Hands On: Multimedia Methods for Large Scale Video Analysis

Dr. Gerald Friedland,
fractor@icsi.berkeley.edu
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 various 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 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 latitude, longitude, and altitude?

Motivation 1

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

<table>
<thead>
<tr>
<th>Portal</th>
<th>%</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>YouTube</td>
<td>3.0</td>
<td>3M</td>
</tr>
<tr>
<td>Flickr</td>
<td>4.5</td>
<td>180M</td>
</tr>
</tbody>
</table>

Allows tackling of tasks of never-before seen difficulty?
Motivation 2

Location-based services are awesome!

Geocoordinates + Time = Unique ID
De-Motivation

Geo-Location enables Cybercasing!

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
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

<table>
<thead>
<tr>
<th>Tag</th>
<th>Matches in Training set</th>
<th>Spatial Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>pavement</td>
<td>2</td>
<td>5.739</td>
</tr>
<tr>
<td>ucberkeley</td>
<td>4</td>
<td>0.132</td>
</tr>
<tr>
<td>berkeley</td>
<td>14</td>
<td>68.138</td>
</tr>
<tr>
<td>greek</td>
<td>0</td>
<td>N/A</td>
</tr>
<tr>
<td>greektheatre</td>
<td>0</td>
<td>N/A</td>
</tr>
<tr>
<td>spitonastranger</td>
<td>0</td>
<td>N/A</td>
</tr>
<tr>
<td>live</td>
<td>91</td>
<td>6453.109</td>
</tr>
<tr>
<td>video</td>
<td>2967</td>
<td>6735.844</td>
</tr>
</tbody>
</table>

- ‘greektheatre’ would be the correct answer: however, it was not seen in the development dataset
- ‘ucberkeley’ is the 2\textsuperscript{nd}–best answer: estimation error : 0.332 km
Visual Approach

• Assumption: similar images $\rightarrow$ 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

Categorization of Natural Scenes

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

<table>
<thead>
<tr>
<th></th>
<th>Coast</th>
<th>Countryside</th>
<th>Forest</th>
<th>Mountain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coast</td>
<td>88.6</td>
<td>8.9</td>
<td>1.2</td>
<td>1.3</td>
</tr>
<tr>
<td>Countryside</td>
<td>9.8</td>
<td>85.2</td>
<td>3.7</td>
<td>1.3</td>
</tr>
<tr>
<td>Forest</td>
<td>0.4</td>
<td>3.6</td>
<td>91.5</td>
<td>4.5</td>
</tr>
<tr>
<td>Mountain</td>
<td>0.4</td>
<td>4.6</td>
<td>3.8</td>
<td>91.2</td>
</tr>
</tbody>
</table>

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)

<table>
<thead>
<tr>
<th></th>
<th>Highway</th>
<th>Street</th>
<th>City Centre</th>
<th>Tall Building</th>
</tr>
</thead>
<tbody>
<tr>
<td>Highway</td>
<td>91.6</td>
<td>4.8</td>
<td>2.7</td>
<td>0.9</td>
</tr>
<tr>
<td>Street</td>
<td>4.7</td>
<td>89.6</td>
<td>1.8</td>
<td>3.4</td>
</tr>
<tr>
<td>Centre</td>
<td>2.5</td>
<td>2.3</td>
<td>87.8</td>
<td>7.4</td>
</tr>
<tr>
<td>Tall Building</td>
<td>0.1</td>
<td>3.4</td>
<td>8.5</td>
<td>88</td>
</tr>
</tbody>
</table>

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.
• Which city was this recorded in?


• Solution: Tokyo, highest confidence score!
Audio-Only Results

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

EER = 22%
Problem: Bias
Geo-tagging: an estimation-theoretic

Observations:

Images:

Tags:

Estimate:

Geo locations:
Interpreting traditional approaches

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

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: $p(x_1, x_2, \ldots, x_N | \{t_1^k\}, \{t_2^k\}, \ldots, \{t_N^k\})$

Mean of the above distribution gives the best estimate of the locations i.e. for each image we want $p(x_i | \{t_1^k\}, \{t_2^k\}, \ldots, \{t_N^k\})$

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:

\[
p(x_1, x_2, \ldots, x_N | \{t_k^1\}, \{t_k^2\}, \ldots, \{t_k^N\}) \propto \prod_i p(x_i | \{t_i^k\}) \prod_{(i,j)} p(x_i, x_j | \{t_i^k\} \cap \{t_j^k\})
\]

Pairwise distribution given at least one common tag

\[
p(x_i | \{t_i^k\}) \quad \text{is obtained from the training set as before}
\]

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

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?

\[
p(x_i | \{t_1^k\}, \{t_2^k\}, \ldots, \{t_N^k\})
\]
Bayesian graphical framework

*Node:* Geolocation of the image

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

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

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

\[
p(x \mid \{ k \})
\]
Belief propagation updates

Iterative algorithm to approximate the posterior distribution

Gaussian modeling

At iteration 0 each node calculates

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

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

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

Posterior mean and variance assuming Gaussian beliefs

$(\mu_1, \sigma_1^2)$

$(\mu_2, \sigma_2^2)$

$(\mu_3, \sigma_3^2)$
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.

\[ p(x_i, x_j | a_i, a_j) \propto \exp \left( -\frac{||x_i - x_j||}{\lambda ||a_i - a_j||} \right) \]

\( a_i \) 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.
The results shown are superior in accuracy than any system described in Section 5. The results are visualized in Figure 3.

For analyzing the algorithm in greater detail, here we also compared by finding how many videos were placed within a threshold distance of 1 km, 5 km, 10 km, 50 km and 100 km. To minimize the distances over all test videos, runs were performed for videos using one or more of: video metadata (tags, titles), visual content, audio content, and social information. Even videos using these attributes can be estimated correctly with an accuracy better than 100 km. The multimodal approach estimates 3989 correctly of the videos uploaded by one user is low on average. Therefore taking into account to which user a video belongs seems to have a higher chance of finding geographical distances of below 100 m and below 10 m. The lowest accuracy in the finer-granularity ranges but increases over the all the images in a 100 km radius around the 1 to 3 coordinates from the previous step. We pick the match with the smallest distance and output its coordinates as a final result.

Visual content, audio content, and social information. Even videos using one or more of: video metadata (tags, titles), visual content, audio content, and social information. Even videos using these attributes can be estimated correctly with an accuracy better than 100 km. The multimodal approach estimates 3989 correctly of the videos uploaded by one user is low on average. Therefore taking into account to which user a video belongs seems to have a higher chance of finding geographical distances of below 100 m and below 10 m. The lowest accuracy in the finer-granularity ranges but increases over the all the images in a 100 km radius around the 1 to 3 coordinates from the previous step. We pick the match with the smallest distance and output its coordinates as a final result.

Multimodal Results: MediaEval 2011 Dataset

![Graph showing the comparison between Visual Only, Tags Only, and Visual+Tags multimodal location estimation algorithms.](Image)
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


• (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