A Multimodal Approach to Automatic Geo-Tagging of Video

Jaeyoung Choi

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2012-109
http://www.eecs.berkeley.edu/Pubs/TechRpts/2012/EECS-2012-109.html

May 11, 2012

A Multimodal Approach to Automatic Geo-Tagging of Video

by Jaeyoung Choi

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

Committee:

Professor Nelson Morgan
Research Advisor

(Date)

* * * * * * *

Professor Kannan Ramchandran
Second Reader

(Date)

* * * * * * *

Dr. Gerald Friedland
Third Reader

(Date)

Abstract

Geo-tags provide an essential support for organizing and retrieving the rapidly growing online video contents captured by users and shared online. Videos present an unique opportunity for automatic geo-tagging as they combine multiple information sources, i.e., textual metadata, visual and audio cues. This report highlights various approaches (data-driven, semantic technology-based, and graphical model-based) to predict the geo-location of online videos. The algorithms make use of each or combinations of textual, visual and audio information sources. All experiments were performed with a geo-coordinate prediction benchmarking corpus containing 10,438 videos. The performance of these algorithm is analyzed, revealing that the textual metadata is particularly more useful than visual or audio contents, but the combination of multiple cues shows better overall performance. The report concludes with a discussion of the impact that the improvement of geo-coordinate prediction will have and the challenges that remain open for future research.

Contents

List of Figures

Acknowledgements

I would like to thank my mentor, Gerald Friedland, who has given motivation, advice, encouragement, and unfailing support throughout my graduate career. He has been very inspirational and changed my view on research. I don't think I can thank him enough.

I'm deeply grateful to my official advisor, Nelson Morgan, who provided me a research home at ICSI. His inputs and feedbacks throughout this project and my graduate years have been most helpful.

I would also like to thank everybody at ICSI who directly or indirectly helped and supported me the last two years, in particular, Trevor Darrell, Adam Janin, Howard Lei, Arlo Faria, Oriol Vinyals, Mary Knox, Lara Stoll, Dan Gillick, Suman Ravuri, Luke Gottlieb, TJ Tsai, Shuo-Yiin Chang, and Michael Ellsworth. I was privileged to be collaborating with gifted minds of Berkeley BASICS Group, Venky Ekambaram, Giulia Fanti and Kannan Ramchandran.

I'm indebted to Jiwon Kim, Choongbum Lee, Jennifer Jinju Lee, Eugene Suh, Michelle Jeung-Eun Lee, and Jun Woo Lim, who were there for me when I needed them the most while going through crises during my graduate years. It would have been a real boring life here if it weren't for dear friends that I've met at Berkeley: Sunghwan Kim, Sun Choi, Kevin Ahn, Ilhyung Lee, Insoon Yang, Shinhye Choi, and Hyojung Shin. All the good and bad moments we've shared and went through together won't be forgotten. Korean folks at CS department have given me precious advices in many aspects. I'm grateful for the time I got to spent with Gunho Lee, Minkyung Kang, Chang-seo Park, and Yunsup Lee.

Lastly, I thank my parents and my sister for their tireless support and encouragement. They have always supported me in everything I do, and for this I want to thank them most of all.

Thank you all.

Chapter 1

Introduction

With the emergence of Web 2.0 and with GPS devices becoming ubiquitous and pervasive in our daily life, location-based services are rapidly gaining traction in the online world. The main driving force behind these services is the enabling of a very personalized experience. Social-media websites such as Flickr, YouTube, Twitter, etc., allow queries for results originating at a certain location. Likewise, the belief is that retro-fitting archives with location information will be attractive to many businesses, and will enable newer applications. Geo-tagging multimedia content has various applications. For example, geo-location services can be provided for media captured in environments without GPS, such as photos taken indoors on mobile phones. Vacation videos and photos can be better organized and presented to the user if they have geo-location information. With the explosive growth of available multimedia content on the Internet (200 million photos are uploaded to Facebook daily), there is a dire need for efficient organization and retrieval of multimedia content, which can be enabled by geo-tagging. Geo-location information further helps develop a better semantic understanding of multimedia content.

Even though many of the high-end cameras and video recorders are retrofitted with GPS chips, it has been estimated that only about 5% of the existing multimedia content on the Internet is actually geo-tagged [21]. Most of the consumer-produced media content are obtained using low-end

Figure 1.1: Geo-tagging: given a database of training images/videos with their geo-coordinates and textual data, estimate the geo-location of a query video given its textual metadata, visual and audio features.

cameras that do not have GPS chips. Further, privacy concerns have motivated users to disable automatic geo-stamping of photos taken on their phones. However, users usually tag their uploaded videos with textual data that can have some geo-location information. Under this scenario, we ask the question, "Given a set of videos and their associated textual tags, how do we determine their locations?"

The task of automatic estimating the geo-coordinates of a media-recordings goes by different names such as "geo-tagging", "location estimation" or "placing". Just as a human analyst uses multiple sources of information and context to determine geo-location, it is obvious that for location estimation, the investigation of clues across different modalities and the combination with diverse knowledge sources from the web can lead to better results than investigating only one stream of sensor input (e.g. reducing the task to an to image retrieval problem).

The task has recently caught the attention of researchers in the multimedia, signal processing, and machine learning communities because of the large amount of available geo-tagged media on the Internet that could be used as training data, allowing algorithms to work on data volumes rarely

seen before. In addition, the task is hard enough to require the collaboration of many different experts and communities, which is a challenge on its own.

1.1 Definition and Motivation

As initially defined in [21], *Multimodal location estimation* denotes the utilization of one or more cues potentially derivable from different media, e.g. audio, video, and textual metadata to estimate the geo-coordinates of content recorded in digital media. Note that the location of the shown content might not be identical to the location where the content was created, in fact in most cases there is a bias because the camera records GPS coordinates of the location where the camera is located not of the objects captured. For practical purposes, the research presented in here focusses on finding one unique location per video file, even if the video happens to be edited to show different locations.

Work in the field of location estimation is currently creating progress in many areas of multimedia research. As discussed in [21], cues used to estimate location can be extracted using methods derived from current research areas. Since found data from the Internet is used, multimodal location estimation work is performed using much larger test and training sets than traditional multimedia content analysis tasks and the data is more diverse as the recording sources and locations differ greatly. This offers the chance to create machine learning algorithms of potentially higher generality. Overall, multimodal location estimation has the potential to advance many fields, some of which we don't even know of as they will be created based on user demand for new applications. However, apart from the academic motivation described here and fast changes in online world to adapt "Geo" capability as described in Section 1, there are several real-world incentives behind the attempt to solving multimodal location estimation.

Except for specialized solutions, GPS is not available indoors or where there is no line of sight with satellites. So multimodal location estimation helps provide geo-location where it is not regularly available. Movie producers have long searched for methods to find scenes at specific locations or

showing specific events in order to be able to reuse them.

After an incident, law enforcement agencies spend many person-months to find images and videos, including tourist recordings, that show a specific address to find a suspect or other evidence. Also, intercepted audio, terrorist videos, and evidence of kidnappings is often most useful to law enforcement when the location can be inferred from the recording. Up to now, however, human expert analysts have to spend many hours watching for clues of the location of a target video.

Recently, privacy issues based on the information published by individuals on social networking sites have raised increased attention both in the research community as well as in the popular press. One particular issue, namely the hidden inclusion of geographical information has been shown to be a major secuirty risk [19] as it enables so-called cybercasing, i.e. it allows potential attackers to track an individual and gain enough information about the person to pursue criminal offenses, such as stalking and burglary. With the automatic geo-tagging, the target of cybercasing can potentially be expanded to every uploaded medias as incorporation of geo-information would not be necessary.

This report describes an approach to determine the geo-coordinates of the recording place of videos based on textual metadata and visual/audio cues. We describe the realization of the system, analyses the different uses of multimodal cues and gazetteer information.

1.2 Collaboration, Previous Publications, and Funding

This report is a result of a collaborative group project. Other people have made direct contributions to the ideas and results that are included in this report. Gerald Friedland, Oriol Vinyals, and Trevor Darrell were among the first members at ICSI to start working on The Berkeley Multimodal Location Estimation Project[1]. Their work [21] demonstrated the possibil-

[1] http://mmle.icsi.berkeley.edu/mmle

ity of approaching the location estimation problem multimodally. The ICSI Multimodal Location Estimation System [10, 11] was developed by Gerald Friedland, Adam Janin, Howard Lei, Luke Gottlieb, and myself from 2010 to 2011. Gerald was responsible for writing several command-line tools including a wrapper script to handle geographic queries to online database, and the evaluation of the system. Adam worked on applying Bloom Filter to the gazetteer-based system to speed up the experiment. Howard took the lead on the audio-based experiments including user identification, and city identification. Luke was responsible for collecting a video corpus for the development of audio-based location estimation system. I started working on the project from June 2010, and later took the lead on developing the system, analyzing dataset, running experiments on the system, and building a web interface for the demonstration of the system. Kannan Ramchandran and Venkatesan Ekambaram made significant contributions to building a theoretical framework in our system [9]. Venkatesan took the lead on writing graphical model-based hierarchical location estimation system. Martha Larson (Delft University of Technology), Pascal Kelm (TU Berlin), Adam Rae (Yahoo! Research), Pavel Serdyukov (Yandex), and Vanessa Murdock (Yahoo! Research) were task organizers of MediaEval Placing Task 2010 and 2011 [36]. They were responsible for collecting and distributing a benchmarking corpus containing 10,438 videos, which allowed our team and other participants to easily compare the algorithms and examine the advantages and disadvantages. Robin Sommers worked with Gerald Friedland on the cybercasing [19], which lead us to our case study on cybercasing using inferred geo-location using the ICSI Location Estimation System. Nelson Morgan provided his input and feedback throughout the project, especially on the audio-based approach to the task. Finally, Gerald Friedland was integral in all aspects of the project.

Some of the figures and content in this report are adapted from previous publications [10, 11, 18, 9]. This work was supported in part by NGA NURI Grant No. HM11582-10-1-0008, NSF EAGER grant IIS-1138599, and KFAS Doctoral Study Abroad Fellowship.

Chapter 2

Related Work

Given the motivation to solve the task described in the previous chapter, it is no wonder that initial approaches to location estimation have already started several years ago. In earlier articles [39, 49], 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. In other work [23], the goal is to estimate just a rough location of an image taken as opposed to close to exact GPS location. For example, many pictures of certain types of landscapes can occur only on certain places on Earth. Krotkov's approach [12] for robot applications, extracts sun altitudes from images while Jacobs' system [25] relies on matching images with satellite data. In both of these settings single images have been used or images have been acquired from stationary webcams. In the work of [29], the geo-location is also determined based on the estimate of the position of the sun. They provide a model of photometric effects of the sun on the scene, which does not require the sun to be visible in the image. The assumption, however, is that the camera is stationary and hence only the changes due to illumination are modeled. This information in combination with time stamps is sufficient for the recovery of the geo-location of the sequence. A similar path is taken in [26].

Previous work that has been carried out in the area of automatic geo-tagging of multimedia has based on tags have also been mostly carried out

on Flickr images. User-contributed tags have a strong location component, as brought out by [41], who reported that over 13 % of Flickr image tags could be classified as locations using Wordnet. Rattenbury et al. [37] and Serdyukov et al. [40] estimate the posterior distribution of the geo-locations given the tags or vice-versa from the training database and use this to estimate the geo-location of a query video.The approach in [22] reports on combining visual content with user tags. However, the accuracy is only reported with a minimum granularity of 200 km.

The 2010 and 2011 MediaEval Placing tasks [34] provided a common platform to evaluate different geo-tagging approaches on a corpus of randomly selected consumer-produced videos. One of the top performing systems proposed by Van Laere et al. [28] used a combination of language models and similarity search to geo-tag the videos purely based on their textual tags. Several other proposed approaches [6, 22, 13, 27] relied on both textual and visual features. However, none of these systems utilized audio features. The author of this report proposed a hierarchical system [18] that uses the spatial variance of the tags' geo-location distribution to find an initial estimate of the query image location, which is used as an anchor point for a visual nearest neighbor search in the next stage. Our enhanced system [11] incorporates audio features as well, motivated by our previous work on location estimation of ambulance videos from different cities [21] using audio features.

Chapter 3

Dataset

3.1 MediaEval Placing Task Dataset

All experiments described in this report were performed using dataset distributed for Placing Task of MediaEval benchmark[1]. The Placing task is part of the MediaEval benchmarking initiative and requires participants to assign geographical coordinates (latitude and longitude) to each provided test video. Participants can make use of metadata and audio and visual features as well as external resources, depending on the run. During the first year in 2010, participants were asked to submit up to five sets of results, with no restrictions on what data or technique they could use. However, some of the submissions for the second year were given criteria to encourage innovation in situations that reflected the constraints of realistic scenarios. For example, one run was required that used only the visual/audio content of the video for placing, which mimics the common situation of needing to locate a video which has not yet has any textual metadata added to it yet. The other runs allowed participants to implement their technique using solely the data provided, the provided data plus their own gazetteers, or using any data source they wished (as long as the test data were not re-crawled as part of their approach).

The MediaEval Placing Task data set consists of Creative Common-

[1] http://multimediaeval.org/

licensed Flickr videos that were uniformly sampled from all over the world. The videos are in MPEG-4 format and include the Flickr metadata in XML format. The meta-data for each video includes user-contributed title, tags, description, comments and also information about the user who uploaded the videos. Additionally, the metadata also include information about the user's contacts, favorites, and all videos uploaded in the past. The data set was divided into training data (10,216 videos) and test data (5,347 videos).

According to [30], videos were selected both to provide a broad coverage of users, and also because they were geo-tagged with a high accuracy at the "street level". Accuracy shows the zoom level the user used when placing the photo on the map. There are 16 zoom levels, and these correspond to 16 accuracy levels (e.g., "region level", "city level", "street level"). The sets of users from the test and the training collections were disjoint in order to not introduce a user-specific bias. This bias will be discussed further in Section 5.2.

In order to allow visual matching as performed in [23] and to improve coverage, the dataset also contained metadata and features extracted from 3,185,258 Flickr images. However, not all the photos had textual metadata and the photos were only guaranteed to have geo-tagging at least region level accuracy. Training data was supplemented with visual features extracted for both photos and frames of the videos. For videos, every fourth second of a video was extracted using FFmpeg[2]and saved as a JPEG-image. For each of the key frames and for each image provided as training data, nine visual features were extracted using the open source LIRE library [33] with the default parameter settings:

- *Color and Edge Directivity Descriptor (CEDD)* [7] combines color and texture information in a histogram.
- *Gabor Descriptor* [17] is a linear filter using frequency and orientation representations for edge detection.
- *Fuzzy Color and Texture Histogram (FCTH)* [8] combines in one his-

[2]http://ffmpeg.org/

togram 3 fuzzy systems of color and texture information.

- *Color histogram (CH)* [43] is a representation of the distribution of colors in an image.
- *Scalable color descriptor (SCD)* [2] uses vector wavelet coefficients of color images.
- *Auto color correlogram (ACC)* [24] extracts the spatial correlation of colors.
- *Tamura texture descriptor (TD)* [44] extracts histograms of low dimensional texture characteristics.
- *Edge histogram descriptor(EHD)* [32] extracts the distribution of 5 types of edges in each sub-image of 4×4 non-overlapping blocks.
- *Color layout descriptor (CLD)* [32] is designed to capture the spatial distribution of color in an image.

3.2 Additional Data

Because of the non-uniformity of the MediaEval training and test set, we used additional data for more coverage and to make the training data more equally distributed over the earth. In addition to the MediaEval data, we also included the data used for the experiments described in [23]. The data originally consists of 6.4 million images from Flickr categorized into countries and states (in case of US). We sampled pictures from each region and used their unique Flickr photo ID to download the metadata from Flickr. 759,249 metadata records were collected in this way. Furthermore, we collected additional photos from Flickr by dividing the area of the earth into 1 km grid cells, counting the number of photos for each grid cell. If the cell contained more than 15 photos, we sampled 15 % of photos. This resulted in about 1,131,698 new metadata records and photos. All metadata was collected and saved in the same format as the MediaEval photo dataset UserID, PhotoID, HTML link to photo, latitude and longitude, tags, date

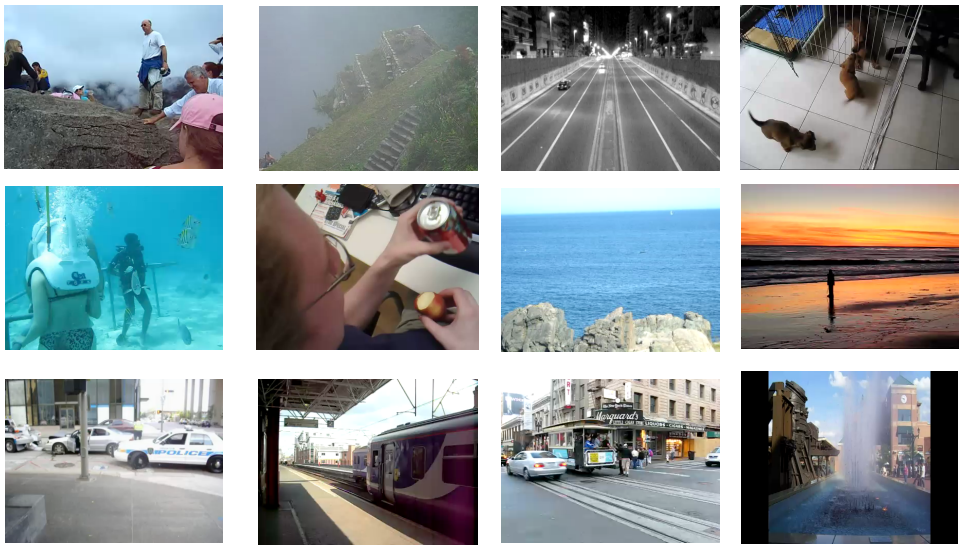

Figure 3.1: Several frames from the MediaEval 2010 test set as described in Section 3.3.

taken, and date uploaded. Again, we ensured that the user set stays disjoint between training and test set.

3.3 Characteristics of the Data

Flickr requires that an uploaded video must be created by its uploader (if a user violates this policy, Flickr sends a warning and removes the video). Manual inspection of the data set lead us initially to conclude that most of visual/audio contents lack reasonable evidence to estimate the location without textual metadata. For example, many videos were recorded indoors or in a private space such as a backyard of a house. This indicates that the videos are not pre-filtered or pre-selected in any way to make the data set more relevant to the task, and are therefore likely representative of videos selected at random.

In order to get an impression of the dataset, we manually watched 84 randomly chosen videos from the training set. Only 2.4 % of them were

recorded in a controlled environment such as inside a studio at a radio station. The other 97.6 % were home-video style with ambient noises and unstable camera settings. 65.5 % of the videos had heavy ambient noises such as crowds chatting in the background, traffic noise, wind blowing into microphone, etc. About 5 % of the videos were edited to contain changed scenes or fast or slow replay. The relatively short lengths of each video should be noted as the maximum length of Flickr videos is limited to 90 seconds. Moreover, about 70 % of videos in our data set have less than 50 seconds playtime. Figure 3.1 shows several sample frames from the MedieEval 2010 test set.

However, metadata provided by the user often provides direct and sensible clues for the task. 98.8 % of videos in the training set were annotated by their uploaders with at least one title, tags, or description, often including location information. For a human, it is a fairly straightforward task to determine from the metadata which keyword or keywords combination indicates the smallest and most accurate geographical entity. However, for a machine, extracting a list of toponym candidate keywords and further choosing a correct single keyword or combination of keywords is a challenging task. Misspelled or compound words concatenated without spaces are commonly found in user-annotated metadata and these add more difficulty to the task. For example, "my trip to fishermanswharf san francisco" should resolve to the "Fisherman's Wharf" in "San Francisco".

Furthermore, partly because of social, political, and economical reasons, in current online video databases (e.g. Flickr and YouTube), videos are not equally distributed over the earth. Therefore downloading a random sample, as performed for MediaEval Placing Task, leads to a large bias towards certain locations. Figure 4.1 shows the distribution of the MediaEval training set. While it will always be difficult to find videos from certain countries or anthropophobic places, a training set that is more equally distributed is desirable for improving global retrieval precision and recall.

Chapter 4

Technical Approach

This chapter is a summary of technical approaches described in previous publications that I (co-)authored [18, 9, 31, 11, 10]. For the ideas from other people, original citations were given in place.

4.1 Textual Features for Geo-Tagging

Most Flickr videos are annotated with a number of descriptive textual tags, which form a valuable source of information for geo-tagging. There are a number of different strategies which may be used to estimate geographic location from text [35].

First, we may attempt to transform this problem into a classification problem by discretizing the locations on Earth into a finite set of disjoint areas, e.g. by using a geodesic grid [40], by clustering the locations of the photos in the training set [46], or by using known administrative boundaries of geographically meaningful regions. By treating each of these areas as a collection of documents, traditional text classification strategies can then be used to find the area in which an unseen video was most likely captured.

Second, by treating the photos and videos in the training set as text documents, standard information retrieval techniques can be used to find the resource which resembles the resource to be geotagged most [18, 16]. An estimation for the location of the latter resource can then be obtained by

using the location of the most similar resource, or a weighted sum of the locations of the k most similar resources.

Third, we may rely on gazetteers to look-up the locations of toponyms which appear among the tags. While very effective for georeferencing standard text documents, there are a number of important challenges when using tags. For example, traditional gazetteer based methods for georeferencing heavily rely on capitalization and context for disambiguation, both of which may be missing in the context of Flickr tags. Other issues are related to the fact that place names are often conjoined on Flickr (e.g. *riodejaneiro*) or broken up into parts (e.g. just *rio*), place names are often misspelled, and place names are concatenated with other nouns (e.g. *halloweenbrazil*) [14].

4.1.1 Geo-Tagging as Classification [35]

A key issue with interpreting the geo-tagging task as a classification problem is how the surface of the Earth is discretized into cellular areas. One approach is to use a geodesic grid [40], which has the advantage of being computationally simple, and of producing areas of roughly the same size (considering that few videos are captured near the poles). Instead, [46] applies a variant of the k-means clustering algorithm to the locations of the resources in the training set. This has the advantage that parts of the world for which a lot of information is available (i.e. where occurrences of photos and videos are denser) are clustered into smaller areas than parts about which very little is known. Moreover, cluster boundaries are more likely to be semantically meaningful, and thus correlate better with the occurrences of certain tags. Another approach is to use mean-shift clustering [13], which shares some of the advantages of k-means, but is more inclined to produce clusters of roughly the same size.

Once a disjoint set of areas $\mathcal{A} = \{a_1, ..., a_n\}$ has been obtained, a standard language modeling approach can be used, estimating the probability that video x was captured in area a as

$$P(a|x) \propto P(a) \cdot \prod_{t \in \mathcal{T}} P(t|a)$$

where $\mathcal{T}$ is the collection of all tags. To estimate the likelihood $P(t|a)$ of seeing tag t in area a, some form of smoothing is needed. Good results are obtained for Bayesian smoothing with Dirichlet priors [40, 46]

$$P(t|a) = \frac{O_{ta} + \mu\, P(t|V)}{O_a + \mu}$$

where O_{ta} is the number of occurrences of tag t in all resources from the training set that are located in area a; O_a is the total number of tag occurrences in area a; $P(t|\mathcal{T})$ is estimated as the percentage of all occurrences of tags in $\mathcal{T}$ that correspond to t. The prior probability $P(a)$ can be taken to be uniform, although slightly better results are obtained when taking $P(a)$ to be proportional to the number of photos in area a (i.e. using maximum likelihood).

Once the most likely area a is obtained, a central element from that area could be used as the location estimation.

4.1.2 Geo-Tagging as Information Retrieval

The concept of 'spatial variance' was used for this approach [18]. For each given tag in the test video record, we determine the spatial variance by searching the training data for an exact match of the tag and creating a list of the geo-locations of the matches. If only one location is found, the spatial variance is trivially small. We pick the centroid location of the top- 3 tags with the smallest spatial variance. This results in 0 to 3 coordinates. In the case of 0 coordinates (e.g. because the video is not tagged or no tags match), we assume the most likely geo-coordinate based on the prior distribution of the MediaEval training set (see Figure 4.1), which is the point with latitude and longitude (40.71257011, -74.10224), a place close to New York City.

For example, if a test videos metadata contains the tags 'Campanile', 'Berkeley', and 'California', the system would match all training videos that contain any of those tags. We then plot the GPS coordinates of the training videos containing the tags Campanile, Berkeley, and California and select the centroid of the tag with the smallest spatial extent (in this case, Cam-

panile) as our final location.

4.1.3 Geo-Tagging using Gazetteers [10]

We used the open service Geonames.org. GeoNames covers all countries and contains 8 million entries of place names and corresponding geo-coordinates. It provides a web-based search engine and an API which returns a list of matching entries ordered by their relevance to the query. A single keyword may cause ambiguity by representing multiple entities (e.g. Paris Texas vs. Paris France). Thus it is crucial to find a combination of keywords that minimizes ambiguity if possible. A computationally inefficient but effective way to do this, is to query the Geonames database exhaustively for every possible combination of keywords. To reduce the run time of the search, we filtered the keywords using a Bloom Filter [3]built over the downloaded database of Geonames. In this method, all compound keywords of every length were tested (e.g. sanfrancisco and San Francisco were both in the Bloom Filter). If the Bloom Filter returned positive, they were added to a candidate list. The Bloom Filter may sometimes return false positives, but these were assumed to be removed by the Geonames search engine. Tags were concatenated into a string in their original order. The order is preserved to handle the context within compound words such as San Francisco or Washington DC. One problem with using a gazetteer is that it has no background model of words that are likely to appear in regular language, i.e. it does give positive results on words such as video and vacation because there is a city of Video in Brazil and a Vacation Island in San Diego. Therefore we filtered out common nouns by using Augmented-WordNet [15]. Augmented-WordNet is an extended version of WordNet [3], that among other things includes annotation for geographical entities. WordNet is a freely available online lexical database of English which contains a network of semantic relationships between words. Note that Flickr videos and photos are annotated in any language so this approach only helped for the English subset. After filtering, we passed the query to the Geonames search engine and retrieved the list of possible matches. We added the entity with the highest relevance

(the first entity in the response list) to the list of candidate entities. Once we obtained the list of candidate entities, we resolved the containment problem (e.g., Fishermans Wharf, San Francisco, CA): Geonames entities provide country code, code of administrative subdivision (typically the city), and feature class parameters. We gave higher priority to entities representing a smaller region (as of Geonames) by removing larger entities containing the smaller entities. Choosing the best match among the list of candidates is similar to the method we used in Section 4.1.2. We plot all candidate entities on a map and pick the one that has the largest count of neighbors with lowest spatial variance. If there is a tie, the coordinate that is closest to the users home location is picked (as described in the videos metadata). If there is no matching entity for all keywords in the metadata of a given video, we apply two backup steps. First, we return the geo-coordinate of the users home location. This is better than a blind guess on the prior, since our observations found that people tend to under-annotate videos about their ordinary everyday life, which tend to have been recorded close to where they reside. If a video did not contain the users home location, we used the default location close to New York City, as explained in Section 4.1.2.

4.2 Visual Features for Geo-Tagging

The impact of the visual modality is of particular interest to the Placing Task. While several previous works [23, 22] have used the visual aspect to estimate the location of images, there has been little work in the literature to automatically place videos on map before MediaEval 2010. The common intuition behind these works is that if two images are similar in the visual feature domain, then they are also likely to be geographically closer.

However, a video offers more visual information than a stationary image in terms of both quantity and quality. These additional information provides both benefits and challenges at the same time. For instance, it would be easier to remove unwanted clutter such as pedestrians and passing cars from the background scenery. On the other hand, a video that is only 10 seconds long typically has between 150 to 300 frames, and picking a frame

(or multiple of frames) that would best represent the geographical aspect of the video is a challenging task. Temporal video segmentation is the first step towards automatic annotation of digital video sequences. Its goal is to divide the video stream into a set of meaningful segments and to find the best representation of a scene. In Kelm et al. [**?**], multiple scenes and longer shots are segmented into smaller part. One representative keyframe with noticeable visual content and in best possible quality is extracted using a visual attention model based on lighting, contrast and camera motion features.

In addition to the provided low-level features, Gist features were extracted from the keyframes and photos which was similar to the approach used in [23]. The k-nearest neighbor is applied to the Gist feature space representation of the development dataset to find a frame (from videos) or a photo that had the most similar and closest-looking scene. Each image was gray-scaled and resized to 128×128 pixels, then a Gist descriptor was extracted with a 5×5 spatial resolution with each bin containing responses to 6 orientation and 4 scales. The Euclidean distance was used to compare the Gist descriptor and nearest neighbor matching was used between the closest pre-extracted frame to the temporal mid-point of a query video and all photos and frames from the development datasets.

Although recent research on automatic geo-location of images using visual features has been promising, the performance of these experiments is not in the same ballpark as the ones that use textual cues. Friedland et al. [18] reports that when cues from multiple modalities (text and visual) were used together, textual cues (annotated by user) played a dominant role. To make visual search effective, it was necessary to narrow down the search boundary to a reasonable degree based on the estimates from the textual cue.

4.3 Audio-based Features for Geo-Tagging

From a previous examination of 84 videos from the evaluation set, we unsurprisingly found that most of the videos' audio tracks are quite "wild". Only

2.4 % of them have been recorded in a controlled environment such as inside a studio at a radio station. As a result, we can expect to be able to exploit background and other noise signatures from the remaining 97.6 % of the videos. We therefore checked the videos for exploitable acoustic properties and found that 65.5 % of the videos have heavy ambient noises. However, 14.3 % of the videos contain music, mostly dubbed afterwards. About 50 % of the videos do not contain human speech, and even for the ones that contain human speech, almost half are from multiple subjects and crowds in the background speaking to one another, making language identification infeasible. Speech recognition is already infeasible given the expected number of different languages and dialects. Fortunately, only 5 % of the videos are edited to contain changed scenes, fast-forwarding, muted audio, or inserted background music.

Given the sparsity of the audio track in videos (as opposed to imagery), we decided to treat the acoustic location estimation as a city identification problem. A video was considered to be located within a city if its geocoordinates were within 5 km of the city center. The following cities were considered for verification, due to the predominance of videos taken in these cities: *Bangkok, Barcelona, Beijing, Berlin, Chicago, Houston, London, Los Angeles, Moscow, NewYork, Paris, Praha, Rio, Rome, San Francisco, Seoul, Sydney, Tokyo.*

We've explored various approaches to city identification [31]. Because of the lack of prior work for the city identification task, there has not been any effective previously-developed technical approaches for the task. Hence, we've decided to approach the city identification task using well-established acoustic modeling-based approaches (i.e. audio-based approaches). The first audio-based approach is derived from the GMM-UBM speaker recognition system [38], with simplified factor analysis and Mel-Frequency Cepstral Coefficient (MFCC) acoustic features C0-C19 (with 25 ms windows and 10 ms intervals), along with deltas and double-deltas (60 dimensions total) [15]. Specifically, for each audio track, a set of MFCC features are extracted and one 128-mixture Gaussian Mixture Model (GMM) is trained for each city, using MFCC features from all audio tracks for the city in the train-

ing set. This is done via MAP adaptation from a universal background GMM model (UBM), which is trained using MFCC features from all audio tracks of all cities in the training set [38]. During testing, the log-likelihood score of MFCC features from test video's audio track are computed for each city-dependent GMM model. Scores for which the city of the test video matches the city of the GMM model are known as true trial scores; scores for which the cities do not match are known as impostor trial scores. Note that the open-source ALIZE speaker recognition system implementation is used for acoustic model training [4], and the MFCC features are obtained via HTK [1].

The second audio-based approach is derived from the GMM-SVM speaker recognition system [5]. In this approach, the same feature extraction is used as in the GMM-UBM approach. A separate GMM model is trained using the audio of each video, via MAP adaptation from a UBM, and the GMM mean parameters are collected into a supervector. Hence, there is one supervector for each video. An SVM model is trained for each city, using the supervectors of the videos belonging to that city in the training data as positive training examples, and supervectors belonging to a set of non-training and testing cities (i.e. development data) as negative training examples. Once an SVM model is trained for each city, a classification score for the supervectors of each test video is obtained for the SVM models of each city.

4.4 Graphical Model Approach to Geo-Tagging

This section is a summary of [9], to which Venkatesan Ekambaram and Kannan Ramchandran made significant contributions.

4.4.1 Data Sparsity

Traditional approaches such as [18, 22] use training sets that are several orders of magnitude larger than the test set. These approaches suffer from the drawback that their accuracies are significantly affected when the training data is sparse. There are two reasons for sparsity in training data. First,

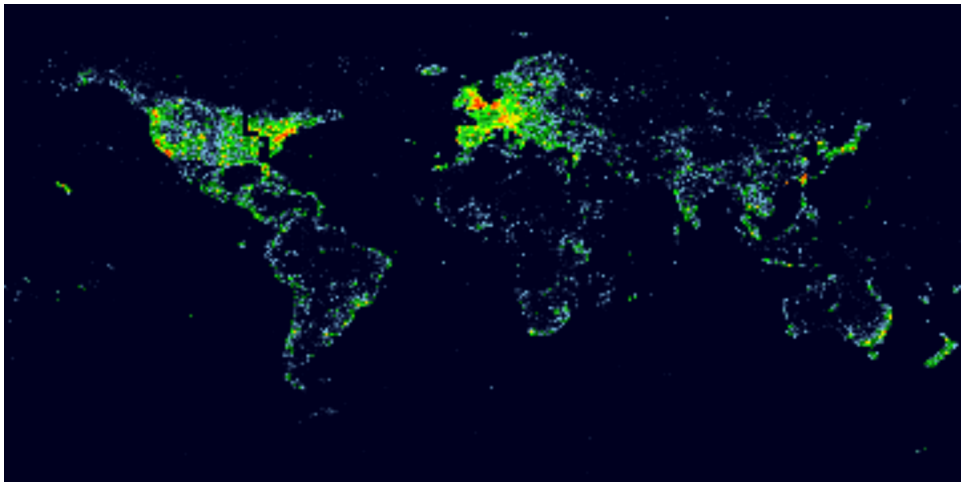

Figure 4.1: The Heatmap shows the distribution of videos and images of the MediaEval Placing Task training set. Randomly sampling videos from Flickr results in a non-uniform geographical prior. The density grows as the color changes from dark blue to red as in rainbow's color order.

it is estimated that only 5% of Internet videos are actually geo-tagged [21], and hence the training set is typically much smaller than the test set, contrary to what is assumed in the literature. Second, the training database is largely skewed toward certain geographical regions (see Fig. 4.1).

For the MediaEval dataset [34], we analyzed the performance of a data-driven algorithm from [18] in different regions with varying data densities. The world map was divided into 64,800 grids of one latitude by one longitude each and the number of training data in each grid was counted. Fig. 4.2 shows the performance of the algorithm for different data densities. The different curves are for different values of the training data density, i.e., we look at the performance in grids with varying quantities of videos: over 6400, 6400, 1600, 400, and 100. The x-axis corresponds to the different error ranges in km and the y-axis to the percentage of geo-tagged videos in these error ranges. Grids with a denser population of training data perform significantly better than those with lesser training data. Thus, estimation

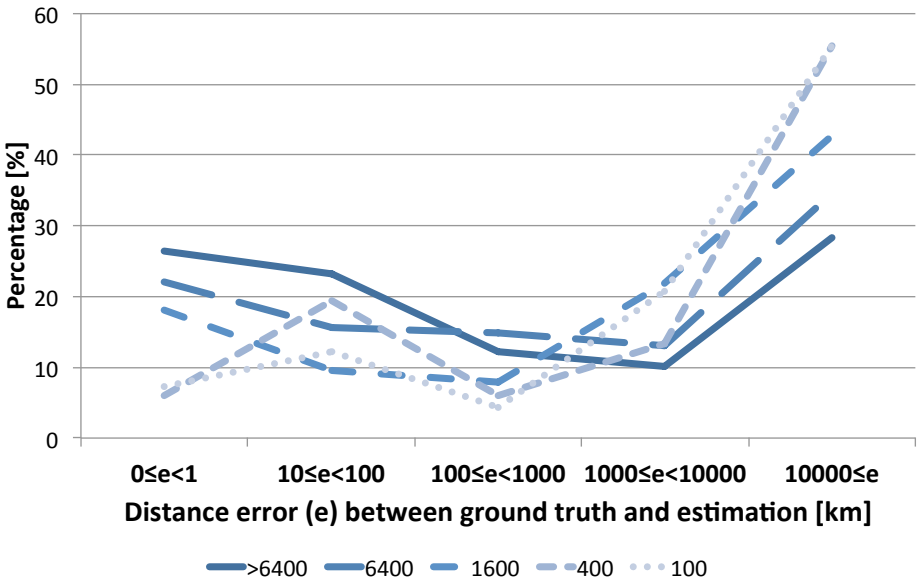

Figure 4.2: Comparison of the performance of a data-driven algorithm[18] on grids with different training data density. Query video from a denser area has higher chance of being estimated with lower error in distance.

models must be developed to handle data sparsity.

The key insight one can hope to exploit is that, even though the training set is smaller, the test set can be potentially much larger than the training set in contrast to what is traditionally assumed in existing algorithms. One can exploit the fact that while the training data may be small, test data may be very large. Existing algorithms do not take this into account. Thus the question of interest is, "Can we intelligently process the test/query videos in such a way that each additional query video not only is placed but also improves the quality of the existing database?" To take a simple example demonstrating our idea, let us suppose that we have two query videos, Q1 and Q2, with associated textual tags {berkeley, sathergate, california} and {sathergate, california} respectively. Assume that the training set only contains the tags, {berkeley, california}. If Q1 and Q2 were to be processed

independently, then Q1's location estimation would be good whereas the location ambiguity of Q2 would be much larger. However, if we jointly process Q1 and Q2, then given that Q1 and Q2 have the tag "sathergate" in common, it is very likely that their locations are also very close by and hence an intelligent algorithm would estimate their locations to be the same, which would improve the location accuracy of Q2. The proposed graphical model framework in the next section applies this principle to a database of query videos and appropriately weighs the edges based on the common tags between the videos.

4.4.2 Overview of the Graphical Model

Graphical models provide an efficient representation of dependencies amongst different random variables and have been extensively studied in the statistical learning theory community [48]. The random variables in our setup are the geo-locations of the query videos that need to be estimated. We treat the textual tags and visual and audio features as observed random variables that are probabilistically related to the geo-location of that video. 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. We give a brief introduction to graphical models and apply the framework to our setup.

An undirected graphical model or a Markov Random Field (MRF) $G(V, E)$ consists of a vertex set V and an edge set E. The vertices (nodes) of the graph represent random variables $\{x_v\}_{v \in V}$ and the edges capture the conditional independencies amongst the random variables through graph separation [48]. The joint probability distribution for a N-node pairwise MRF can be written as follows [48],

$$p(x_1, ..., x_N) \quad = \quad \prod_{i \in V} \psi(x_i) \prod_{(i,j) \in E} \psi(x_i, x_j). \tag{4.1}$$

$\psi(.)$'s are known as potential functions that depend on the probability distribution of the random variables. A typical problem of inference over a graphical model involves finding the marginal distribution of the random variables $p(x_i)$. Finding the exact marginals is in general an NP-hard problem [48] and approximation algorithms such as the sum-product algorithm are used in practice. In the sum-product algorithm, messages are passed between nodes that take the following form:

$$m_{j\to i}(x_i) \propto \int_{x_j} \psi(x_i, x_j)\psi(x_j) \prod_{k\in N(j)/i} m_{k\to j}(x_j)dx_j, \tag{4.2}$$

where $m_{j\to i}(x_i)$ is the message passed from node j to node i and $N(j)$ is the set of neighbors of j . The messages are iteratively passed until convergence and the final estimate of $p(x_i)$ is obtained as follows, $\hat{p}(x_i) \propto \psi(x_i) \prod_{j\in N(i)} m_{j\to i}(x_i)$. This algorithm is seen to work well in practice for many applications.

In order to obtain a graphical model representation for our problem setup, we need to model the joint distribution of the query video locations given the observed data. Since it is hard to obtain the exact probability distribution, we will 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 (i.e., latitude and longitude). 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 set. Fig. 4.3 shows an example of one such graph. The edge potential functions appropriately weigh the edge

based on the common textual tag. The model could be further improved using the audio and visual features as well.

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, under a pairwise MRF assumption, the joint probability distribution factorizes as follows:

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

Given the potential functions, one could use the sum-product iterates (4.2), to estimate $p(x_i|\{t_1^k\}, ..., \{t_N^k\})$.

We now need to model the node and edge potential functions. The literature provides numerous techniques for modeling potential functions and adaptively learning them from the given data [42]. We use the following simple model for the 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 two geographic places). To estimate the parameters of the GMM, we use an algorithm based on Expectation Maximization [47] that adaptively chooses the number of components for different tags using a likelihood criterion. Assuming conditional independence for different tags, 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

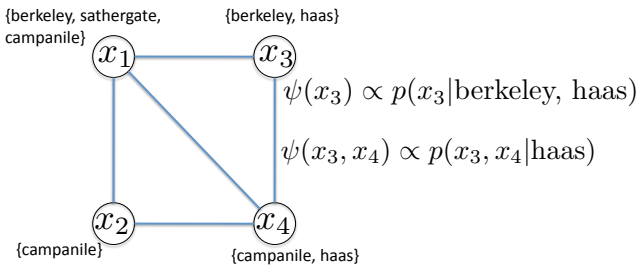

Figure 4.3: An example graphical model for geo-tagging.

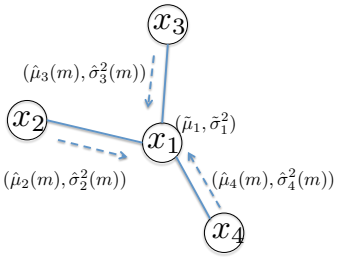

Figure 4.4: Illustration of messages passed along the edges.

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,

$$\psi(x_i, x_j|\{t_i^k\}, \{t_j^k\}) \quad = \quad \begin{cases} 1 & \text{if } x_i = x_j, \\ 0 & \text{otherwise.} \end{cases} \tag{4.3}$$

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 dis-

tribution with the mean ($\tilde{\mu}_i$) and variance ($\tilde{\sigma}_i^2$) given by,

$$(\tilde{\mu}_i, \tilde{\sigma}_i^2) = \left(\frac{\sum_{k=1}^{n_i} \frac{1}{\sigma_i^{k2}} \mu_i^k}{\sum_{k=1}^{n_i} \frac{1}{\sigma_i^{k2}}}, \frac{1}{\sum_{k=1}^{n_i} \frac{1}{\sigma_i^{k2}}} \right), \tag{4.4}$$

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,

$$\hat{\mu}_i(m) = \frac{\frac{1}{\tilde{\sigma}_i^2}\tilde{\mu}_i + \sum_{j \in N(i)} \frac{1}{\hat{\sigma}_j^2(m-1)} \hat{\mu}_j(m-1)}{\frac{1}{\tilde{\sigma}_i^2} + \sum_{j \in N(i)} \frac{1}{\hat{\sigma}_j^2(m-1)}}, \tag{4.5}$$

$$\hat{\sigma}_i^2(m) = \frac{1}{\frac{1}{\tilde{\sigma}_i^2} + \sum_{j \in N(i)} \frac{1}{\hat{\sigma}_j^2(m-1)}}. \tag{4.6}$$

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 has converged. The variance $\hat{\sigma}_i^2(m)$ provides a confidence metric on the location estimate. Fig. 4.4 provides an illustration of the algorithm.

Given the graphical model framework, it is now easy to incorporate visual and audio features. These features can be used to modify the potential functions $\psi(x_i, x_j)$ on the edges. The intuition is that if two images are similar in some feature space, then they are also most likely geographically closer. However, this intuition holds true only when we already have a coarse estimate of their locations and we want to further refine it. For this purpose, we adopt a hierarchical approach. We first obtain a coarse estimate of the query video's location using only the tags in the graphical model. We then

choose a subgraph for each query video consisting only of query and training videos within some particular radius of each other. Visual and acoustic features are obtained for each video using GIST and MFCC features similar to what Friedland et al. use in [11]. The probability distribution of the geographic distance between two videos given the closeness of the videos in the visual and audio feature space is modeled as a mixture of exponentials and is incorporated in the edge potential function $\psi(x_i, x_j)$.

Chapter 5

Experimental results

This chapter is a summary of experimental results and analysis of the technical approach from previous publications that I (co-)authored [18, 9, 31, 11, 10].

5.1 Evaluation [35]

To evaluate the performance of each technique, the geodesic distance between the ground truth coordinates and those of the output from a participants system were compared. To take into account the geographic nature of the task, the Haversine distance was used. This measure is calculated thus:

$$d = 2 \cdot r \cdot \arcsin\left(\sqrt{h}\right) \tag{5.1}$$

$$h = sin^2\left(\frac{\phi_2 - \phi_1}{2}\right) + cos(\phi_1)cos(\phi_2)sin^2\left(\frac{\psi_2 - \psi_1}{2}\right) \tag{5.2}$$

where d is the distance between points 1 and 2 represented as latitude (ϕ_1, ϕ_2) and longitude (ψ_1, ψ_2) and r is the radius of the Earth (in this case, the WGS-84 standard value of 6,378.137km is used).

The following results should be considered with the following points in mind:

- The scope of possible video placement is considered to be the entire planet.

- This implies that the maximum possible distance between any two points is half the equatorial circumference, which is 20,037.51km (2 d.p.) according to WGS-84 standard. This provides an upper bound to any distance error. However, this can be improved by assuming a trivial video placing approach that assigns a test video the location of a randomly chosen training video. This would then provide an average upper bound distance of 12,249 km using the 2011 training and test data.

While it was important to minimize the distances over all test videos, runs were compared by finding how many videos were placed within a threshold distance of 1 km, 5 km, 10 km, 50 km and 100 km. For analyzing the algorithm in greater detail, here we also show distances of below 100 m and below 10 m. The lowest distance category is about the accuracy of a typical GPS localization system in a camera or smartphone.

5.2 Text-based Geo-Tagging using Spatial Variance

First we discuss the results as generated by the algorithm described in Section 4.1.2. The results are visualized in Figure 5.1. The results shown are superior in accuracy than any system presented in MedieEval Placing Task. Also, although we added additional data to the MediaEval training set, which was legal as of the rules explained above, we added less data than other systems in the evaluation, e.g. [45]. Compared to any other system, including our own, the system presented here is the least complex.

5.3 Using a Geographical Gazetteer

We found that incorporating gazetteer information can help significantly with sparse datasets. However, with enough sample records, tag matching as described in Section 4.1.3 outperforms the gazetteer approach, even when incorporating the Flickr-specific home location as described above. Figure 5.2 shows the results comparing tag matching and using Geonames plus a user's home location.

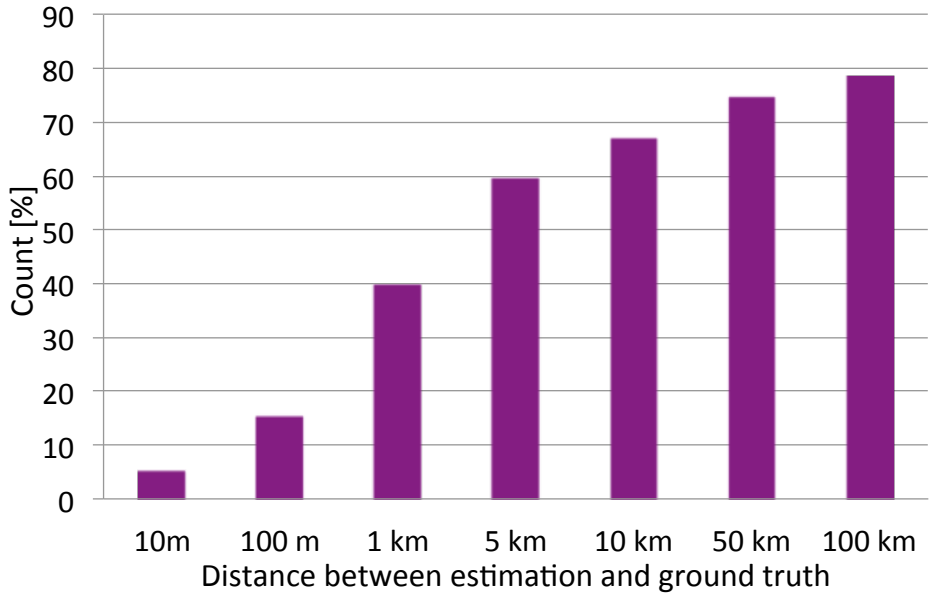

Figure 5.1: The resulting accuracy of the algorithm as described in Section 4.1.2.

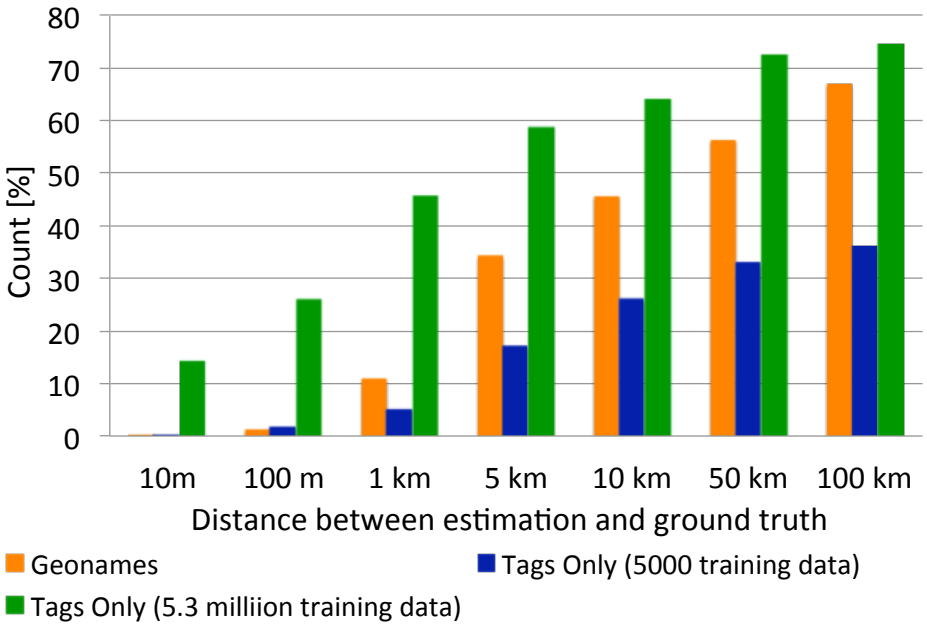

Figure 5.2: Comparing the use of a geographical gazetteer versus the technical approach in Section 4.1.2 with different training data volumess. See also discussion in Section 4.1.3.

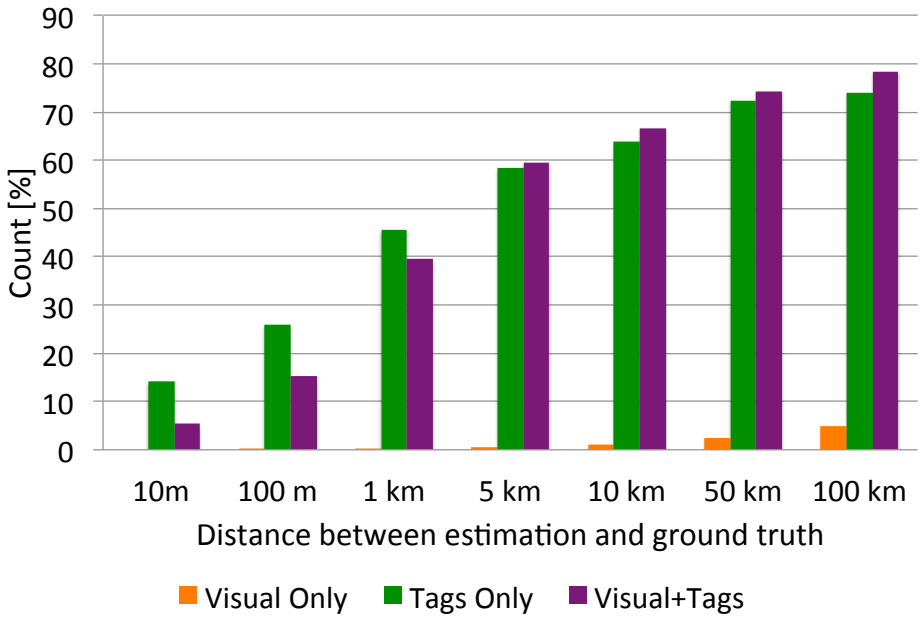

Figure 5.3: The resulting accuracy when comparing tags-only, visual-only, and multimodal location estimation as discussed in Section 5.5.

5.4 Using the Visual Cue

As a comparison, the image-matching based location estimation algorithm in [23] started reporting accuracy at the granularity of 200 km. As can be seen in Figure 5.3, this is consistent with our results: Using the location of the 1-best nearest neighbor in the entire database compared to the test frame results in a minimum accuracy of 10 km. In contrast to that, tag-based localization reaches accuracies of below 10 m. For the tags-only localization we modified the alogorithm from Section 4.1.2 to output only the 1-best geo-coordinates centriod of the matching tag with lowest spatial variance and skip the visual matching step. While the tags-only variant of the algorithm performs already pretty well, using visual matching on top of the algorithm decreases the accuracy in the finer-granularity ranges but increases overall accuracy, as in total more videos can be classified below 100 km. Out of

the 5091 test videos, using only tags 3774 videos can be estimated correctly with an accuracy better 100 km. The multimodal approach estimates 3989 correctly in the range below 100 km.

Despite that the result was from using only one visual feature, the performance is worse when compared to [23, 22]. However, it should be noted that the test video set in this task was not filtered or selected for the task. This 'wildness' of the test set made the task much more difficult.

5.5 Using the Audio Cue

We used a different error metric for the audio-based approach as the experiment was set up to classify audio as one of selected cities. During scoring, a threshold is established for distinguishing the true trial scores from the impostor trial scores. The system performance is based on Equal Error Rate (EER), which is the false alarm rate (percentage of impostor trial scores above the threshold) and miss rate (percentage of true trial scores below the threshold) at a threshold where the two rates are equal. We had 32.3 % EER with a test set of 285 videos (random would be 50 % EER). In other words, almost 70 % of the test videos, when tested against its correct target city, were identified as belonging to that city. The results demonstrate the feasibility of using the audio tracks of videos to identify their cities of origin. However, when the result of audio cue was combined with other modalities, it contirubted very little to the accuracy of the overall system. Still, it does provide an interesting area for further investigation as we have not yet been able to exploit audio to its full potential.

5.6 The Influence of Non-Disjoint User Sets

Each individual person has his own idiosyncratic method of choosing a keyword for certain events and locations when they upload videos to Flickr. Furthermore, the spatial variance of the videos uploaded by one user is low on average. At the same time, a certain amount of users uploads many videos. Therefore taking into account to which user a video belongs seems

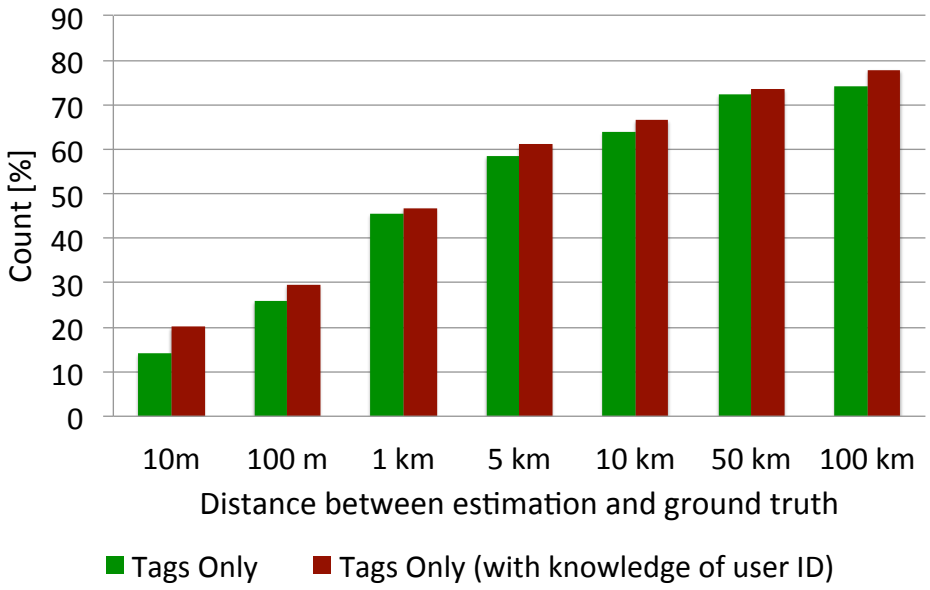

Figure 5.4: The resulting accuracy when taking into account user locality as discussed in Section 5.6.

to have a higher chance of finding geographically related videos. For this reason, videos in the MediaEval Placing Task test dataset were chosen to have a disjoint set of users from the training dataset. However, the additional training images provided for MediaEval Placing Task are not user disjoint with the test videos. Therefore we are able to run an experiment exploiting the user overlap. Instead of searching for a matching keyword in all videos within the dataset, we limit the search to just the videos uploaded by the same user, cutting down on confusion. If the user is not in the training image set, we use the tags-only algorithm as described in the previous paragraph. The results are shown in Figure 5.4. As can be seen, the accuracy is increased significantly, especially in the regions below 1 km. While exploiting user locality is legal as of the rules of MediaEval Placing Task, it is generally considered bad practice. The Flickr dataset often contains many videos and photos by the same individual and exploiting this property of the database might not be helpful to solve the multimodal location estimation problem in general.

Results for GMM-UBM experiments demonstrate up to a 28.9% relative EER improvement (32.3% EER vs. 23.0% EER) if the training and test sets have common users (albeit on a different set of trials). This demonstrates that implicit user-specific effects, such as channel artifacts from the recording device, and the user's preferred video-recording environment, contribute significantly to accuracy.

5.7 Graphical Model Approach [9]

Performance evaluation is carried out for different values of training and test data. Fig. 4.2 shows the performance of [18]. We evaluate the performance of the proposed algorithm for different subsets of the training videos in comparison with this system. In order to understand the performance improvements obtained in the data sparse case, we use 500 training videos and plot the performance improvement as more and more query videos are added to the system. Fig. 5.5 shows the performance improvement in different categories. The number of test videos for this plot is fixed at 5347.

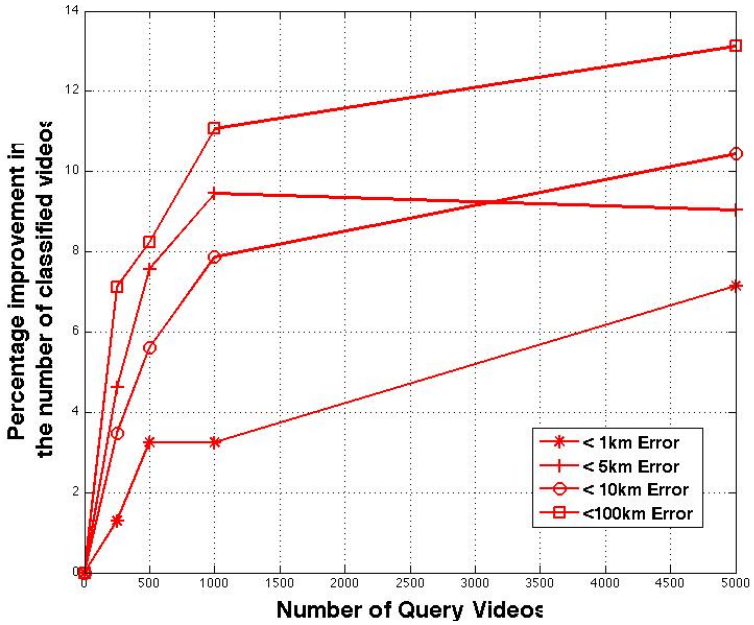

Figure 5.5: Performance improvement in geo-tagging 5347 videos using a training set of 500 videos as a function of the number of query videos used in the graphical model.

The x-axis shows the number of query videos that were used in forming the graph. For example, the point 1000 on the x-axis means that, while building the graph for all the 5347 videos, each video had a neighbor only from these 1000 videos. The best case is when any video in 5347 videos could be a neighbor of any other video. The y-axis is the performance improvement over the baseline system. The different curves correspond to different error categories i.e. $< 1km$ error, $10km - 100km$ error etc. The performance improvement in each category is calculated as follows,

$$A = \text{Num. of videos with} < 1km \text{ error using our alg,}$$

$$B = \text{Num. of videos with} < 1km \text{ error using the baseline alg,}$$

$$\text{Perf. improvement in} < 1km \text{ error category} = \frac{A - B}{B} \times 100.$$

This extends to the other error categories. The performance improvement can be over 10% and increases with the number of test videos. However, one can clearly see the diminishing returns with increasing number of query videos.

Fig. 5.6 shows the performance improvement from the test videos as the number of training videos increases. In this plot, the underlying graph was generated with all the 5347 test videos and the performance improvement was observed as a function of the number of training videos. Though there are gains initially, as the number of training videos increases, the performance improvement obtained by using the query videos decreases. This is to illustrate that in the sparse data case with fewer training videos, the performance improvements can be large. However, with a larger training set, the performance improvement can be very marginal. In practice, given that most of the videos are not geo-tagged, i.e., the test set can be orders of magnitude larger than the training set, one can hope to achieve significant performance gains.

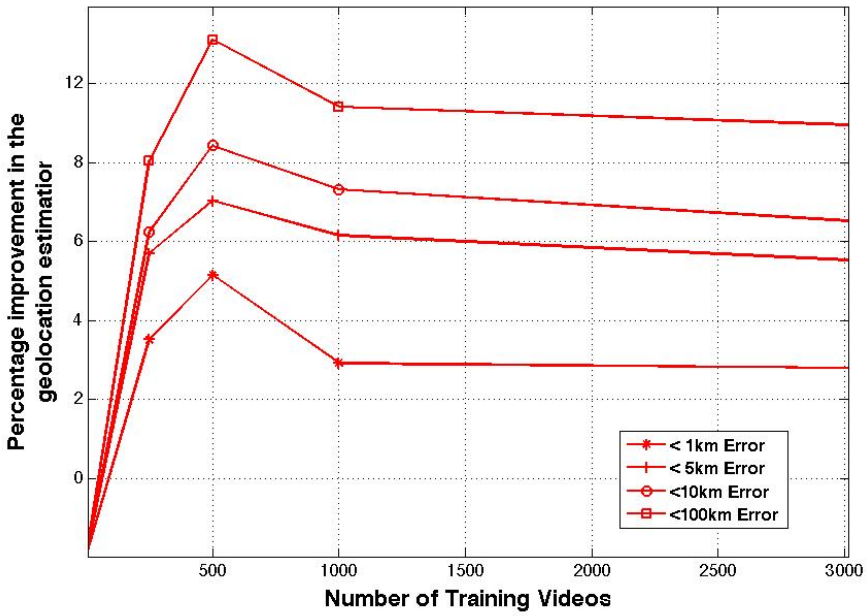

Figure 5.6: Performance improvement in geo-tagging 5347 videos as a function of the number of training videos.

Chapter 6

Conclusion

The percentage of photos uploaded from cell phones is growing quickly, and more and more photos will be annotated with the location automatically at the time of recording using GPS receiver or triangulation of cellular towers. Nevertheless, geo-tagging of photographs and video recordings is still an important problem. Only 5% of the content in the internet is geo-referenced and the automatic geo-tagging would be very useful for browsing and organizing media recordings where location information has not been collected or cannot be collected.

6.1 Summary of Work

In this report we gave an overview of the state of the art approaches in automatic geo-tagging of videos using textual, audio-based and visual features. The report discusses several experiments that contribute to the understanding and validity of the task. Even though audio or visual information alone does not seem to be competitive compared to a tags-only approach, the performance of using each cue is still much better than random and the combination of the multiple media sources does improve the overall performance. Note that there are many cases where audio or visual content plays a dominant role even when a textual cue is given. In fact, the utilization of the visual and audio content is crucial for the robustness of the system, as

the common situation of needing to locate a video may not have any textual metadata added yet.

We verified the claim that overlapping users in test and training sets will introduce a bias which might hinder the generalizability of approaches. We addressed the case where the training data set is sparse. We discussed the use of gazetteer data and found that semantic technologies can be helpful but mostly in situations where not enough training data is available. Finally, we proposed a graphical model framework, posed the problem of geo-tagging as one of inference over this graph, and showed that performance improvements can be achieved by processing the test data set in an intelligent way.

When the user provides sufficient tags in the metadata and the tags are location-specific to where the video was taken, our approach shows potential to return the location very accurately. In fact, our algorithm already outperforms the availabilty of explicitly geo-tagged multimedia (e.g. as EXIF data), as only about 5 % of Flickr videos and images are geo-tagged [20] and our tags-only approach is already able to classify about 14 % of the videos within the 10 m range.

6.2 Future Research Directions

Various issues remain to be explored. The most important question is how to combine all modalities so that they would complement each other when used jointly. The current method uses text as the primary cue to set search boundary for visual or acoustic similarity search. While this hierarchical approach seems ad-hoc, it works well enough. The accuracy of audio/visual approaches without the use of textual tags is low, and needs to be improved for practical purposes. However, there are many cases where the visual or audio cues contain more direct and useful clue than textual cue. For instance, one of the video in the dataset shows the name of a restaurant on the wall whereas the tags doesn't provide any useful information about the place. The recorder of a video might narrate enough information about the place to allow a very accurate estimation of the location. In the future, we may run an OCR system on the frames of the videos to capture texts

or run a speech recognition system on audio tracks to dictate whatever the uploader might have said. These will be among many examples that would benefit from an algorithm that would put more weight on visual or audio cue if the system is confident about the evidence found in those cue.

Data sparsity is another important issue that needs to be addressed. We proposed a graphical model framework, that achieved a performance improvement by processing the test data set. However, even with bigger dataset, the population bias inherent in the dataset will still cause data sparsity, which will be the root of a performance bottle-neck for data-driven approach. Semantic technologies (i.e., gazetteer-based approach) may complement this.

There are some videos that confusingly contain toponyms in their metadata to describe an incident or an object which is not proximal to where the video was recorded (e.g. "Goodbye Oregon, hello San Francisco"). While not an exception, these cases are much more difficult. We expect that further integration with other media will help here.

The modeling of edge potentials in the graphical model is very naive, and one can explore richer models such as hierarchical models (e.g., latent dirichlet topic models) to model the correlations on the edges. The node potentials are further modeled as a product of the distributions given each tag individually. However, the distribution of the locations given multiple tags is not independent. For example, the location distributions of the tags "berkeley" and "sathergate" are clearly not independent. Hence a better correlation model needs to be explored for these distributions.

Bibliography

[1] Hmm toolkit (htk), http://htk.eng.cam.ac.uk.

[2] E. Albuz, E. Kocalar, and A.A. Khokhar. Scalable color image indexing and retrieval using vector wavelets. *Knowledge and Data Engineering, IEEE Transactions on*, 13(5):851–861, 2001.

[3] B. H. Bloom. Space/time trade-offs in hash coding with allowable errors. *Communications of the ACM*, 13(7):422–426, 1970.

[4] J. F. Bonastre, F. Wils, and S. Meignier. ALIZE, a free toolkit for speaker recognition. pages 737–740.

[5] W. M. Campbell, D. E. Sturim, and D. A. Reynolds. Support vector machines using gmm supervectors for speaker verification. *IEEE Signal Processing Letters*, 13:308–311, 2006.

[6] 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.

[7] 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.

[8] 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.

[9] J. Choi, V. Ekambaram, G. Friedland, and K. Ramchandran. Multimodal location estimation of consumer media: Dealing with sparse training data. *To appear in the proceedings of the IEEE International Conference on Multimedia and Expo (ICME 2012), Melbourne, Australia*, 2012.

[10] J. Choi, A. Janin, and G. Friedland. The 2010 ICSI Video Location Estimation System. In *Proceedings of MediaEval*, October 2010.

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

[12] F. Cozman and E. Krotkov. Robot localization using a computer vision sextant. In *IEEE international conference on robotics and automation*, pages 106–106, 1995.

[13] 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.

[14] D. Ferrés and H. Rodríguez. TALP at MediaEval 2010 Placing Task: Geographical focus deterction of Flickr textual annotations. In *Working Notes of the MediaEval Workshop*, 2010.

[15] S. B. Davis and P. Mermelstein. Readings in speech recognition. chapter Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, pages 65–74. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990.

[16] F. Krippner, G. Meier, J. Hartmann, and R. Knauf. Placing Media Items Using the Xtrieval Framework. In *Working Notes of the MediaEval Workshop*, 2011.

[17] H.G. Feichtinger and T. Strohmer. *Gabor analysis and algorithms: Theory and applications*. Birkhauser, 1998.

[18] 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.

[19] G. Friedland and R. Sommer. Cybercasing the joint: On the privacy implications of geo-tagging. *International Computer Science Insitute*, Technical Report TR-10-005, May 2010.

[20] G. Friedland and R. Sommer. Cybercasing the Joint: On the Privacy Implications of Geo-Tagging. In *Proc. USENIX Workshop on Hot Topics in Security*, August 2010.

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

[22] 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.

[23] J. Hays and A.A. Efros. IM2GPS: estimating geographic information from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008*, pages 1–8, 2008.

[24] J. Huang, S.R. Kumar, M. Mitra, W.J. Zhu, and R. Zabih. Image indexing using color correlograms. In *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*, pages 762–768. IEEE, 1997.

[25] N. Jacobs, S. Satkin, N. Roman, R. Speyer, and R. Pless. Geolocating static cameras. In *IEEE international conference on computer vision*, pages 1–6, 2007.

[26] I. Junejo and H. Foroosh. Estimating geo-temporal location of stationary cameras using shadow trajectories. *Computer Vision–ECCV 2008*, pages 318–331, 2008.

[27] 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.

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

[29] J.F. Lalonde, S. Narasimhan, and A. Efros. What does the sky tell us about the camera? *Computer Vision–ECCV 2008*, pages 354–367, 2008.

[30] M. Larson, M. Soleymani, P. Serdyukov, S. Rudinac, C. Wartena, V. Murdock, G. Friedland, R. Ordelman, and Gareth J.F. 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.

[31] H. Lei, J. Choi, and G. Friedland. Multimodal city-verification on flickr videos using acoustic and textual features. *To appear in the proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2012), Kyoto, Japan*, 2012.

[32] B.S. Manjunath, J.R. Ohm, V.V. Vasudevan, and A. Yamada. Color and texture descriptors. *Circuits and Systems for Video Technology, IEEE Transactions on*, 11(6):703–715, 2001.

[33] L. Mathias and S. A. Chatzichristofis. Lire: Lucene Image Retrieval – An Extensible Java CBIR Library. In *In proceedings of the 16th ACM International Conference on Multimedia*, pages 1085–1088, October 2008.

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

[35] A. Rae, M. Larson, S. Schockaert, O. Van Laere, P. Kelm, J. Choi, and G. Friedland. Our media's place in the world: The state of automatic geotagging. *Unpublished*, 2012.

[36] Adam Rae, Vanessa Murdock, Pavel Serdyukov, and Pascal Kelm. Working notes for the placing task at mediaeval 2011. In *MediaEval*, 2011.

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

[38] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. Speaker verification using adapted gaussian mixture models. In *Digital Signal Processing*, page 2000, 2000.

[39] 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.

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

[41] B. Sigurbjoernsson and R. Van Zwol. Flickr Tag Recommendation based on Collective Knowledge. In *ACM WWW*, pages 327–336, April 2008.

[42] L. Song, A. Gretton, D. Bickson, Y. Low, and C. Guestrin. Kernel belief propagation. *Preprint arXiv:1105.5592*, 2011.

[43] M.J. Swain and D.H. Ballard. Indexing via color histograms. In *Computer Vision, 1990. Proceedings, Third International Conference on*, pages 390–393. IEEE, 1990.

[44] 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.

[45] O. Van Laere, S. Schockaert, and B. Dhoedt. Ghent University at the 2010 Placing Task. In *Proceedings of MediaEval*, October 2010.

[46] O. Van Laere, S. Schockaert, and B. Dhoedt. Finding locations of Flickr resources using language models and similarity search. In *Proceedings of*

the 1st ACM International Conference on Multimedia Retrieval, pages 48:1–48:8, 2011.

[47] N. Vlassis and A. Likas. A greedy em algorithm for gaussian mixture learning. *Neural Processing Letters*, 15(1):77–87, 2002.

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

[49] 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.