Human vs Machine: Establishing a Human Baseline for Multimodal Location Estimation

Jaeyoung Choi[1], Howard Lei[1], Venkatesan Ekambaram[2], Pascal Kelm[3], Luke Gottlieb[1], Thomas Sikora[3], Kannan Ramchandran[2] and Gerald Friedland[1]

[1]International Computer Science Institute, Berkeley, CA, USA
[2]Department of Electrical Engineering and Computer Sciences, UC Berkeley, Berkeley, CA, USA
[3]Communication Systems Group, Technische Universität Berlin, Germany

[1]{jaeyoung,hlei,luke,fractor}@icsi.berkeley.edu
[2]{venkyne,kannanr}@eecs.berkeley.edu
[3]{kelm,sikora}@nue.tu-berlin.de

ABSTRACT

Over the recent years, the problem of video location estimation (i.e., estimating the longitude/latitude coordinates of a video without GPS information) has been approached with diverse methods and ideas in the research community and significant improvements have been made. So far, however, systems have only been compared against each other and no systematic study on human performance has been conducted. Based on a human-subject study with 11,900 experiments, this article presents a human baseline for location estimation for different combinations of modalities (audio, audio/video, audio/video/text). Furthermore, this article compares state-of-the-art location estimation systems with the human baseline. Although the overall performance of humans' multimodal video location estimation is better than current machine learning approaches, the difference is quite small: For 41 % of the test set, the machine's accuracy was superior to the humans. We present case studies and discuss why machines did better for some videos and not for others. Our analysis suggests new directions and priorities for future work on the improvement of location inference algorithms.

Categories and Subject Descriptors

H.3.3 [**Information Systems and Retrieval**]: Retrieval models

General Terms

EXPERIMENTATION, HUMAN FACTORS

Keywords

Location Estimation, Multimodal, Crowdsourcing

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MM'13, October 21–25, 2013, Barcelona, Spain.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2404-5/13/10 ...$15.00.
http://dx.doi.org/10.1145/2502081.2502123.

1. INTRODUCTION

With the widespread use of GPS-equipped handheld devices, location metadata (aka geo-tagging) has rapidly become an integral part of photos and videos shared over the Web. This trend has enabled location based multimedia organization, search and retrieval on many Internet services such as Google, Facebook, and Flickr. The main driving force behind these services is the creation of highly personalized user experiences, allowing for better recommendations and targeted advertisements.

Even with this trend, it has been estimated that only about 5 % of the existing multimedia content on the Internet is actually geo-tagged [8]. A significant amount of consumer-produced media content is still obtained using devices that do not have GPS functionality. Privacy concerns have motivated users to disable automatic geo-tagging of media. Furthermore, even GPS-enabled devices cannot provide accurate location information when the photo or video is captured in an indoor environment.

Nevertheless, the volume of high quality geo-tagged videos and photos on the Web represent a quantity of training data for machine learning on an unprecedented scale, giving rise to the idea of creating an automated task that would try to locate non-geo-tagged media from the web using models obtained through the geo-tagged subset. Put simply: Given a video and its associated textual metadata, can we infer the location where it was taken? This idea of "multimodal video location estimation", was proposed three years ago as a Brave New Idea in [8]. Since then, the "Placing Task" of the MediaEval evaluation [22] has evaluated the task on a global scale [19]. Therefore, the problem has been approached with diverse methods and ideas in the research community and significant improvements have been made. So far, however, approaches have only been compared against each other and there is little intuition on how humans would perform at this task. Some researchers even assume that the automated algorithms would probably always perform better on this task compared to humans. In this paper, we establish a human baseline for video location estimation and present a comparative analysis with automatic location inference systems. The baseline was created by asking qualified humans to perform a total of 9,000 video localizations. With the human baseline in our hand, we are able to analyze different cases of when machines perform better than humans, humans perform better than machines or when both fail.

Our paper is organized as follows. Section 2 provides a brief overview of the existing work in the field and positions our work in comparison to the available literature. Section 3 describes the task and the characteristics of the dataset that render the task difficult. Section 4 describes the experimental setup for establishing the human baseline using a crowdsourcing platform. Section 5 describes our technical approaches to utilizing audio, audio/visual, and audio/visual/textual metadata for automatic location inference. Section 6 provides a comparison of the performance of location estimation between machines and humans. In section 7, we present case studies and discuss why machines perform better for some videos and not for others. Section 8 concludes with a summary of the paper and future research directions based on our analysis.

2. RELATED WORK

Initial approaches to location estimation started several years ago. In earlier articles [28, 33], the location estimation task is reduced to a retrieval problem on self-produced, location-tagged image databases. The idea is that if the image is the same then the location must be the same too. Hays and Efros [12] estimate a rough location of an image with several visual descriptors and represent the estimated image location as a probability distribution over earth's surface. A comprehensive overview of this early research work is presented in [21].

Previous work in the area of automatic geotagging of multimedia based on tags has mostly been carried out on Flickr images. In [26], the geo-locations associated with specific Flickr tags are predicted using spatial distributions of tag use. A tag which is strongly concentrated in a specific location has a semantic relationship with that location. User-contributed tags are exploited for geotagging by [29], who used tag distributions associated with locations as represented by grid cells on a map of the earth which is then used to infer the geographic locations of Flickr images.

Evaluations on multimodal location estimation on randomly selected consumer-produced videos were carried out in the 2010, 2011, and 2012 MediaEval Placing tasks [22]. One of the notable participants [18] used a combination of language models and similarity search to geo-tag the videos using their associated tags. Many participants in the Placing task try to utilize both visual and textual features for their location estimations. Friedland et al. [7] propose a hierarchical system that first uses spatial variance of tags to find initial estimates which are used as anchor points to set the search boundary for visual nearest neighbor search in the later stage. Kelm et al. [16] propose another hierarchical, multi-modal approach which first classifies a query video's location into possible regions and then applies a visual nearest neighbor method to find corresponding training images in those regions. Crandall et al. [6] propose a model that produces enhanced geographical annotation for web images using visual and tag features and an annotation lexicon. The approaches in [1] and [9] report on combining visual content with user tags as well.

Multimodal location estimation on videos that utilize audio was first attempted in [8] in which the authors matched videos containing audio from ambulance sirens in different cities, without the use of textual tags. Other works such as [5] reports the results of a location estimation system that incorporates all three modalities that are usable in video location estimation, i.e. textual, visual, and audio features.

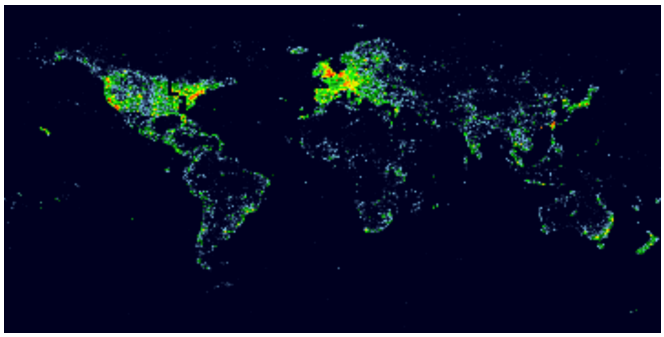

Figure 1: Distribution of the videos and images of the MediaEval 2010 Placing Task training dataset. Jet colormap was applied to show the density.

All of the approaches described above have the common feature of processing each query video independently and estimating its geo-location based on textual, visual, and audio features using a geo-tagged training database. Clearly, the performance of these systems largely depends on the size and quality of the training database. However, data sparsity is one of the major issues that can adversely affect the performance of these systems. The work presented in [4] differs from the existing work in the literature in the respect that they *jointly* estimate the geo-locations of all of the input query images. Each query image added to the database enhances the quality of the database by acting as "virtual" training data and thereby boosts the performance of the algorithm. The system presented in this paper is an improvement over that work as the system presented here uses all three modalities (visual, audio, and textual features) whereas [4] only uses textual features collected from user-annotated metadata.

Crowdsourcing is currently used for a range of applications such as exploiting unsolicited user contributions, for spontaneous annotation of images for retrieval [27], etc. Systematic crowdsourcing platforms, such as Amazon Mechanical Turk have been used to mass-outsource artificial intelligence jobs [14]. Further, crowdsourcing has also been used for surveying and evaluating user interfaces [17], designs, and other technical approaches.

However, as the name coincidentally implies, platforms such as Mechanical Turk are often best used for mechanical tasks, i.e. tasks that only require simple intuition. Therefore, for a task like the one presented here, where there is a suspicion that humans might perform worse than machines and there is no clear intuition as to how to solve the task, one has to be very careful about how to approach it properly. Apart from [10], there is no previous work on using Mechanical Turk for geo-tagging videos, moreover there seems to be no previous work on how to use Mechanical Turk for a task that is not straightforward to solve.

So far, no human baseline exists for location estimation, and systems have only been compared against each other. At the same time, these systems differ quite dramatically. This paper establishes a human baseline and compares it against two state of the art systems from the literature to draw conclusions for future directions of this research. To ensure the quality of the human baseline, we rely on the qualification methodology described in [10] in combination with redundancy [15].

Figure 2: Several frames from the randomly selected videos that were used in the experiment. Most of them are very difficult to specifically geolocate.

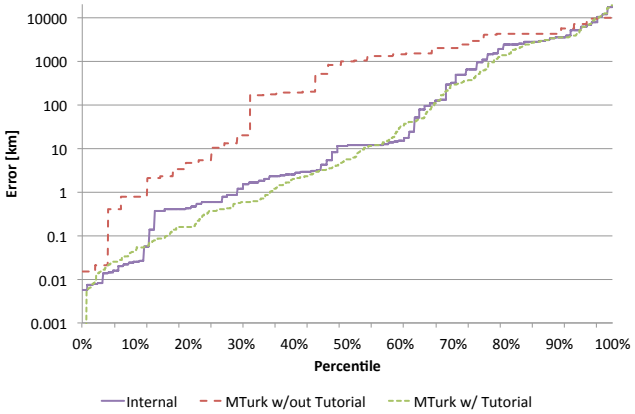

Figure 3: Comparison of performance results for Ideal10 set [10]. Error distance (y-axis) is in a log scale.

3. TASK AND DATASET

All experiments described in this article were performed using the dataset distributed for the Placing Task of the 2010 MediaEval benchmark[1]. The Placing Task is part of the MediaEval benchmarking initiative that requires participants to assign geographical coordinates (latitude and longitude) to each test video. Participants can make use of textual metadata, audio and visual features as well as external resources, depending on the test run.

The dataset consists of 3,185,258 photos and 10216 videos. All are Creative Commons-licensed and from Flickr[2]. The metadata for each video includes a user-annotated title, tags, and a description among others. The videos are not filtered or selected in any way to make the dataset more relevant to the task, and are therefore likely to be representative of videos selected at random [19]. Figure 1 shows the non-uniform distribution of Flickr videos and images due to geographical, economical, and political reasons.

Flickr requires that an uploaded video must be created by the uploader, and thus almost all videos on Flickr are home-video style. The relatively short lengths of each video should be noted, as the maximum length of a Flickr video was limited to 90 seconds when the dataset was collected. Moreover, about 70 % of videos in this data set have less than 50 seconds of playtime. Manual inspection of the randomly sampled 150 videos from the dataset shows that if given with only the audio and visual contents, 8% of the videos contained enough information for accurate guesses, and 10% with rough hints that would lead to city or country level estimations. The rest of the videos had very little cue for location estimation. Videos of indoor settings consist 36% of the videos and about half of them are private space such as one's house or in the backyard. 24% of the videos contained human speech in various languages including English, Spanish, Swedish, Japanese, etc. The metadata provided by the user often provided direct and sensible clues for location estimation. 98.8 % of videos in the training dataset were annotated by their uploaders with at least one title, tag, or description, which often included location information.

[1]http://multimediaeval.org/
[2]http://www.flickr.com

4. ESTABLISHING A HUMAN BASELINE

Collection of human baseline was performed in two steps. In the first step we qualified people, by filtering out incompetent or unmotivated workers and ensuring that the quality of submissions were high enough for the second stage. In the second stage, we collected the human baseline for 1000 videos. We used Amazon Mechanical Turk as the crowdsourcing platform.

4.1 Qualification

The task of location estimation is different from a standard Mechanical Turk task in that it is difficult for both humans and machines, whereas a standard Mechanical Turk task is usually easy for humans and difficult or impossible for machines. There are several notable challenges to finding skilled workers for this task: First, we must find what we term "honest operators" i.e., people who will seriously attempt to do the task and not just click quickly through it to collect the bounty. Second, we need to develop a meaningful qualification test set that is challenging enough to allow us to qualify people for the real task, but is also solvable by individuals regardless of their culture or location, although English language understanding was required for instructions. For example, in the process of selecting videos, there were videos of tourists in Machu Picchu, which our annotator immediately recognized, however there were no clues to reveal this location that would be usable to someone who had not heard or seen this location previously. These videos were ruled out for the qualification. In the end, ten videos, which we called 'Ideal10 set', were carefully chosen and presented to the workers. We created an in-depth tutorial which presented the workers with the basic tools and skills for approaching this challenging task. The workers are allowed to use any applicable resource from the Internet, including Google Maps and Streetview.

Our previous study show that, after the qualification process, workers on the crowdsourcing platform achieved almost equal level of accuracy as an internal expert volunteer

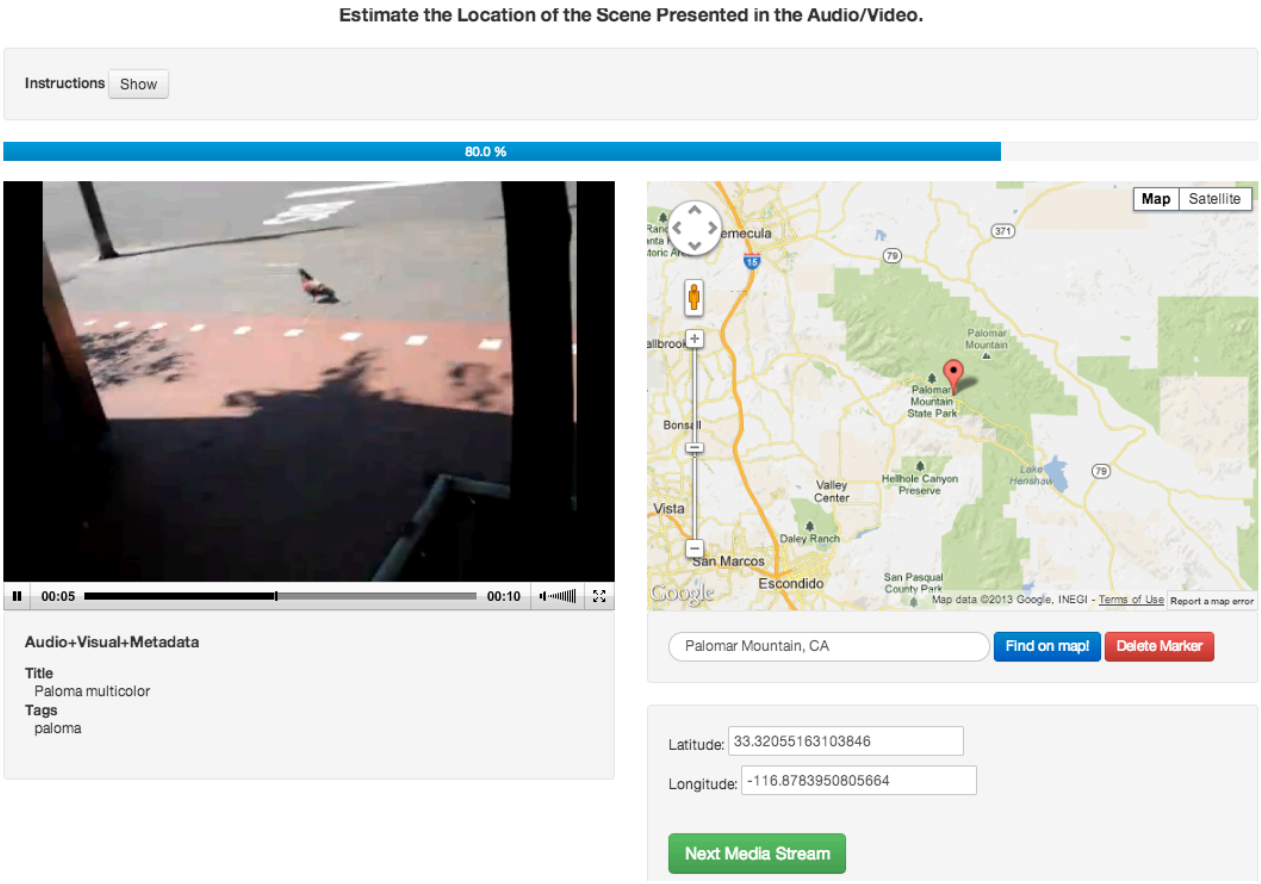

Figure 4: Screenshot of web interface used in the Amazon Mechanical Turk experiments described.

testers which was composed of highly-educated, well-trained and motivated researchers. In Figure 3, we have a comparison of the performance results for our internal testers, the initial test results of MTurkers, and their results with the tutorial. The internal tester and initial test results were derived by using the classification results from the first round for the videos in the Ideal10 set, as those videos are a subset of our larger annotation. We can see that while the internal testers still perform better than the Mechanical Turk workers, the addition of a tutorial greatly narrowed the performance margin. This led us to conclude that a worker who is adequately trained with a tutorial can be considered reasonably qualified.

However, to select only the highest quality workers, we applied a screening process to filter out not only bad but even workers who performed averagely well. Out of 290 workers on Mechanical Turk, who had participated in the qualification task, we qualified only 84 workers (29% acceptance rate), those who were able to achieve very high accuracy i.e., were able to put at least 8 of 10 qualification videos within a 5km radius of the ground truth location. We considered 5km margin of error as a city-level accuracy. When applied with the same evaluation criteria, internal volunteers showed similar acceptance rate of 34%. Additionally, time spent for each human video localization was recorded and often directly correlated with accuracy, giving us a further accuracy indicator for the actual localization test. After successful qualification, we paid USD 1.50 for each HIT (Human Intelligence Task), which consisted of 10 video localizations.

4.2 The Web Interface

Here, we describe our user interface which the Amazon Mechanical Turk workers used for the task. We went through several rounds of internal testing and feedback to enhance the usability of the tool.

Figure 4 shows the final version of this interface. The instructions on the top of the screen can be expanded and shrunk with a 'Show/Hide' button. It was shrunk by default to make the whole interface fit in a normal-sized window to minimize unnecessary scrolling of the screen. A progress bar was shown below the instructions box to let workers know where they are along the progress of a HIT. A video was played automatically once the page was loaded. All videos were re-uploaded to a private file server without the metadata so that simply following the link on the player would not reveal any additional information about the video. A Google Maps instance was placed to the right of the video. A marker would be dropped where the map was clicked, and it could be dragged around the map. The marker's position was automatically translated to the latitude and longitude and printed to the 'Latitude' and 'Longitude' boxes. A location search form was placed under the map to aid the search of the location. The form had an auto-completion feature which would help in cases where the worker did not know the exact spelling of the place, etc. At the end of the HIT, we asked participants to leave comments about the HIT. This enabled us to filter out submissions with incidents and other exceptions.

4.3 Collection of Human Intelligence

When we collected the human performance for a different set of modalities, we presented the media with least information to all i.e., first we gave audio then audio/visual, and lastly audio/visual and textual information (tags). For each HIT, a worker was given five videos with three different media combinations, thus the total of 15 media streams. For the same set of media streams, three identical HITs were generated redundantly for the comparison and to filter out the possible bad results. HITs were assigned with first-come first-served basis to the pool of 84 qualified workers. To distribute the HITs to as many as workers and as evenly as possible, we applied some throttle control to limit the number of HITs that an individual worker can accept. We started the reward for each HIT with USD 0.25 but, based on feedback, quickly increased it to USD 1.50, which resulted in a much faster collection result. Within 18 days, we were able to collect a total of 11,900 localizations (2900 from qualification, 9000 from localizing 1000 videos). Many workers have left comments that the task was challenging, especially for the cases when only the audio was given. At the same time, many of them reported the HIT to be fun and we believe that this motivated people to try to submit better results.

5. MACHINE-BASED LOCATION ESTIMATION

In this section, we describe the technical approach and our experimental setup for automatic location inference. We first describe the audio-based approach, then describe how visual feature was added to the modality, and finally, the method that uses all three modalities (audio, visual, and textual metadata).

5.1 Audio-based Location Estimation

We used the city identification system reported in [20] as our machine baseline for location estimation. We describe the main idea of the system as follows. The system involves training a total variability matrix T to model the variability (both city- and channel-related) of the acoustic features of all audio tracks, and using the matrix to obtain a low-dimensional vector characterizing the city of where each audio track was from. Specifically, for each audio file, a vector of first-order statistics M - of the acoustic feature vectors of the audio centered around the means of a GMM world model - is first obtained, and can be decomposed as follows:

$$M = m + T\omega \tag{1}$$

where m is the GMM world model mean vector, and ω are low-dimensional vectors, known as the identity vectors or i-vectors.

The system then performs Probabilistic Linear Discriminant Analysis (pLDA) [13] and Within-Class Covariance Normalization (WCCN) [11] on the i-vectors. pLDA linearly projects the i-vectors ω onto a set of dimensions to maximize the ratio of between-user scatter to within-user scatter of the i-vectors, producing a new set of vectors. WCCN then whitens the pLDA-projected vectors via a second linear projection, such that the resulting vectors have an identity covariance matrix. For our city identification system, 1,024 mixtures are used for the GMM world model, and a rank of 400 is used for the total variability matrix T, such that the i-vectors ω have 400 dimensions. pLDA projects the i-vectors onto a set of 200 dimensions. The cosine distance is used to obtain the city-similarity score of a pair of i-vectors ω between two audio tracks of user-uploaded videos [30]:

$$score(\omega_1, \omega_2) = \frac{(A^T\omega_1)^T W^{-1}(A^T\omega_2)}{\sqrt{(A^T\omega_1)^T W^{-1}(A^T\omega_1)}\sqrt{(A^T\omega_2)^T W^{-1}(A^T\omega_2)}} \tag{2}$$

where A and W are the LDA and WCCN projection matrices respectively, and ω_1 and ω_2 are i-vectors from the two audio tracks being compared against. The acoustic features consist of MFCC C0-C19+Δ+$\Delta\Delta$ coefficients of 60 dimensions, computed using 25 ms windows and 10 ms shifts, across 60 to 16,000 Hz.

Since we could not train the model to cover all regions of the earth due to the data sparsity, we clustered the distribution of videos into the 40 cities in the training dataset, and reduced the location estimation to a city-identification problem. We trained the system with models for each city using the collection of audio tracks extracted from each city. A video is defined to belong to a city when it is in a 50 km radius of the geographical city center. We then tested the audio tracks extracted from the test videos against the trained models and picked the city with the highest likelihood. For comparability, we converted the city labels to the (latitude, longitude) format with the geo-coordinates of the center of the city. Note, that this creates a slight disadvantage for the machine.

5.2 Visual Location Estimation

In order to utilize the visual content of the video for location estimation, we reduce location estimation to an image retrieval problem, assuming that similar images mean similar locations as in [12]. We used several visual descriptors extracted from sample frames of both query and training videos along with the images given as the Placing Task dataset and ran a k-nearest neighbor search on the training dataset to find the video frame or an image that is most similar. We used FCTH (Fuzzy Color and Texture Histogram) [3], CEDD (Color Edge Directivity Descriptor) [2], and Tamura [31] visual descriptors that were given as a part of the Placing Task dataset. In addition to these descriptors, we extracted Gist features [24] as it was shown to be very effective at scene recognition in [12]. Weighted linear combination of distances was used as the final distance between frames. The scaling of the weights was learned by using a small sample of the training dataset and normalizing the individual distance distributions so that each the standard deviation of each of them would be similar. We used L^2 norm to compare the combination of descriptors and used 1-nearest neighbor matching between the closest pre-extracted frame to the temporal mid-point of a query video and all photos and frames from the videos in the training dataset. In order to handle the large amounts of development data efficiently, we split the reference data set into chunks of 100,000 images, ran 1-NN in parallel on each subset to get intermediate results, and ran 1-NN once again on the intermediate results to get the final nearest neighbor. We used an approximate nearest neighbor library [23] for the experiment.

While videos with soundtrack were shown to the crowd, due to data sparsity, our comparison system did not use the acoustic modality and relied solely on the visual modality.

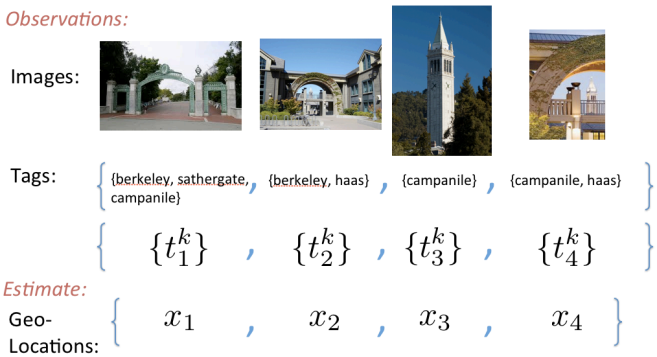

Figure 5: A theoretic viewpoint of the geo-tagging problem.

5.3 Multimodal Location Estimation

To integrate textual and visual data, we combined the visual search method with the system reported in [4]. Due to data sparsity, again, while the soundtrack was available to humans, the acoustic modality was not used. The approach based on graphical models is summarized as follows.

Graphical models provide an efficient representation of dependencies amongst different random variables and have been extensively studied in the statistical learning theory community [32]. The random variables in our setup are the geo-locations of the query videos that need to be estimated. We treat the textual tags as observed random variables that are probabilistically related to the geo-location of that video. Figure 5 illustrates the idea. The goal is to obtain the best estimate of the unobserved random variables (locations of the query videos) given all the observed variables. We use graphical models to characterize the dependencies amongst the different random variables and use efficient message-passing algorithms to obtain the desired estimates. In order to obtain a graphical model representation for our problem setup, we model the joint distribution of the query video locations given the observed data. We use a simplistic conditional dependency model for the random variables as described below. Each node in our graphical model corresponds to a query video and the associated random variable is the geo-location of that query video. Intuitively, if two images are nearby, then they should be connected by an edge since their locations are highly correlated. The problem is that we do not know the geo-locations a priori. However, given that textual tags are strongly correlated to the geo-locations, a common textual tag between two images is a good indication of the proximity of geo-locations. Hence, we will build the graphical model by having an edge between two nodes if and only if the two query videos have at least one common textual tag. Note that this textual tag need not appear in the training dataset.

Let x_i be the geo-location of the ith video and $\{t_i^k\}_{k=1}^{n_i}$ be the set of n_i tags associated with this video. Based on our model the joint probability distribution factorizes as follows:

$$p(x_1, ..., x_N|\{t_1^k\}, ..., \{t_N^k\}) \propto \prod_{i \in V} \psi(x_i|\{t_i^k\}) \prod_{(i,j) \in E} \psi(x_i, x_j|\{t_i^k\}, \{t_j^k\}).$$

We now need to model the node and edge potential functions. Given the training data, we fit a Gaussian Mixture Model (GMM) for the distribution of the location given a particular tag t, i.e., $p(x|t)$. The intuition is that tags usually correspond to one or more specific locations and the distribution is multi-modal (e.g., the tag "washington" can refer to the State of Washington or Washington D.C., among other locations). To estimate the parameters of the GMM, we use an algorithm based on Expectation Maximization that adaptively chooses the number of components for different tags using a likelihood criterion. Although distribution of the locations given multiple tags is not independent, for this experiment, we start with a naive assumption that different tags are conditionally independent. We take the node potential as follows, $\psi(x_i) \propto \prod_{k=1}^{n_i} p(x_i|t_i^k)$. For the potential functions, $\psi(x_i, x_j|\{t_i^k\}, \{t_j^k\})$, we use a very simple model. Intuitively, if the common tag between two query videos i and j occurs too frequently either in the test set or the training set, that tag is most likely a common word like "video" or "photo" which does not really encode any information about the geographic closeness of the two videos. In this case, we assume that the edge potential is zero (drop edge (i, j)) whenever the number of occurrences of the tag is above a threshold. When the occurrence of the common tag is less frequent, then it is most likely that the geographic locations are very close to each other and we model the potential function as an indicator function,

$$\psi(x_i, x_j|\{t_i^k\}, \{t_j^k\}) = \begin{cases} 1 & \text{if } x_i = x_j, \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

This model is a hard-threshold model and we can clearly use a soft-version wherein the weights on the edges for the potential functions are appropriately chosen.

Further, we propose the following simplification, which leads to analytically tractable expressions for the potential functions and message updates. Given that for many of the tags, the GMM will have one strong mixture component, the distribution $\psi(x_i)$, can be approximated by a Gaussian distribution with the mean $(\tilde{\mu}_i)$ and variance $(\tilde{\sigma}_i^2)$ given by,

$$(\tilde{\mu}_i, \tilde{\sigma}_i^2) = \left(\frac{\sum_{k=1}^{n_i} \frac{1}{\sigma_i^{k2}} \mu_i^k}{\sum_{k=1}^{n_i} \frac{1}{\sigma_i^{k2}}}, \frac{1}{\sum_{k=1}^{n_i} \frac{1}{\sigma_i^{k2}}} \right), \quad (4)$$

where μ_i^k and σ_i^{k2} are the mean and variance of the mixture component with the largest weight of the distribution $p(x_i|t_i^k)$. Under this assumption, the iterations of the sum-product algorithm take on the following simplistic form. Node i at iteration m, updates its location estimate $(\hat{\mu}_i(m))$ and variance $(\hat{\sigma}_i^2(m))$ as follows,

$$\hat{\mu}_i(m) = \frac{\frac{1}{\tilde{\sigma}_i^2}\tilde{\mu}_i + \sum_{j \in N(i)} \frac{1}{\hat{\sigma}_j^2(m-1)} \hat{\mu}_j(m-1)}{\frac{1}{\tilde{\sigma}_i^2} + \sum_{j \in N(i)} \frac{1}{\hat{\sigma}_j^2(m-1)}}, \quad (5)$$

$$\hat{\sigma}_i^2(m) = \frac{1}{\frac{1}{\tilde{\sigma}_i^2} + \sum_{j \in N(i)} \frac{1}{\hat{\sigma}_j^2(m-1)}}. \quad (6)$$

The location estimate for the ith query video $\hat{x}_i$ is taken to be $\hat{\mu}_i(m)$ at the end of m iterations, or when the algorithm

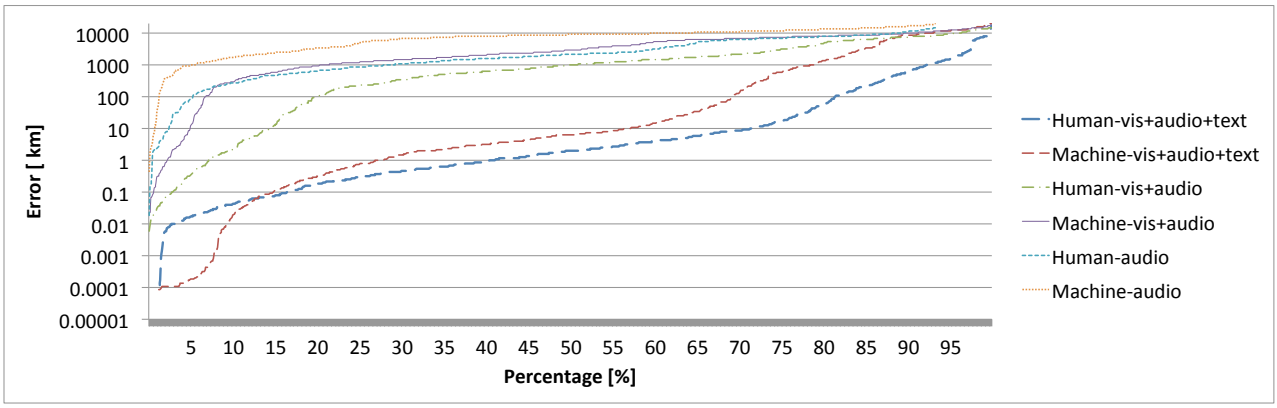

Figure 6: The human baseline of location estimation vs current stat-of-the-art algorithms on all modalities (audio, audio+visual, and audio+visual+text)

has converged. The variance $\hat{\sigma}_i^2(m)$ provides a confidence metric on the location estimate.

So far, only the textual features are used to estimate the location. We performed the same visual search described in section 5.2 around the location estimate $\hat{\mu}_i(m)$ with the search boundary set dynamically according to the variance $\hat{\sigma}_i^2(m)$. The intuition is that more we are certain about the location estimate from the graphical framework, we search a narrower range for similar images. Thus, if the variance is low, the search boundary would be set low as well, and vice versa. Since the visual search was limited to much smaller number of images and video frames dynamically adjusted from the confidence of the textual-based location estimate, the results were improved over searching naively across the whole training dataset or searching within the fixed range from the text-based location estimate.

All of the algorithms described above, except for the acoustic system that seems to be unique, achieve among the highest scores in the MediaEval 2012 evaluation and we therefore consider them a state-of-the-art machine baseline.

6. RESULTS

To evaluate the performance of both the online workers and the machine, the geodesic distance between the ground truth coordinates and those of the outputs from participants or the machine, respectively, are compared. To take into account the geographic nature of the evaluation, the Haversine [25] distance is used.

Figure 6 shows the performance of humans vs the machine given three different combinations of modalities: audio, audio+visual, and audio+visual+text. Human performance was measured by having 3 different qualified workers redundantly locate the same video. A total number of 1000 geotagged videos were presented in the 3 different forms (as discussed in section 4), resulting in a total number of 9000 experiments. The results of the qualification experiments are not included in this chart. To obtain a conservative baseline for this chart, the best answer out of the 3 was picked. For representing machine performance, we used the system as described in section 5 based on the combination of media tested.

Overall, humans are better at this task than machines. However, in the very accurate (below 50 m region) current algorithms outperform humans by about a 12 % margin (number of correctly located videos). The difference between the algorithm and human intelligence is quite low though, since overall only 59 % of the videos are located more accurately by the human than by the algorithm when all modalities are utilized. An unqualified worker base would most certainly have resulted in machine dominance of this task. As a control, choosing the second-best of the 3 results, resulted in the humans being less accurate than the algorithm in 60 % of the videos. Both the machine and humans do better once more modalities are available. Using only audio, the machine was better for only 16.4 % of the videos. In the visual case, the machine was only better in 25.5 % of the videos.

We will qualify the results in the following section.

7. DISCUSSION

In this section, we analyze the cases where humans were successful in inferring location while the machine failed and vice versa. We also investigate the videos where humans and computers have both failed to infer location. The threshold for determining the 'success' and 'failure' was set differently for each of the modalities given. We were more generous when only the audio was given, as it is much more difficult than the two other cases.

7.1 Machine vs. Human using only Audio

For audio only experiments, the threshold to be considered a successful estimate was set at 150km, and the failure threshold was set at 1000 km.

7.1.1 When humans are better

Out of 1000 audio-only test videos, there were 62 cases in which the audio provided enough information for humans to infer location while the machines failed. Close analysis of all of these audio tracks revealed that the cases where humans were better than machines can be categorized into the following three classes:

1. Humans were able to identify location based on the kind of language spoken or the distinctive accent or variation of the speakers in the video. With only this information at hand and no other clues, human annotators picked the capital city of the country or region where that language is mainly spoken or was originally from (Paris for French, Glasgow for Scottish accented English, London for British accented English,

Lisbon for Portuguese). We have no way to investigate whether the workers were able to use any additional information from understanding the contents of the speech.

2. Humans were able to pick up keywords from the speech that they were able to understand either entirely or at least in part providing a clue sufficient to estimate the location. For example, a Finnish singer saying, "Hello, Helsinki!" at the opening of a concert was the only understandable portion for people who don't speak Finnish, but this is a sufficient clue for estimating the location.

3. Humans were also able to infer location information from the context of the text spoken. In one video, the speech contained the keywords 'California', 'grapes', 'harvest', and 'wine' which could be inferred to be in the Napa Valley using a Google search or geographical knowledge.

On the machine side, the first category of videos could be localized with an audio-based language or dialect identification system. The second category of videos could be localized as if the textual metadata were given using the keywords extracted from the transcript of the speech obtained from passing the audio track to the automatic speech recognition system. The major challenge would be to deal with the noisy transcript from the 'wild' audio. The third kind of inference is the most difficult.

7.1.2 When machines are better

While humans were in general better at the task using only the soundtrack, there were 10 cases when machines did reasonably well (estimating location under 100km, which is the city-level boundary used for the audio-based approach) while all of the human annotators failed. One notable finding from the analysis was that three of these videos were from Prague, CZ. All three videos were from different users and contained different scenes and events, however, a close inspection of the audio revealed that three of the videos contained a musical noise in the background. Also, two of the training videos used to train the model for the city of Prague had music playing in the background. We believe that our city model from i-vector system picked up the common musical chords that are often played in the touristy locations in Prague.

The system also gave the highest score to Tokyo to a video that shows the Shinkansen train leaving from Tokyo station. Similarly with the above example, we believe that the i-vector system has learned the very specific sound of the Shinkansen train track.

These results are good news for automatic approaches: They show that machine learning approaches can exploit very specific sounds that are hard to spot for humans and use them for location matching.

7.1.3 When both fail

When a video is edited and the audio track is altered such as when its dubbed with background music, or if there's no audible speech, both humans and machines usually failed. However, human workers did tend to converge on certain locations for the audio tracks that can be inferred to be a stereotype of a broader category of locations such as "beach", "fireworks", or "farm". For example, for beaches, all three human annotators picked a beach in Los Angeles, CA when the audio track contains the sound of sea gull and the sound of breaking waves, whereas the ground truth was a beach in Liverpool, UK. Two of the annotators picked New York for the audio track that contained the sound of fireworks.

This scheme is actually applicable to machine learning as well. For example, the machine could be trained to classify the scene into a broader category to aid in the estimation of location. Classifying the scene of a beach at night could benefit from using audio as the visual features would not work well due to poor lighting conditions.

7.2 Machine vs. Human using Audio and Video

With the added visual feature, both machines and humans were able to get much better results than using audio alone. For these experiments, the threshold to be considered a successful estimation was set at 50 km, and the failure was set at 1000 km. We excluded cases where audio-only had already given a sufficient clue for humans to get below the 50 km error range as we could not independently evaluate the effectiveness of the visual feature. We also investigated cases where the audio and visual features complement each other whereas only one modality would have failed.

7.2.1 When humans are better

In 179 cases humans were better than machines. We could separate out about 4 classes.

1. The majority of cases where human workers perform extremely well (getting under 5 km or even 100 m error) belong to the class that contain textual information in the video. These can be in the form of captions added by the user, signs, or even messages written on buildings or machinery in the video. This category of videos could possibly be located with a video OCR system that extracts textual information.

2. When the visual and audio modalities complement each other but when used separately are less effective, humans are usually better. In other words, multimodal integration in machines is not yet successful. For example, one video showed a TV broadcast of an American Football game with the name of the team and the score shown on the screen. The uploader makes a cheering sound as the game ends and the all three human workers inferred that the the uploader is a resident of the winning team's region, which is true.

3. When the scene contains a famous landmark humans perform very well. However, our location estimation system was not specifically trained to recognize landmarks.

4. The atmosphere and context of the scene can be understood by the human workers. For instance, a video taken from a moving car shows the clothing style of the people on the street, the status of the road, and the shape of buildings. In one particular case, all three human workers inferred from this collected information that this could be a specific rural town in India, which was in fact was the right answer.

We did not find reasonable evidence that temporal information of the visual features impacted the location inference of human workers.

7.2.2 When machines are better

We found 81 cases where the machine was better than the humans using only visual features. Most of these cases consisted of specific tourist spots where the machine had many

training videos of but the locations are not well known to a lot of people. For example, all three human workers failed to recognize the foggy Machu Picchu (mislabeled as pyramids in Mexico) or a mountain scene of Patagonia (mislabeled as Himalaya or Canadian Rockies).

7.2.3 When both fail

Most of the cases where both humans and machines failed was when videos were taken indoors such as the inside of a night club (poor lighting, poor audio), videos of babies in a house, and so on. Other cases were some generic scenes such as an unpopular mountain, outskirts of large cities with no landmarks or signs, etc.

7.3 Machine vs. Human using all Modalities

The threshold for success in this case is 5 km as the textual information is very effective at allowing inference of location for both humans and the machine. The threshold for failure is set at 1000 km. In 39 cases, the machine achieved less than 5 km error while all three qualified human annotators failed to estimate the location with an error of more than 1000 km. For the opposite case where the human does better than the machine, we found 162 cases.

7.3.1 When humans are better

In 162 cases, humans were successful at picking a correct location while the machines failed. Many of the errors were from the system failing to pick up a single keyword that represents the location within the list of tags.

1. We believe the critical advantage in some of these 162 cases was from the misspelling of a tag or that the tag was written in a foreign language (which was not included in the training). Human locators did not have a problems with the misspelled words.

2. The bias in the distribution of the training dataset results from the failure of the system to correctly process keywords if they were not seen in the training dataset. Although our system tries to address the problem of sparsity using the graphical framework, it is still bound by the quantity and quality of the data in both the training and test sets. The use of semantic computing-based approaches as done in [16] can be an effective solution in these cases.

7.3.2 When machines are better

1. Some of the videos contained multiple tags that were not helpful in inferring location but were repeatedly seen in other videos in the training dataset. For example, "iPhone, 3gs, or iphone3gs" does not have a specific meaning related to the location in one of the test videos. However, our system was able to pick up the repeated common appearance of these tags in the training data and was able to estimate the location under 0.5 km error. This is due to similar users using the same "language model" when tagging their videos. Keep in mind that test and training set had different sets of users.

2. Humans failed to pick up clues from combination of words when too many tags were given, whereas the machine was able to implicitly incorporate n-gram using the graphical framework.

3. Language barrier: In some cases the tags were written in a foreign language (not in the training dataset). Worker populations on Amazon Mechanical Turk are mostly English-based, thus the presence of non-English tags presents a language barrier. Some human workers managed to get over this by using translation services such as Google Translate.

7.3.3 When both fail

There were 26 cases where both the machine as well as all humans failed to get a location even with all possible sources of information present. This is expected because sometimes there are just no useful cues to estimate location. For example, a scene which is a closeup with no distinguishable sounds, and no textual description to indicate location. In the end, we were surprised though that only 2.5 % of all videos fell into this category (where the location was not estimated with under 1000 km accuracy by either the human or the machine). This indicates a high growth potential for future location estimation research.

8. CONCLUSION AND FUTURE WORK

In this article, we establish a human baseline for multimodal location estimation of random consumer-produced videos with textual descriptions. Even though algorithms work on low-level statistics, we show that humans outperform the algorithm sometimes and in other cases the algorithm outperforms humans. The difference between human performance and algorithmic performance is so close that we speculate that in a relatively short time, algorithms will become better than the human baseline. Surprisingly enough only about 2.5 % of the videos could not be located at all. This suggests a huge potential for future research in the field even though for some of the videos, the algorithm would already pass the "Turing test" as it was already better than humans for 41 % of the videos. The analysis of human vs machine errors suggests complementarity which implies future work might be very successful when concentrating on interactive systems. For example, for humans the acoustic modality works quite well when language and speech content can be picked up. Machines are better at picking up specific sounds that might be the fingerprint of a location. Similarly, in the visual domain, humans are very good at localization based on written text, signs, architecture, and vehicle styles while machines are very good at finding specific locations that appear often enough in the training data. Humans can remember or search for specific locations based on context, while machines can pick up patterns of tags or textual descriptions that indicate the location of a specific social group unknown to most humans. In summary, all three modalities (audio, video, and text) are very powerful at determining location both for humans and the machine even though recent researches mostly have concentrated on text-based systems with the aid of visual information. Future research on location estimation might gain improved results from including optical character recognition and sign interpretation as well as language identification and language-independent keyword spotting. The use of semantic information from Gazetteers and other knowledge bases will be very helpful for locations with limited training data.

9. ACKNOWLEDGMENTS

This research is supported in part by NSF EAGER grant IIS-1128599 and KFAS Doctoral Study Abroad Fellowship. Human subjects experiments are authorized under IRB approval CPHS 2011-06-3325.

10. REFERENCES

[1] L. Cao, J. Yu, J. Luo, and T. Huang. Enhancing semantic and geographic annotation of web images via logistic canonical correlation regression. In *Proceedings of the 17th ACM international conference on Multimedia*, MM '09, pages 125–134, New York, NY, USA, 2009. ACM.

[2] S. Chatzichristofis and Y. Boutalis. Cedd: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval. *Computer Vision Systems*, pages 312–322, 2008.

[3] S. Chatzichristofis and Y. Boutalis. Fcth: Fuzzy color and texture histogram-a low level feature for accurate image retrieval. In *Image Analysis for Multimedia Interactive Services, 2008. WIAMIS'08. Ninth International Workshop on*, pages 191–196. Ieee, 2008.

[4] J. Choi, G. Friedland, V. Ekambaram, and K. Ramchandran. Multimodal location estimation of consumer media: Dealing with sparse training data. In *2012 IEEE International Conference on Multimedia and Expo (ICME)*, pages 43–48. IEEE, 2012.

[5] J. Choi, H. Lei, and G. Friedland. The 2011 ICSI Video Location Estimation System. In *Proc. of MediaEval*, 2011.

[6] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In *Proc. of WWW '09*, pages 761–770, New York, NY, USA, 2009. ACM.

[7] G. Friedland, J. Choi, H. Lei, and A. Janin. Multimodal Location Estimation on Flickr Videos. In *Proc. of the 2011 ACM Workshop on Social Media*, pages 23–28, Scottsdale, Arizona, USA, 2011. ACM.

[8] G. Friedland, O. Vinyals, and T. Darrell. Multimodal Location Estimation. In *Proceedings of ACM Multimedia*, pages 1245–1251, 2010.

[9] A. Gallagher, D. Joshi, J. Yu, and J. Luo. Geo-location inference from image content and user tags. In *Proceedings of IEEE CVPR*. IEEE, 2009.

[10] L. Gottlieb, J. Choi, G. Friedland, P. Kelm, and T. Sikora. Pushing the Limits of Mechanical Turk: Qualifying the Crowd for Video Geo-Location. *Proceedings of the 2012 ACM Workshop on Crowdsourcing for Multimedia (CrowdMM)*, 2012.

[11] A. Hatch, S. Kajarekar, and A. Stolcke. Within-class covariance normalization for SVM-based speaker recognition. In *Proceedings of ISCA Interspeech*, volume 4, 2006.

[12] J. Hays and A. Efros. IM2GPS: estimating geographic information from a single image. In *IEEE CVPR 2008*, pages 1–8, 2008.

[13] S. Ioffe. Probabilistic linear discriminant analysis. *Computer Vision–ECCV 2006*, pages 531–542, 2006.

[14] P. G. Ipeirotis. Analyzing the amazon mechanical turk marketplace. *XRDS*, 17(2):16–21, Dec. 2010.

[15] D. Karger, S. Oh, and D. Shah. Budget-optimal crowdsourcing using low-rank matrix approximations. In *Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on*, pages 284 –291, sept. 2011.

[16] P. Kelm, S. Schmiedeke, and T. Sikora. A hierarchical, multi-modal approach for placing videos on the map using millions of flickr photographs. In *Proc. of SBNMA '11*, pages 15–20, New York, NY, USA, 2011. ACM.

[17] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, CHI '08, pages 453–456, New York, NY, USA, 2008. ACM.

[18] O. V. Laere, S. Schockaert, and B. Dhoedt. Ghent university at the 2011 placing task. In *Proc. of MediaEval*, 2011.

[19] M. Larson, M. Soleymani, P. Serdyukov, S. Rudinac, C. Wartena, V. Murdock, G. Friedland, R. Ordelman, and G. J. Jones. Automatic Tagging and Geo-Tagging in Video Collections and Communities. In *ACM International Conference on Multimedia Retrieval (ICMR 2011)*, pages 51:1–51:8, April 2011.

[20] H. Lei, J. Choi, and G. Friedland. City-Identification on Flickr Videos Using Acoustic Features. Technical report, ICSI Technical Report TR-11-001, 2011.

[21] J. Luo, D. Joshi, J. Yu, and A. Gallagher. Geotagging in multimedia and computer vision–a survey. *Multimedia Tools Appl.*, 51:187–211, Jan. 2011.

[22] Mediaeval web site. *http://www.multimediaeval.org*.

[23] D. M. Mount and S. Arya. ANN: A library for approximate nearest neighbor searching. In *CGC 2nd Annual Fall Workshop on Computational Geometry*, pages 153–, 1997.

[24] A. Oliva and A. Torralba. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155:23–36, 2006.

[25] M. C. Palmer. Calculation of distance traveled by fishing vessels using gps positional data: A theoretical evaluation of the sources of error. *Fisheries Research*, 89(1):57–64, 2008.

[26] T. Rattenbury and M. Naaman. Methods for extracting place semantics from Flickr tags. *ACM Transactions on the Web (TWEB)*, 3(1), 2009.

[27] B. Russell, A. Torralba, K. Murphy, and W. Freeman. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77:157–173, 2008. 10.1007/s11263-007-0090-8.

[28] G. Schindler, M. Brown, and R. Szeliski. City-scale location recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07*, pages 1–7, 2007.

[29] P. Serdyukov, V. Murdock, and R. van Zwol. Placing Flickr photos on a map. In *ACM SIGIR*, pages 484–491, 2009.

[30] M. Soufifar, M. Kockmann, L. Burget, O. Plchot, O. Glembek, and T. Svendsen. iVector approach to phonotactic language recognition. In *Proc. of Interspeech*, pages 2913–2916, 2011.

[31] H. Tamura, S. Mori, and T. Yamawaki. Textural features corresponding to visual perception. *Systems, Man and Cybernetics, IEEE Transactions on*, 8(6):460–473, 1978.

[32] M. Wainwright and M. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1:1–305, 2008.

[33] W. Zhang and J. Kosecka. Image based localization in urban environments. In *3D Data Processing, Visualization, and Transmission, 3rd Intl. Symposium on*, pages 33–40, 2006.