City-Identification on Flickr Videos Using Acoustic Features

Howard Lei[†], Jaeyoung Choi[†]*, and Gerald Friedland[†]

TR-11-001

April 2011

Abstract

This article presents an approach that utilizes audio to discriminate the city of origin of consumer-produced videos – a task that is hard to imagine even for humans. Using a sub-set of the MediaEval Placing Task's Flickr video set, we conducted an experiment with a setup similar to a typical NIST speaker recognition evaluation run. Our assumption is that the audio within the same city might be matched in various ways, e.g., language, typical environmental acoustics, etc., without a single outstanding feature being absolutely indicative. Using the NIST speaker recognition framework, a set of 18 cities across the world are used as targets, and Gaussian Mixture Models are trained on all targets. Audio from videos of a test set is scored against each of the targets, and a set of scores is obtained for pairs of test set files and target city models. The Equal Error Rate (EER), which is obtained at a scoring threshold where the number of false alarms equals the misses, is used as the performance measure of our system. We obtain an EER of 32.3% on a test set with no common users in the training set. We obtain a minimum EER of 22.1% on a test set with common users in the training set. The experiments show the feasibility of using implicit audio cues (as opposed to building explicit detectors for individual cues) for location estimation of consumer-produced “from-the-wild” videos. Since audio is likely complementary to other modalities useful for the task, such as video or metadata, the presented results can be used in combination with results from other modalities.

[†] International Computer Science Institute, 1947 Center St., Ste. 600, Berkeley, CA 94704
[*] University of California, Berkeley, EECS Department, Berkeley, CA 94720

This work was partially supported by funding provided through the National Geospatial-Intelligence Agency University Research Initiatives program (NGA NURI). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors or originators and do not necessarily reflect the views of NGA.

1. INTRODUCTION

With more and more multimedia data being uploaded to the web, it has become increasingly attractive for researchers to build massive corpora out of "from-the-wild" videos, images, and audio files. While the quality of consumer produced content on the internet is completely uncontrolled, and therefore imposes a massive challenge for current highly-specialized signal processing algorithms, the sheer amount and diversity of the data also promises opportunities for increasing the robustness of approaches on an unprecedented scale. Moreover, new tasks can be tackled that couldn't even be attempted before, even tasks that couldn't easily be solved by human subjects. In the following article, we present the task of city identification using the audio tracks of a random sample of Flickr videos.

Using only the audio tracks from the videos, the NIST speaker recognition [11] framework is used to perform the experiments, where audio from videos in a test set are scored against pre-trained city models using audio from city-labeled videos in a training set. While we discard much of the information in the videos by using only the audio, our approach provides a simple way to use well-established techniques such as the speaker recognition system, and reduces computational requirements.

Using only audio information also gives us insight into the extent to which city-scale geo-locations of videos is correlated with their audio features. Listening to the audio tracks of a random sample of videos, we do not find any city-specific sounds that would enable a human listener to accurately perform city identification of the videos. Hence, this work demonstrates the power of the machine learning algorithm in performing a task that would ordinarily be difficult for humans. Because the audio modality is likely complementary to the video and metadata modalities, achieving success using only the audio modality suggests the potential for further improvements when different modalities are combined together.

While speaker recognition evaluations usually follow strict guidelines concerning the quality and the channel of the recording, the experiments described here use random videos, which contain audio tracks with a large variance in length, content, and quality. Nevertheless, the results of our experiments are far from random. According to the Equal Error Rate (EER) measure used in a speaker recognition system, a scoring threshold can be set such that the vast majority of the scores for which a test set video is correctly classified at the city-scale are above the threshold, and the vast majority of the scores for which a test set video is incorrectly classified at the city-scale are below the threshold.

This article is structured as follows: Section 2 presents related work; section 3 describes the publicly available dataset; section 4 describes the technical approach used for the experiment; section 5 describes the results; section 6 discusses the implications of the results, and section 7 presents a conclusion and outlook to future work.

2. RELATED WORK

Recent articles [13, 14] indicate initial results already showing 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 images are the same then the locations must be the same too. In other recent work [5], the goal is to estimate a rough location of an image as opposed to its exact GPS location. For example, images of certain landscapes can occur only in certain places on Earth. Jacobs' system [7] relies on matching images with satellite data. The above work (along with other work) relies on the detection or matching of a set of explicit visual features (e.g. landmarks or sun altitudes) rather than performing an implicit matching of unknown cues as performed in this article.

Multimodal location estimation, in contrast to visual location estimation, was first defined and attempted in [4] where the authors match ambulance videos from different cities. The first evaluation on multimodal location estimation was performed in the 2010 MediaEval Placing task [8]. While the accuracies achieved there were better than city-scale, none of the systems presented at MediaEval used audio and all systems relied on textual (metadata) cues.

3. DATASET

The audio tracks for the experiments are extracted from videos distributed as a training data set for the Placing Task of MediaEval 2010 [10], a multimedia benchmark evaluation. The Placing Task involves automatically estimating the location (latitude and longitude) of each test video using one or more of: metadata (e.g. textual description, tags), visual/audio content, and social information.

The data set consists of 5125 Creative Commons licensed Flickr videos uploaded by Flickr users. Flickr requires that a video 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 leads us to initially conclude that most visual/audio contents of the videos lack reasonable information for estimation of their origin [2]. For example, some videos are recorded indoors or in private spaces such as the backyard of a house, which makes the Placing Task nearly impossible if we examine only the visual and audio contents. 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.

From an examination of 84 videos from the data set, we find that most of videos' audio tracks are quite "wild". Only 2.4 % of them were recorded in a controlled environment such as inside a studio at a radio station. The other 97.6 % are home-video style with ambient noise. 65.5 % of the videos have heavy ambient noises such as crowds chatting in the background, traffic noise, wind blowing into microphone, etc. 14.3 % of the videos contain music, either played in the background of the recorded scene, or inserted at the editing phase. About 50 % of the videos do not contain any form of human speech at all, and even for the ones that contain human speech, almost half are from multiple subjects and crowds in the background speaking to one another, often at the same time. 5 % of the videos are edited to contain changed scenes, fast-forwarding, muted audio, or inserted background music. While there are some audio features that may hint at the city-scale location of the video – features such as the spoken language in cases where human speech exist, type and genre of music, etc – such factors are not prevalent, and are often mixed with heavy amounts of background noise and music.

For the task of city identification, all videos in the data

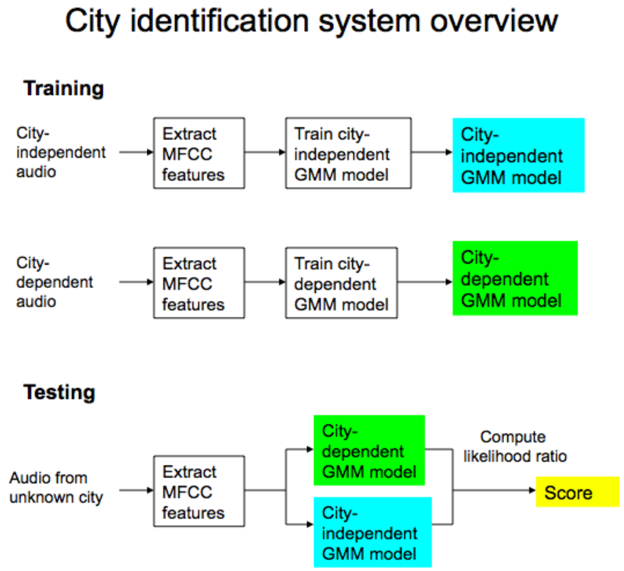

Figure 1: A diagram of the city identification system as described in Section 4.

set have been labeled by their city location according to the video's metadata, and a video is considered to be located within a city if it's geo-coordinates are within 5 km of the city center.

4. TECHNICAL APPROACH

The city identification system is derived from a GMM-UBM speaker recognition system [12], 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) [3]. The GMM-UBM system is a widely-used approach to speaker recognition, and is easy to implement. Specifically, for each audio track, a set of MFCC features are extracted and one Gaussian Mixture Model (GMM) is trained for each city, using MFCC features from all its audio tracks (i.e. city-dependent audio tracks). 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 (i.e. city-independent audio) [12]. For testing, the log-likelihood of MFCC features from the audio tracks of each test video are computed using the pre-trained GMM models of each city. A total of 128 mixtures are used for each GMM. Figure 1 illustrates the GMM-UBM system.

A likelihood score for each test video as a match to each of the cities is obtained as follows: 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. During scoring, a threshold is established for distinguishing the true trial scores from the impostor trial scores. The system performance is based on 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. The open-source ALIZE speaker recognition system implementation is used [1], and the 60-dimensional MFCC features are obtained via HTK [6].

Table 1: Results of different runs of the city identification system, as explained in Section 5

Training set	Testing set	Common users	EER (%)
trn_all	*tst*	No	32.3
trn_s1	*trn_s2*	Yes	22.6
trn_s2	*trn_s1*	Yes	**22.1**
trn_s1	*tst*	No	31.0
trn_s1noSFLon	*tst_noSFLon*	No	37.8
trn_s1rand	*tst*	No	46.4

5. EXPERIMENTS AND RESULTS

Experiments are run using the GMM-UBM system to obtain city identification results. The entire duration of each audio track is used, and MFCC features are mean- and variance-normalized prior to GMM training. We've performed a series of experiments examining different combinations of data used for training and testing. Our main experiment uses 1,080 videos in the training set (denote as *trn_all*) and a 285-video test set (denote as *tst*) with no common users in the training set. The 285-video test set gives 5,130 trials (with 285 true trials). The reason for not having common users in the training and test sets is that previous work showed that one can match videos of the same user with better-than-chance-accuracy based on the audio track of a video [9], and we thus want to eliminate this factor.

We've also performed experiments examining the effect of having common users for the training and test set videos, having randomized city labels, and removing the two cities (San Francisco and London) with roughly half the videos in the training and test sets. Removing the two cities results in the videos being roughly uniformly distributed across approximately 70% of the cities. We've created two random splits of the training set, with 542 videos in split 1 and 541 in split 2. Among the $542 \times 541 = 293,222$ pairs of videos across both splits, 3,967 pairs (1.35 % of total pairs) have the same user. One experiment uses split 1 (denote as *trn_s1*) for UBM and city model training, and split 2 (denote as *trn_s2*) for testing, with a total of 9,738 trial scores (539 true trial scores). A second experiment uses split 2 for UBM and city model training, and split 1 for testing, with a total of 9,756 trials (434 true trial scores). A third experiment uses split 1 for UBM and city model training, and the 285-video test set (denote as *tst*) – with no common users in split 1 of the training set – for testing. A fourth experiment uses split 1 of the training videos and the 285-video test set with San Francisco and London removed from both training and testing (denote as *trn_s1noSFLon* and *tst_noSFLon* respectively). A fifth experiment uses split 1 with randomized city labels (denote as *trn_s1rand*) for UBM and city model training, and the 285-video test set. This last experiment should approximate random chance. Table 1 shows all results.

Our main experiment (using the training and test sets *trn_all* and *tst* respectively) gives a 32.3 % EER. A Detection Error Tradeoff (DET) curve (which plots the false alarm vs. miss rates at different scoring thresholds) for this experiment is shown in Figure 2. We can also obtain a measure of raw accuracy for this experiment by setting the scoring threshold

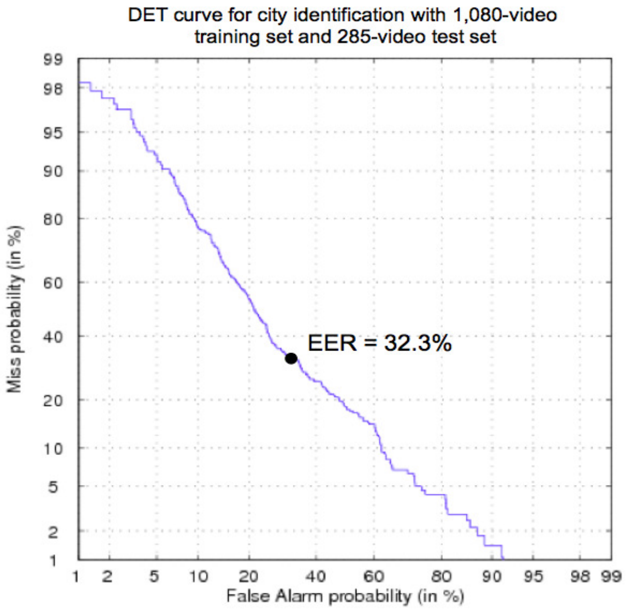

Figure 2: The DET curve of the results described in Section 5.

to the level for the EER, and simply tallying the number of trials that are correctly identified (i.e. true trial scores that fall above the scoring threshold, and impostor trial scores that fall below the threshold). 67.7 % (193 out of 285) of the true trial scores are correctly classified; 67.7 % (3279 out of 4845) of the impostor trial scores are correctly classified. To demonstrate the consistency of this result, we created 100 random splits amongst the 5,130 trial scores, with roughly 2,500 trial scores (roughly 128 true trial scores) per split, and computed the EER on each of the 100 splits. The average of the EERs is 32.4 %, with a standard deviation of 1.16 %. The fact that the EER average is close to the original EER (32.4 % vs. 32.3 %), along with the low standard deviation, shows the consistency of our EER result across all trials.

Results for other experiments demonstrate up to a 28.7% relative EER improvement (31.0% EER vs. 22.1% EER) if the training and test sets have common users. This demonstrates that implicit effects, such as channel artifacts from the recording device, contribute significantly to the accuracy. When removing San Francisco and London (the cities with roughly half the videos), we obtain a 37.8% EER, and with randomizing the training city labels, we obtain a 46.4% EER (the result obtained by random chance). Note that while the result when San Francisco and London are removed (37.8% EER) is worse than the corresponding result where the cities are included (31.0% EER), the result is still significantly better than the random chance result (46.4% EER). One possible reason why including San Francisco and London significantly improves results is the availability of more training data for its city models. This shows that overall results would likely improve even more if additional training data is used for other cities as well. Overall, the results demonstrate the feasibility of using the audio tracks of videos to identify their city of origin.

6. DISCUSSION

Audio is one of several possible media to use for this task and is likely complementary to other modalities, such as video and textual metadata. Hence, potential improvements in city identification can be obtained by combining audio with other media. It is likely that the reason audio performs well in this task, even by itself, is that different cities have different types of noises, music, languages, as well as loudness levels at different times during the day. The GMM models of each city may have learned such distinctive features of each city, enabling our system to perform reasonably.

Our result is even more interesting considering that even after listening to a random sample of the videos across different cities, we did not get the sense that there are any clear, distinctive audio features for each city. For instance, there are no sounds that would clearly identify audio as belonging to the city of San Francisco. We suspect that it would be difficult for humans to perform the same task using the same experimental setup. Hence, we believe that the GMM-UBM approach may well be better than humans at performing city identification of videos based on their audio.

7. CONCLUSION AND FUTURE WORK

This work demonstrates the applicability of a GMM-UBM speaker recognition system to city-identification of Flickr videos, based only on audio information. Moreover, it shows the feasibility of using implicit audio cues (as opposed to building explicit detectors for individual cues) for location estimation of consumer-produced, "from-the-wild" videos. Therefore, an EER of 32.3% (meaning that among all the trials, we obtain a 32.3% false alarm and a 32.3% miss rate) on a test set of 285 videos, with no common users in the training set (which is a self-imposed constraint to eliminate direct recording device matching), is a significant result, and is far from our random baseline result (46.4 % EER). For test sets with common users in the training set, we obtain an EER as low as 22.1%. Our result is rather surprising, given that human observation of the audio tracks of the videos has not revealed any distinctive characteristics for whether a video is from one city versus another. However, a conglomeration of factors, such as differences in music, language, loudness, among others, may have been taken into account by our machine learning approach.

We also understand that audio is only one of the modalities that can be used. Future work may involve improving our system to better handle the audio modality, as well as incorporating other modalities such as video and metadata to this task. We suspect that, because audio is likely complementary to the other available modalities, using these other modalities as well will result in considerable improvements to the initial results presented here.

8. ACKNOWLEDGMENTS

This work was partially funded by an NGA NURI grant.

9. REFERENCES

[1] J. F. Bonastre, F. Wils, and S. Meignier. ALIZE, a free Toolkit for Speaker Recognition. *ICASSP*, 1:737–740, 2005.

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

[3] S. Davis and P. Mermelstein. Comparison of Parametric Representations of Monosyllabic Word Recognition in Continuously Spoken Sentences. *ICASSP*, 1980.

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

[5] 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.

[6] HMM Toolkit (HTK). *http://htk.eng.cam.ac.uk*.

[7] N. Jacobs, S. Satkin, N. Roman, R. Speyer, and R. Pless. Geolocating Static Cameras. In *IEEE international conference on computer vision*, 2007.

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

[9] H. Lei, J. Choi, A. Janin, and G. Friedland. User Verification: Matching the Uploaders of Videos across Accounts. *accepted to IEEE ICASSP*, 2011.

[10] MediaEval Web Site. *http://www.multimediaeval.org*.

[11] NIST Speaker Recognition Evaluation. *http://www.itl.nist.gov/iad/mig/tests/sre/*.

[12] D. A. Reynolds, T. F. Quatieri, and R. Dunn. Speaker Verification using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10:19–41, 2000.

[13] 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.

[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.