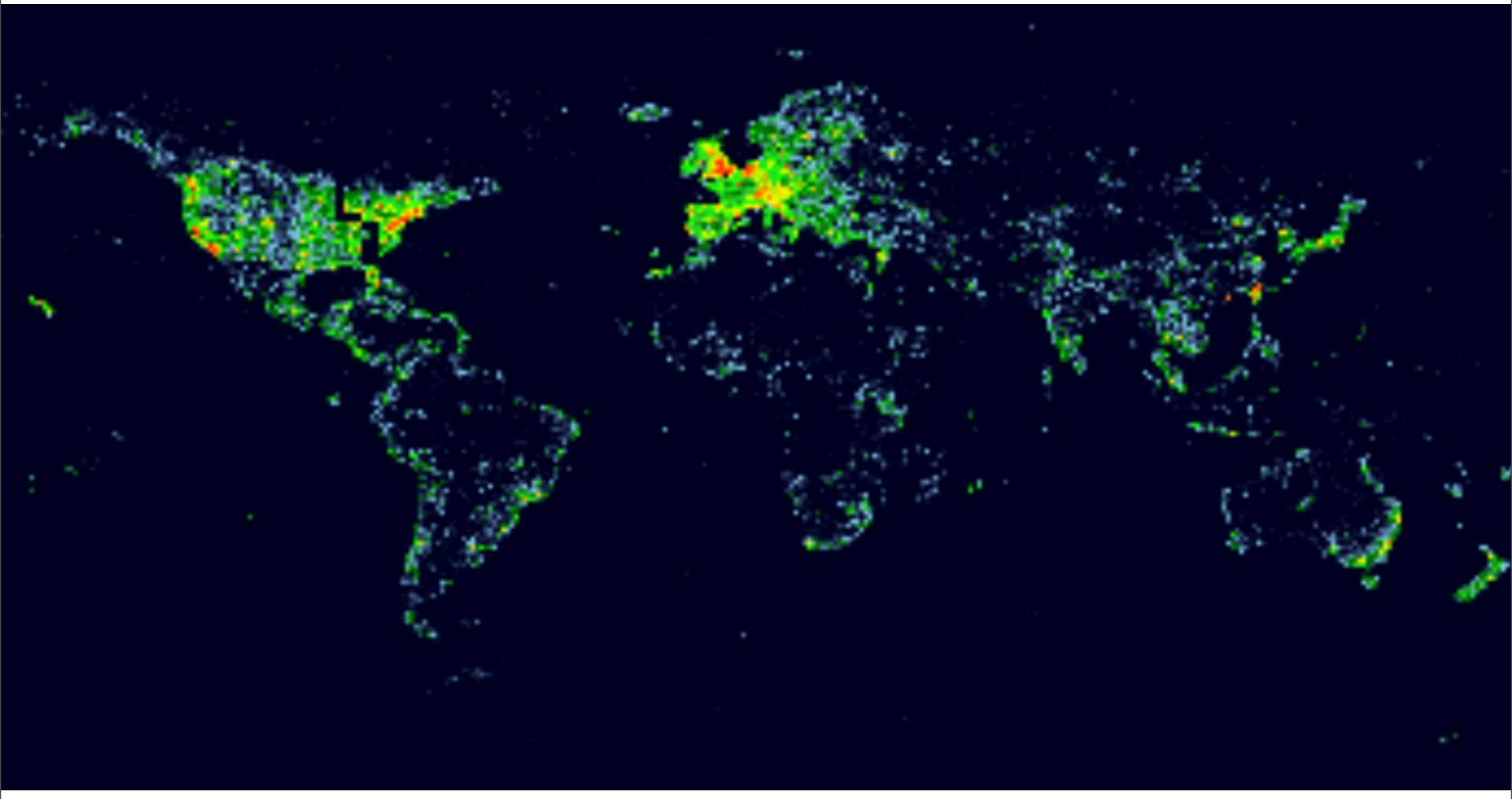
Listen!

- 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!

Audio-Only Results



Problem: Bias



Geo-tagging: an estimation-theoretic

Observations:

Images:




Tags: {berkeley, sathergate, campanile}, {berkeley, haas}, {campanile}, {campanile, haas}

$$\left\{ \{t_1^k\}, \{t_2^k\}, \{t_3^k\}, \{t_4^k\} \right\}$$

Estimate:

Geo locations: $\left\{ x_1, x_2, x_3, x_4 \right\}$

Interpreting traditional approaches

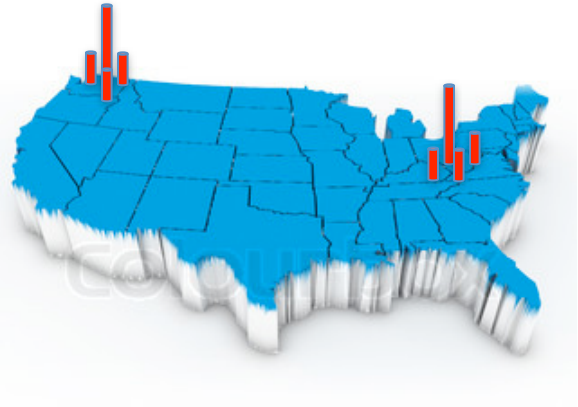
Locations are random variables: $\{x_1, x_2, \ldots\ldots, x_N\}$

Probability of location given tags

Traditional approaches estimate: $p(x_i|\{t_i^k\}) \approx \prod_k p(x_i|t_i^k)$

where $p(x_i|t_i^k)$ is obtained from the training set

Example: the distribution for the tag "washington" is depicted here



Location estimate: $\int x_i \; p(x_i|\{t_i^k\}) dx_i$

Drawbacks

Data sparsity:

Not all tags in test set are available in training set.
Hence estimate of $p(x_i|t_i^k)$ can be bad

Sub-optimality:

The approaches are suboptimal given the data.

What we ideally want: $p(x_1, x_2, ..., x_N|\{t_1^k\}, \{t_2^k\}, ..., \{t_N^k\})$

Mean of the above distribution gives the best estimate of the locations

i.e. for each image we want $p(x_i|\{t_1^k\}, \{t_2^k\}, ..., \{t_N^k\})$

Traditional algorithms only give: $p(x_i|\{t_i^k\})$

Cooperative geo-tagging

Intuition: Images in the training set having common tags have correlated geo-locations captured by the joint distribution

Joint probability modeling:

$$p(x_1, x_2, ..., x_N | \{t_1^k\}, \{t_2^k\}, ..., \{t_N^k\}) \propto \prod_i p(x_i | \{t_i^k\}) \prod_{(i,j)} p(x_i, x_j | \{t_i^k\} \cap \{t_j^k\})$$

Pairwise distribution given at least one common tag

$p(x_i | \{t_i^k\})$ is obtained from the training set as before

$p(x_i, x_j | \{t_i^k\} \cap \{t_j^k\})$ Modeled as an indicator function $\mathbb{I}(x_i = x_j)$

If the common tag has low spatial variance or occurs infrequently, e.g. if the common tag is “haas”, its very likely the locations are the same

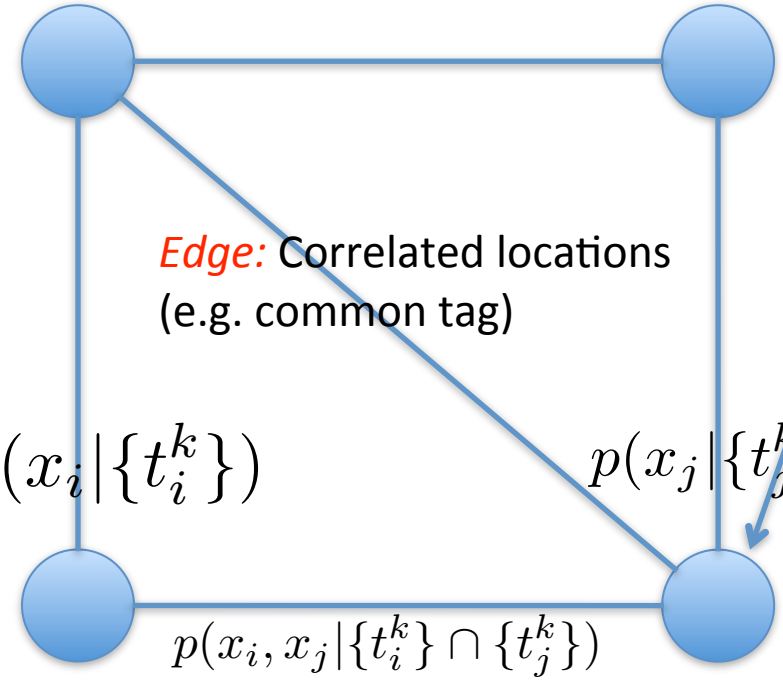
Question: How to estimate to optimal marginal distribution ?

$$p(x_i | \{t_1^k\}, \{t_2^k\}, ..., \{t_N^k\})$$

Bayesian graphical framework

{berkeley, sathergate, campanile}

{berkeley, haas}



Edge: Correlated locations (e.g. common tag)

Node: Geolocation of the image

{campanile}

{campanile, haas}

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

Belief propagation updates

Iterative algorithm to approximate the posterior distribution

$$p(x_i|\{t_1^k\},\{t_2^k\},....,\{t_N^k\})$$

Gaussian modeling

$$p(x_i|\{t_i^k\}) \approx \mathcal{N}(\mu_i, \sigma_i^2)$$

At iteration 0 each node calculates

$$(\mu_i, \sigma_i^2)$$

At iteration t each node updates its location as a weighted mean of its previous location and that of its neighbors

$$\mu_i^{(t)} = \frac{\frac{1}{(\sigma_i^{(t)})^2}\mu_i^{(t-1)} + \sum_{k\in\mathcal{N}(i)} \frac{1}{(\sigma_k^{(t)})^2}\mu_k^{(t)}}{(\sigma_i^{(t)})^2}$$

$$\frac{1}{(\sigma_i^{(t)})^2} = \frac{1}{(\sigma_i^{(t-1)})^2} + \sum_{k\in\rangle} \frac{1}{(\sigma_k^{(t-1)})^2}$$

The weights reflect the confidence in that measurements, i.e. higher the spatial variance lower is the weight

Belief propagation

(μ_2, σ_2^2)

(μ_3, σ_3^2)

Posterior mean and variance assuming Gaussian beliefs

(μ_1, σ_1^2)

Audio visual features are incorporated in modeling the edge and node potentials

Incorporating Audio-Visual features

- GIST features are extracted for the images.
- MFCC features are extracted for the audio.
- These are now incorporated into the node and edge potentials as exponential distributions.

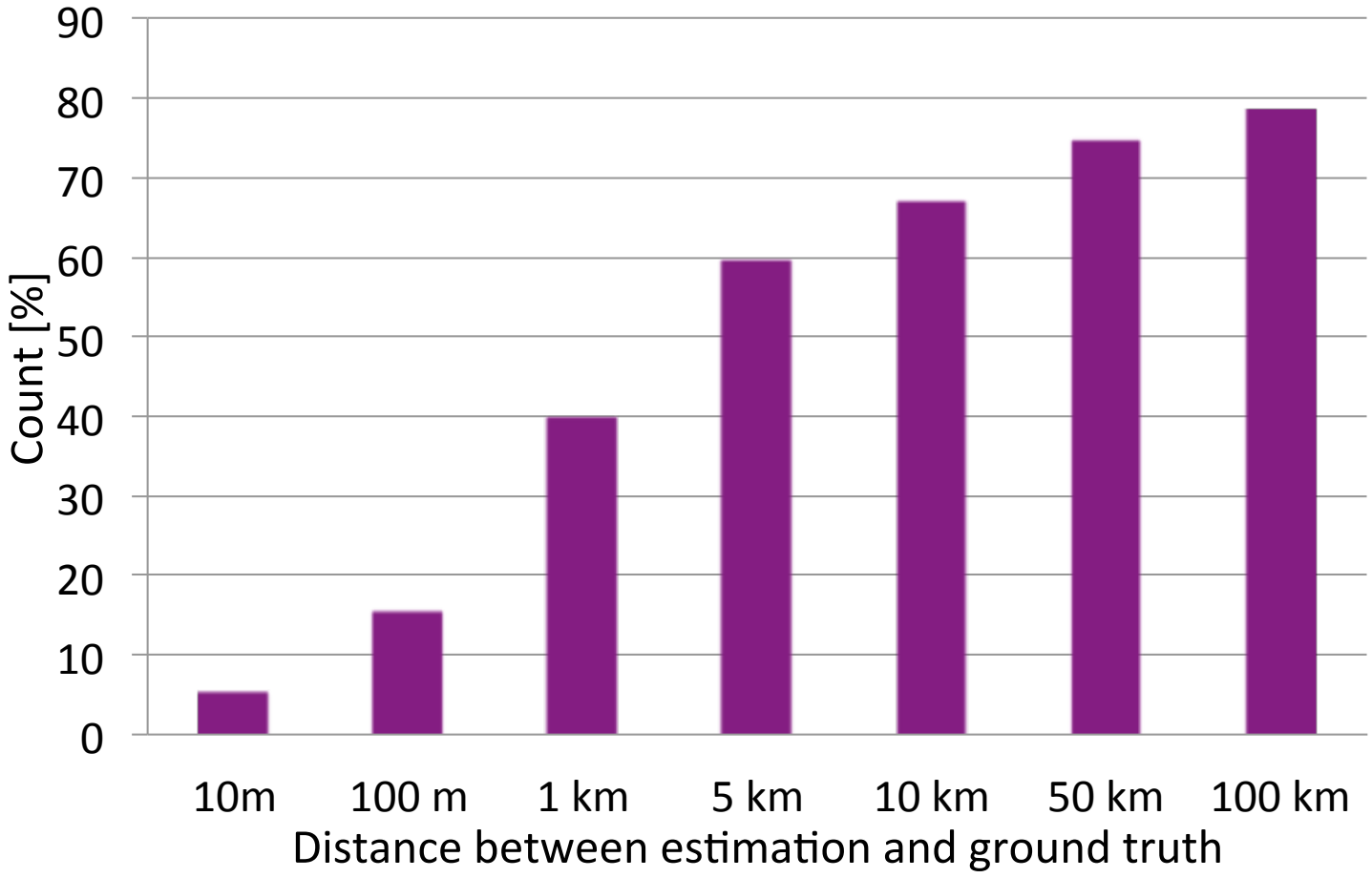
$$p(x_i, x_j | a_i, a_j) \propto \exp(-\frac{||x_i - x_j||}{\lambda ||a_i - a_j||})$$

a_i are the audio features associated with image i

The intuition is that closer the audio features are, higher the probability that the geo-locations are closer.

Similarly this can be included in the node potentials as well as for the visual features.

Multimodal Results: MediaEval 2011 Dataset



How is the Class Organized?

- Once a week: lectures fundamental introduction (from me)
- Once a week: Project meetings
- Guest lectures by YouTube, Yahoo!, Intel, TU Delft.

Hands-On Part

- Do a project on a large scale data set
- Data accessible through ICSI accounts
- Additional computation: $100 in Amazon EC 2 money

Lecture Material

- Background material for lectures: G. Friedland, R. Jain: Introduction to Multimedia Computing, to appear at Cambridge University Press, 2012.

- (Constantly changing) draft available http://www.mm-creole.org

How do you receive Credit?

- 3 Credits (graded or ungraded) based on:
  - 2 quizzes
  - a project