Multimodal Location Estimation

Gerald Friedland
International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704-1198
fractor@icsi.berkeley.edu

Oriol Vinyals
International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704-1198
vinyals@icsi.berkeley.edu

Trevor Darrell
International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704-1198
darrell@icsi.berkeley.edu

ABSTRACT

In this article we define a multimedia content analysis problem, which we call multimodal location estimation: Given a video/image/audio file, the task is to determine where it was recorded. A single indication, such as a unique landmark, might already pinpoint a location precisely. In most cases, however, a combination of evidence from the visual and the acoustic domain will only narrow down the set of possible answers. Therefore, approaches to tackle this task should be inherently multimedia. While the task is hard, in fact sometimes unsolvable, training data can be leveraged from the Internet in large amounts. Moreover, even partially successful automatic estimation of location opens up new possibilities in video content matching, archiving, and organization. It could revolutionize law enforcement and computer-aided intelligence agency work, especially since both semi-automatic and fully automatic approaches would be possible. In this article, we describe our idea of growing multimodal location estimation as a research field in the multimedia community. Based on examples and scenarios, we propose a multimedia approach to leverage cues from the visual and the acoustic portions of a video as well as from given metadata. We also describe experiments to estimate the amount of available training data that could potentially be used as publicly available infrastructure for research in this field. Finally, we present an initial set of results based on acoustic and visual cues and discuss the massive challenges involved and some possible paths to solutions.

Categories and Subject Descriptors

H3.1 [**Information Storage and Retrieval**]: Indexing methods; I4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*Sensor Fusion*

General Terms

Experimentation

MM'10, October 25–29, 2010, Firenze, Italy.

Keywords

Video, Audio, Multimodal, Location Estimation, Content Analysis

1. INTRODUCTION

In the last few decades, branches of machine learning have been divided along the types of data that were to be processed because research communities have developed as soon as a certain data type could be captured, stored, and processed in a reasonable amount of time. As a result, artificial intelligence is split into speech, computer vision, natural language processing, and so on. Today's computers have begun to have the computational power and memory to be able to process a large amount of data in different sensory modalities. This, in combination with the large amount of multimedia data freely accessible in the Internet, provides an opportunity to improve the robustness of current machine learning approaches and attack problems that are impossible to solve satisfactorily using only a single modality.

In this article, we introduce a new multimedia content analysis task that has only recently become even remotely possible to tackle: the estimation of the location of a media recording that lacks geo-location metadata. We call this task multimodal location estimation. Just as human analysts use multiple sources of information to determine geo-location, it is obvious that for location detection, the investigation of clues across different sensory modalities and their combination can lead to better results than investigating only one stream of sensor input. Therefore, approaches to tackle this task should be inherently multi-media.

Let's imagine a video in which the location is unknown. Acoustic event detection on the audio track reveals a siren usually only found in American police cars, and automatic language identification detects English language spoken with a southern-state dialect. An image object recognizer finds several textures that are typical to a specific terrain with vegetation found only in a humid, sub-tropical area. The classification of birds singing in the background indicates that the recording might be from the southern portion of the US. For a couple of frames, a building is observed that matches Flickr photos of the Parthenon. The combination of these clues is sufficient evidence to conclude that the video is from the Nashville, TN area. Location estimation is an inherently hard problem, since in many cases it is completely impossible to assign the location of a piece of video as there are simply no indicators.

In this article we describe our idea of growing multimodal location estimation as a research field in the multimedia

Figure 1: A figurative description of multimodal location detection.

community. Based on examples and scenarios, we propose different research directions to leverage cues from the visual and the acoustic portions of a video as well as from any given metadata. We describe experiments to estimate the amount of available training data and argue that the research has now become feasible. An initial set of results is presented based on acoustic and visual cues. It again indicates the general feasibility of the task but also serves as a base to discuss the massive challenges involved and some possible paths to solutions.

The article is organized as follows. We start with the definition of multimodal location estimation in Section 2, followed by a comparison of our definition with prior work in Section 3. Section 4 then describes why we think this is an interesting field to work on and the potential applications of location estimation. Section 5 continues with proposed initial steps and research directions towards solving the task. Section 6 then reports on our experiments estimating how much training data would be available for this task in the Internet before Section 7 presents a very first attempt of a multimodal location estimation algorithm. Section 8 concludes the article with final remarks.

2. DEFINITION

We define location *estimation* as the task of estimating the geo-coordinates of all the content recorded in digital media. Figure 1 figuratively describes the idea. Note that the location of the shown content might not be identical to the location where content was created. Also, use of split screen, cutting, and other techniques might allow a video, for example, to show multiple locations. For practical purposes, research will likely concentrate on finding one unique location per file. *Multimodal* location estimation denotes the utilization of one or more cues potentially derivable from different media, e.g. audio and video. Importantly, location estimation as defined above is only one possible research direction. In many cases, slight variations of the task might also provide valuable information. Location *detection*, for example can be defined as the task of finding whether a video contains any cue that might help find a location. For example, the detection of a bird singing without actually classifying the bird would be a first step in a chain of (automatic and non-automatic) analysis steps towards identifying the coarse location of a video. Likewise, location *verification* is the task of finding whether a video has been recorded at a given place. This is not only very valuable for search and retrieval ("find all videos from Times Square in Manhattan, NY") it is also interesting for the validation of existing databases, i.e. verifying whether a given description of a video is true; it thus has direct connections with the fields of cybersecurity and forensics. Location estimation itself can either be interpreted as a *classification* or as a *regression task*. While the accurate estimation of concrete geo-coordinates is a regression task, due to practical concerns with data sparsity and maintaining tractability, initial work in the field will surely start as a classification task (compare Section 7). The classification task takes the following form: Given training data at m locations assign n test recordings to these locations. The *closed* classification tasks would only include test data recorded at the given trained locations, the *open* classification task would include test recordings from different locations. In the latter case systems must therefore be able to identify unknown locations. Finally, we define *relative* location estimation as the task of detecting whether two recordings were recorded at the same or similar place. Tasks include whether videos have been located outdoors/indoors, in a city/outside a city, or near a train station/far away from a train station. Any of these tasks might be researched targeting a fully automatic approach, in combination with (partially descriptive) metadata, or as interactive approaches.

3. PRIOR ART

Recent articles [12, 14] indicate initial results that already show that location estimation is solvable by computers to some extent. The approaches presented in the referenced articles reduce the location detection task to a retrieval problem on a self-produced, location-tagged image database. The idea is that if the image is the same then the location must be the same too. As discussed in Section 1, we think that only a very small part of the location recognition problem can be solved using image retrieval techniques. In other recent work [7], 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. All of these cues, together with acoustic counterparts, could potentially be fused together into a single robust estimate of location under our proposed framework. Krotkov's approach [3] extracts sun altitudes from images while Jacobs' system [8] 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 [10], 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 geolocation of the sequence. A similar path is taken in [9].

There are potentially many artificial intelligence tasks that could assist in determining geo-location, such as keyword spotting, language identification, and sign recognition. In general, however, the systematic investigation of automatic location estimation has a very short research history. As far as we know, the problem of automatically estimating geo-location has been considered only for images and only under specific constrained conditions. Despite the potential, described in the next Section, there has never been an at-

tempt on video or audio data and a multimodal attempt has never even been considered.

4. POTENTIAL IMPACT AND USES

4.1 Research Impact

Work in the field of location estimation will create progress in many areas of multimedia research. As discussed in Section 5 cues used to estimate locations can be extracted using methods derived from current research areas. Acoustic processing fields that could contribute mostly would be speech recognition, language recognition, and acoustic event detection. From computer vision, optical character, sign, and general object recognition methods will be very useful. We already described the use of image retrieval methods in Section 3. Similarly, natural language processing methods would be helpful in many regards as well. In addition, knowledge from geography, for example used to calculate distances, will shape the field as much as new HCI methods for building interfaces that allow semi-automatic location estimation applications. The rather young field of multimodal integration in computer science will develop further as new methods for the combination of cues and media will be demanded. New classification tasks, similar to the one described in Section 7 on ambulances, will gain attention. 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 (sic!) 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 users demanding applications. Some of these are discussed in the following two paragraphs.

4.2 Media Organisation and Retrieval

Location-based services are rapidly gaining traction in the online world. An extensive and rapidly growing set of online services is collecting, providing, and analyzing geo-information. Besides major players like Google and Yahoo!, there are many smaller start-ups in the space as well. The main driving force behind these services is the enabling of a very personalized experience. *Foursquare* for example encourages its users to constantly "check-in" their current position, which they then propagate on to friends; *Yowza!!* provides an iPhone application that automatically locates discount coupons for stores in the user's current geographical area; and *SimpleGeo* aims at being a one-stop aggregator for location data, making it particularly easy for others to find and combine information from different sources. In a parallel development, a growing number of sites now provide public APIs for structured access to their content, and many of these already come with geo-location functionality. Flickr, YouTube, and Twitter all allow queries for results originating at a certain location. Likewise, we believe retro-fitting archives with location information will be attractive to many businesses and enables usage scenarios we don't even think of yet. Also, except for specialized solutions, GPS is not available indoors or where there is no line of sight with the satellites. So multimodal location estimation would help enabling geo-location where it is not regularly available. For example, vacation videos and photos could now be grouped even if location isn't 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. This would partly be enabled by retrofitting location information.

4.3 Law Enforcement

After an incident, law enforcement agancies spend many person-month 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. Until today, however, human expert analysts have to spend many hours watching for clues on the location of a target video. Even when there is an obvious clue that could easily be identified by a computer, humans have to pay attention and watch the video carefully until the point where the hint is revealed. If the human expert happens not to pay attention at the particular set of frames where the audio or image clue appears, the location might never be determined. There are many clues that are hard to perceive for a human being, such as a masked sound, a small object, or slight variations on lighting conditions that are the result of a unique landscape not captured by the camera. Therefore, even only partially successful semi-automatic location detection would reduce the work for human analysts to detect the location of videos, especially in cases that are obvious. Human experts could concentrate on the more difficult cases. The computer might provide confidence output and suggestions that might be judged by the analyst, which will save workload, even on videos that are not completely classifiable by the computer.

5. DIRECTIONS OF RESEARCH

In this section we indicate some potential directions and first steps for location estimation research by breaking up the tasks by media type, i.e. the search for visual and acoustic cues as well as the cues from accompanying metadata.

5.1 Visual Location Estimation

As discussed in Section 3, research on image-based location estimation has already begun with an approach of reducing the location estimation problem to an image retrieval problem in a large database of environmental images. In order to tackle the location estimation problem at a larger scale, using a broader class of media (image, video, audio, text), a hierarchy of tasks and associated techniques needs to be developed. In addition to feature matching and large scale indexing techniques at a fine scale, a variety of visual/non-visual clues (such as text, street signs, landmarks, specific architecture) can be used for determining the location at an intermediate scale, for example at the level of specific countries or certain county regions (urban, rural). At the coarsest scale, broader image/video categories can be determined and correlated with various geographical locations based on whether they have been taken in urban areas, suburban areas, mountainous landscape, etc. The following is a non-exhaustive list of visual cues that could be exploited for location detection:

- Visual landmarks: "Eiffel tower" or "Berlin Reichstag", architecture styles, structure and color of buildings

- Landscapes: Mountain and river shapes, desert illuminations, sand color, street shapes, urban/non-urban
- Written text: Recognition of character-types, language recognition, word recognition (e.g., street names), localized information (e.g., how dates and times are expressed)
- Signs: Traffic signs, car license plates
- Lighting: Indoor/outdoor, night/day, weather, position of the sun (related to time stamp of the video)

For written text recognition, it is well known that state-of-the-art video OCR methods can be applied to cellphone imagery; coarse illumination detection and direction estimation (e.g., for time-of-day constraints on location) may also be feasible – this approach is especially appealing when rich camera metadata is available in the image file (see below).

5.2 Acoustic Location Estimation

A similar taxonomy of acoustic cues is available to infer location. At the scale of a city, speech recognition of named entities and environmental sound classification, such as the presence or absence of car sound or the presence or absence of noise produced by a crowd, will help to determine location. For example, a farmers' market might include car noise in the background, crowd noise, and spoken words such as the names of fruits and vegetables. At an intermediate scale, dialect identification, as well as noise classification (police siren, bird calls) could be very useful. At a large scale, language and localized information (what are the units for dates, times, distances, volumes, mass, temperature?) are among the cues that will contribute to an overall confidence score. Acoustic landmarks, such as the sound of London's Big Ben or the playing of the UC Berkeley Campanile, should be among the top providers of a high-confidence level at all scales. The following is a non-exhaustive list of acoustic cues that could be exploited for location estimation:

- Acoustic landmarks: Specific church bell, specific reverberation inside a certain building, 50/60Hz power hum
- Recorded noise: Cars/no-cars, police car siren types, birds, water flowing, crowd noise
- Recorded speech: Language and dialect identification, word recognition of named entities, recognition of directions
- Environments: Jungle (fauna), street noise (frequency and types of vehicles), urban/non-urban (acoustically), airport proximity, room shape through reverberation

5.3 Metadata-based Estimation

Internet multimedia repositories such as YouTube, Flickr, and WikiMedia Commons, store (sometimes exhaustive) accompanying metadata close to the media object. The metadata might sometimes contain the actual location or a vague descriptions of it (e.g. "Berlin" or "USA"). Of course, metadata description might be wrong and then location verification needs to be developed (see Section 2). Other metadata might indirectly give hints to a possible location, including:

- Words used: Terms used to describe the video might clearly indicate locations, such as landmarks, localized information, street and city names
- Language used: Combinations of words together with specific language can identify location, e.g. a Finish description of finish traffic laws is most likely pointing to a video in Finland
- Relative location is often implicitly described in metadata, e.g. garden party, will most likely point to an outdoor video as do activity words such as "sailing", "driving", "boating".

In addition, embedded metadata, such as EXIF might be helpful even if geo-coordinates are not present: Indoor and outdoor camera settings, time and date, and other specific information might be able to limit the search domain further. Also, GPS coordinates, even when embedded are often only embedded with a certain accuracy and might be refined using location estimation.

5.4 Multimodal Integration

As described previously, location detection is inherently multimodal since the output of individual classifiers will often only result in vague assumptions. Given a video, a typical output would consist of a bag of categories and their associated probabilities. Example output could have the following structured form:

1. Outdoor: 70 %,
2. Urban area: 80 %,
3. Language: East German dialect: 35 %,
4. Landmark similarity to Brandenburg Gate: 35 %,
5. Recording channel: amateur camera 70 %

In order to enable fully automatic location estimation, i.e. in order to interpret the bag of categories and probabilities, an appropriate scheme for multimodal integration is a key challenge in this approach. Traditional schemes for "late fusion" (see for example [6]) may be inappropriate, as the specific set of candidate locations may not be obvious a-priori, and/or there may be an extremely large number of them, rendering a classic product or sum late fusion inaccurate. On the other hand, it is unfeasible to adopt a pure early fusion approach, as the image and video measurements come from distinct spaces with differing observation properties; a naive concatenation of features from different modalities will likely be biased inappropriately to one modality or the other. The multimodal location estimation problem is interesting and somewhat unique in that the fusion required can change depending on the situation: When a Boston accent is heard and a Boston landmark image is observed, our confidence of the video being in Boston should be high. However, the presence of a German voice is not necessarily a significant negative, as it may well be the voice of the tourist. So fusion schemes must amplify when there is agreement, but when there is disagreement, it may be appropriate to maintain distinct location estimates to fill different "roles" in the video interpretation.

6. TRAINING DATA

A major distinguishing point of this task as proposed is the availability of directly useable training data "in the wild". In 2006, our planet hosted about two billion cell phones of which about 50 million had a built-in video camera. As these numbers grow, more and more videos are uploaded to the Internet for public access on sites like YouTube, Flickr, and Liveleak. For a significant amount of these data, corresponding geolocations in the form of GPS coordinates exist. This represents a massive amount of annotated training data for the task that can be taken from the Internet, i.e. there is no need for explicit recording and hand-annotation. In this Section, we discuss experiments, also presented in [5], that quantify our claim about the availability of geo-tagged data.

6.1 Background

The most common mechanism to associate locations with photos are *EXIF* records, which were originally introduced by the *Japan Electronic Industry Development Association* for attaching metadata to images such as exposure time and color space. Since then EXIF has been extended to also cover geographical coordinates in the form of latitude and longitude. Currently, EXIF is used only with JPEG & TIFF (image) and WAV (audio) files. However, most other multimedia formats can contain metadata as well, often including geo-tags. In addition, most camera manufacturers specify proprietary metadata formats. For videos, these "maker notes" are the most common form for storing locations. Both Flickr and YouTube have comprehensively integrated geolocation into their infrastructure, and they provide powerful APIs for localized queries. Leveraging these APIs, we can estimate the number of public geo-tagged photos/videos they offer.

6.2 Flickr

Flickr's API allows to directly query for the number of images that are, or are not, geo-tagged during a certain time interval. Examining all 158 million images uploaded during the first four months of 2010, we found that about 4.3% are geo-tagged. We also examined the brands of cameras used for taking the photos that have geo-information, derived from their EXIF records which can be retrieved via Flickr's API as well. Doing so however requires one API request per image, and hence we resorted to randomly sampling a 5% set of all geo-tagged images uploaded in 2010. We found that the top-five brands were Canon (31%), Nikon (20%), Apple (6%), Sony (6%), and Panasonic (5%). A closer look at the individual models reveals that today mostly devices at the higher end of the price scale are geo-tagging. Historically, it has often been observed that high-end models become the commonly used one and their features become standard even for the lower end at some point in time. We therefore think that the amount of geo-tagged information is going to accumulate rapidly.

6.3 YouTube

With YouTube, due to restrictions of the API, it is not possible to directly determine the number of geo-tagged videos, as we could with Flickr. YouTube restricts the maximum number of responses per query to 1,000; and while it also returns an (estimated) number of total results, that figure is also capped at 1,000,000. Furthermore, the granularity for time-based queries is coarse: YouTube only allows to specify the attributes `all_time`, `this_month`, `this_week`, and `today`. Still, we believe we can estimate the number of geo-tagged videos in the following way: We submitted an unconstrained query, which results in an estimation of 1,000,000 results. The query was then refined by filtering for all videos that contain geo-location. Repeating the experiment a number of times resulted in total result estimates ranging from about 30,000 to 33,000 videos. In other words, out of what we assume to be a random sample of 1,000,000 YouTube videos, roughly 3 % have geo-location. While this number is clearly just an estimate, it matches with what we derived for Flickr. A note: YouTube's API distinguishes between videos *without* location, with *coarse* location (usually manually added, e.g. "Berlin"), and with *exact* location. For our experiments, we only considered the latter.

If one takes YouTube and Flickr as two samples representative of the Internet one can say that about 3 % of the consciously uploaded multimedia in the Internet is geo-tagged. Of that, many media might not be useful as training data for location estimation because of manual editing, dubbed music, or simply because they do not contain any clues. However, given the accumulation effect of persistent storage and the increasing number of geo-enabled capturing devices, even if only 1 % of the entire geo-tagged multimedia on the Internet is useable, this represents a training set of never before-seen magnitudes in the field. Also, we clearly see that location estimation is needed as at least 97 % of all videos and photos are not yet location enabled.

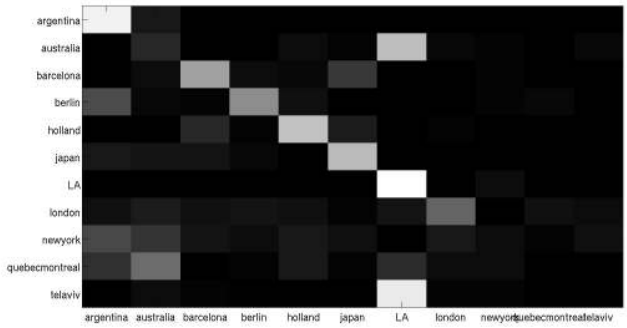

Figure 2: Confusion matrix of our GMM/SVM audio-based ambulance classifier (described in Section 7).

7. A FIRST EXPERIMENT

This section exemplifies an ambulance classifier that has been created as an initial approach towards multimodal location estimation.

7.1 Input Data

As a first task, we considered a scenario that would be a common case for city-level location estimation: the classification of distinctive objects commonly found in cities, and, as an initial detailed case study, we focused on the classification of ambulances. Therefore, we collected 200 YouTube videos filmed in 11 cities, manually chosen to contain an ambulance. The data is inherently challenging as it derives from real users and is not recorded under controlled condi-

tions. Our first task towards understanding location detection is thus limited to classifying which city an ambulance comes from. The amount of data we have collected so far is small, making the training of models challenging. Furthermore, some cities do not have enough data, and thus we had to make some classes broader than a city (e.g. Argentina, or Quebec/Montreal area).

7.2 Methods

The first system that we considered contained only audio information. Given the nature of the data, we expected this system to perform significantly better than chance on video data. We extracted 19-dimensional Mel frequency cepstrum coefficients (MFCCs). A Gaussian mixture model (GMM) was trained on a per-city level on the acoustic feature space. Classification based on likelihood was performed on an independent set of videos. The split between training and testing data was 70% and 30% respectively. Besides this generative approach baseline, we also considered SVM classification on the Gaussian mixture space (a system with state-of-the-art performance on the speaker identification task[2]).

Since the audio features we extracted are optimized for speech recognition, they may be a poor match to our data as, *a priori*, ambulance sound is quite different from natural speech. Thus, we created another baseline system based on vector quantization: we form a codebook of 20 clusters using k-means on the MFCC feature space, and extract the histogram of these feature occurrences on a per video basis, similar to bag of words (BoW) approaches that are typically used in natural language and computer vision [1]. The histogram obtained is used as the observation vector for training a Support Vector Machine (SVM) classifier.

Lastly, we extracted features based on color SIFT [11] on a uniform grid on each frame in the videos. A codebook of 1000 clusters is then extracted, and histogram features are extracted and fed into a SVM classifier similarly as in the previously described procedure.

For the fusion systems, we performed both early and late fusion as baselines for multimodal processing. In the early fusion system, we concatenated the features prior to the SVM classifier. For the late fusion, we used the SVM classifier scores and fed them as features for a third SVM, as described in [13].

7.3 Results

Table 1 shows the accuracies of the various systems, as well as what a random classifier would output (since all the classes are balanced, chance would give us an accuracy of $\frac{1}{11}$). We see that even the GMM model performs significantly better than chance, even though it is wrong more than half the time. The simpler bag of words system performed worse than the GMM approach, which leads us to conclude that GMM based clustering for audio data is better than simpler k-means (albeit slower). It is worth noting than on a smaller development dataset containing only three cities, the BoW approach performed better with the same number of clusters. It appears that more than 20 clusters may be necessary for the more complex classification task, and thus other clustering techniques that scale better with number of clusters and samples should be used. The obtained multimodal results favor the early fusion scheme, although the performance is dissimilar for both modalities, making multimodal combination more challenging and a topic to further work on in the future.

System	Accuracy
Random	9.1%
GMM (audio)	45.20%
GMM SVM (audio)	47.72%
BoW SVM (audio)	35.5%
BoW SVM (video)	23.1%
BoW early fusion SVM (audio+video)	37.5%
BoW late fusion SVM (audio+video)	36.9%

Table 1: Results on the testing set for the ambulance detection task on a set of 11 cities/regions. See Section 7 for details.

Other lines of future work could include the training of purified models. This can be achieved by means of temporal clustering to avoid fitting non informative frames in the video (e.g. when someone is speaking on top of an ambulance sound, or when the ambulance sound is not present). Different clustering techniques other than k-means or finite mixture models for codeword generation could be explored, such as Latent Dirichlet Allocation [1] or Dirichlet Process mixture models [4], and features other than MFCC or SIFT will be explored as we gain more knowledge on which aspects of data classification are challenging.

Interestingly, our classifier has significantly different performance across cities. As can be seen in Figure 2, the best performing cities/regions are Argentina, Barcelona, Berlin, Holland, Japan, LA, and London. Australia and Telaviv get confused with LA, partially due to the fact that there are several ambulance companies operating in LA, which may cause the class to be too broad. Quebec/Montreal and New York get confused with Argentina and Australia, and we cannot explain this behavior. It is worth noting that, even though the classifier based on BoW features had worse overall accuracy, the behavior per city was more uniform. Again, an indication of how hard it is to work with heterogenous data from YouTube.

8. FINAL REMARKS

This article describes a new research problem, possible directions for tackling it, and our initial work in the field. While at first glance it is almost impossible, and indeed for many media unsolveable, the multimodal location estimation task offers research opportunities in many fields connected to multimedia. As the solution can be mostly described as a search for cues, the task is inherently multimodal. With the large amounts of training data available on the Internet, the task offers a chance to tackle machine learning problems using more and more heterogeneous input, which in turn might lead to better understanding and more generalizable solutions. Therefore, we want to encourage multimedia researchers to actively engage in the tasks involved and create a brand new community working on a very challenging but exiting problem. We want to encourage readers to contact us and visit our project website `http://mmle.icsi.berkeley.edu`, where we will post updates on our progress and, more importantly, continuously develop publicly available training and test sets for benchmarking.

Acknowledgments

This research is supported by an NGA NURI grant #HM11582-10-1-0008.

9. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[2] 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.

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

[4] J. M. David, D. M. Blei, and M. I. Jordan. Nonparametric empirical bayes for the dirichlet process mixture model. *Statistics and Computing*, 2006, 2004.

[5] G. Friedland and R. Sommer. Cybercasing the Joint: On the privacy implications of geo-tagging. *Proceedings of the Fifth USENIX Workshop on Hot Topics in Security (HotSec 10)*, Washington, D.C., August 2010.

[6] G. Friedland, O. Vinyals, Y. Huang, and C. M "Uller. Prosodic and other Long-Term Features for Speaker Diarization. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(5):985–993, 2009.

[7] J. Hays and 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.

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

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

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

[11] D. Lowe. Object recognition from local scale-invariant features. In *Proceedings of ICCV*, pages 1150–1157, 1999.

[12] 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.

[13] C. Snoek, M. Worring, A. Smeulders Early versus late fusion in semantic video analysis. In *Proceedings of ACM Multimedia*, pages 399–402, 2005.

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