MULTIMODAL CITY-VERIFICATION ON FLICKR VIDEOS USING ACOUSTIC AND TEXTUAL FEATURES

Howard Lei[1], *Jaeyoung Choi*[1,2], *and Gerald Friedland*[1]

[1]International Computer Science Institute
1947 Center Street, Suite 600
Berkeley, CA 94704, USA

[2]University of California, Berkeley
Dept. of EECS
Berkeley, CA 94720, USA

ABSTRACT

We have performed city-verification of videos based on the videos' audio and metadata, using videos from the MediaEval Placing Task's video set, which contain consumer-produced videos "from-the-wild". 18 cities were used as targets, for which acoustic and language models were trained, and against which test videos were scored. We have obtained the first known results for the city verification task, with an EER minimum of 21.8%, suggesting that ∼80% of test videos, when tested against a correct target city, were identified as belonging to that city. This result is well above-chance, even as the videos contained very few city-specific audio and metadata features. We have also demonstrated the complementarity of audio and metadata for this task.

Index Terms— City verification, acoustic models, N-gram language models, multimodal processing

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 videos "from-the-wild", 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 be attempted before, even tasks that couldn't easily be solved by humans.

In this article, we present the task of city verification, where we attempted to determine the likelihood of a set of cities being the city from which a Flickr video is taken, without relying on the video's geo-tagged GPS location. This task is a subset of the related task of location estimation. Recent articles [17, 19] on location estimation of images have indicated that the task may be approached as a retrieval problem on a location-tagged image database. In the work of [6], the goal was to estimate a rough location of an image as opposed to its exact GPS location. For example, images of certain landscapes could occur only in certain places on Earth. Jacobs' system [8] relied on matching images with satellite data. The above work (along with other work) relied 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. Works have also been performed in the most recent MediaEval Placing task evaluation [15], where participants must obtain the geo-location of Flickr videos based on textual metadata, video, and audio. While the accuracies achieved there were better than city-scale, audio usage had been virtually ignored in all systems. Mertens et al. [14] performed work on acoustic event classification using audio features, in a framework similar to this work.

Based on such previous work, and our areas of expertise, we approached the city-verification task using the audio tracks and textual metadata of the Flickr videos. While much of the information in the videos was discarded by using only the audio and metadata, our approaches demonstrated the cross-domain adaptability of well-established techniques such as acoustic and language modeling. Audio and/or metadata from videos in a test set were scored against pre-trained city models from city-labeled videos in a training set (identical to the NIST speaker recognition framework).

Using the audio information also gave us insight into the extent to which city-scale geo-locations of videos are correlated with their audio features. Listening to the audio tracks of a random sample of videos, we rarely found city-specific sounds that would enable a human listener to accurately perform city verification of the videos based on their audio. In addition, while location and/or language-specific metadata tags could sometimes be found in the Flickr videos to aid in the city-verification task, the useful metadata was sparse. Hence, this work demonstrated the power of machine learning algorithms in performing a task that would likely be difficult for humans. Because the audio and metadata modalities are likely complementary to other modalities (i.e. video), achieving success using only the audio and metadata modalities would suggest the potential for further improvements when additional modalities are incorporated.

The experiments described here used videos, whose audio tracks had large variances in length, content, and quality, and whose metadata were often not indicative of the video's location. This is in contrast, for example, to the NIST speaker recognition task, where the audio usually follow strict guidelines concerning its quality and recording channel.

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

2. DATASET

2.1. Characteristics

The audio tracks for the experiments were extracted from videos distributed as a training dataset for the Placing Task of MediaEval 2010 [13], a multimedia benchmark evaluation. The dataset consists of 5125 Creative Commons licensed Flickr videos uploaded by Flickr users. Manual inspection of the dataset led us to conclude

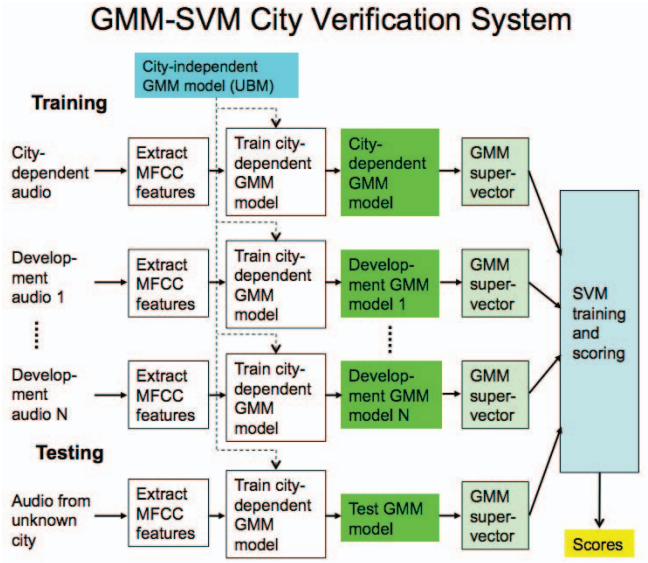

Fig. 1. A simplified view of the GMM-SVM city verification system as described in Section 3.

that most visual/audio contents of the videos lack reasonable information for estimation of their origin. For example, some videos have been recorded indoors or in private spaces such as the backyard of a house, which makes the Placing Task nearly impossible if only the visual and audio contents were examined. This indicates that the videos have not been pre-filtered or pre-selected in any way to make the dataset more relevant to the city-verification task.

From an examination of 84 videos from the dataset, we 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. The other 97.6 % are home-video style with ambient noise. 65.5 % of the videos have heavy ambient noises; 14.3 % of the videos contain music. 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. 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. The maximum length of Flickr videos is limited to 90 seconds. About 70 % of videos are less than 50 seconds.

For the task of city verification, a video was considered to be located within a city if its geo-coordinates were within 5 km of the city center. The following cities were considered for verification, because of the predominance of videos belonging to these cities: *Bangkok*, *Barcelona*, *Beijing*, *Berlin*, *Chicago*, *Houston*, *London*, *LosAngeles*, *Moscow*, *NewYork*, *Paris*, *Praha*, *Rio*, *Rome*, *San Francisco*, *Seoul*, *Sydney*, *Tokyo*.

3. TECHNICAL APPROACHES

We explored various approaches to city verification. Because of the lack of prior work for the city verification task, there were no effective previously-developed technical approaches for the task. Hence, we decided to approach the city verification task using well-established acoustic modeling-based approaches (i.e. audio-based approaches), as well as approaches using language models built from the metadata of the Flickr videos. The first audio-based approach was derived from the GMM-UBM speaker recognition system [16], 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) [5]. Specifically, for each audio track, a set of MFCC features was extracted and one 128-mixture Gaussian Mixture Model (GMM) was trained for each city, using MFCC features from all audio tracks for the city in the training set. This was done via MAP adaptation from a universal background GMM model (UBM), which was trained using MFCC features from all audio tracks of all cities in the training set [16]. During testing, the log-likelihood score of MFCC features from test video's audio track was computed for each city-dependent GMM model. Scores for which the city of the test video matched the city of the GMM model were known as true trial scores; scores for which the cities do not match were known as impostor trial scores. The GMM models were trained using the open-source ALIZE toolkit [2], and the MFCC features were obtained via HTK [7].

The second audio-based approach was derived from the GMM-SVM speaker recognition system [3]. In this approach, the same feature extraction was used as in the GMM-UBM approach. A separate GMM model was trained using the audio of each video, via MAP adaptation from a UBM, and the GMM mean parameters were collected into a supervector. Hence, there was one supervector for each video. An SVM model was 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 development data as negative training examples. A classification score for the supervectors of each test video was obtained for the SVM models of each city. The SVMs were implemented using the SVM^{light} toolkit [9], with wrapper scripts from SRI. Figure 1 illustrates the GMM-SVM system.

The language-modeling based approach involved training back-off language models, implemented using the SRILM toolkit [18]. Uni-, bi-, and trigram word language models were trained for each city using the metadata (keywords and descriptions) of all videos for the city in the training data. The likelihoods of the metadata of test videos were then computed using each city's language model to determine classification scores of each test video versus each city.

During scoring, a threshold was established for distinguishing the true trial scores from the impostor trial scores. The system performance was 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.

4. EXPERIMENTS AND RESULTS

Experiments were run using the GMM-UBM, GMM-SVM, and Uni-, bi-, and trigram word language model systems to obtain city verification results. For audio-based experiments, the entire duration of each audio track was used, and MFCC features were mean- and variance-normalized prior to GMM training. Different combinations of data was used for training and testing. The main experiment used a 117-video development set, a 1,080-video training set (denote as *trn_all*), and a 285-video test set (denote as *tst*) with no common users in the training set. The city-specific distribution of videos in

Training set	Testing set	System	Common users	EER (%)
trn_all	*tst*	GMM-UBM	No	32.3
trn_all	*tst*	GMM-SVM	No	32.3
trn_s1	*trn_s2*	GMM-UBM	Yes	**23.0**
trn_s1	*tst*	GMM-UBM	No	31.0

Table 1. Results for the GMM-UBM and GMM-SVM audio-based approaches to city verification. The result for the experiment with common users between the training and test sets demonstrated greater city verification accuracy in terms of a lower EER (random EER being 50%) than the experiments without common users.

the 1,080-video training set was such that 43% of videos were from *San Francisco*, 17% were from *London*, and each remaining city had 7% or less of the total number of videos. The distribution in the 285-video test set was such that 25% of videos were from *San Francisco*, 22% were from *London*, and each remaining city had 7% or less of the total number of videos. The 285-video test set gave 5,130 trials (with 285 true trials).

Experiments were also performed examining the effect of having common users in the training and test set videos (previous work showed that one can match videos of the same user with better-than-chance-accuracy based the audio tracks [11]). To simulate the effect of having common training and test users, we 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) had the same user. Split 1 (denote as *trn_s1*) was used 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). We also used split 1 for training, and the 285-video test set (with no common users with split 1 of the training set) for testing. Table 1 shows results for the GMM-UBM and GMM-SVM audio-based approaches.

To combine the audio and metadata-based approaches at the score level, a Multi Layer Perceptron (MLP) with 2 hidden nodes and 1 hidden layer, implemented using Lnknet [12], was used. The EER results represented averaged EER values over 100 splits amongst the training cities and test videos, where each split contained training and testing sub-splits. For each of the 100 splits, MLP weights were trained using the training sub-split, and applied to the testing sub-split. The combination was done for experiments using *trn_s1* for training data and *trn_s2* for testing data. The EER averaging was done for all results using this training and testing data combination, even if only one system was used, so that standalone results would be consistent with combination results.

Table 2 shows these metadata-based results, along with its combination with the GMM-UBM approach. The Unigram LM, Bigram LM, and Trigram LM systems represent systems with uni-, bi-, and trigram city language models respectively. The GMM-UBM result is also shown for purposes of comparison.

According to the results in tables 1 and 2, the audio experiment using the training and test sets *trn_all* and *tst* respectively gave a 32.3 % EER, for both the GMM-UBM and GMM-SVM systems. Because the two systems gave statistically similar results, we used the GMM-UBM approach for all other experiments due to its computational efficiency over the GMM-SVM approach. Results for other GMM-UBM experiments demonstrated up to a 28.9% relative EER improvement (32.3% EER vs. 23.0% EER) if the training and test sets had common users (albeit on a different set of trials). This showed that implicit user-specific effects, such as channel artifacts

Training set	Testing set	System	Common users	EER (%)
trn_all	*tst*	Unigram LM	No	33.5
trn_s1	*trn_s2*	Unigram LM	Yes	23.9
trn_s1	*trn_s2*	Bigram LM	Yes	29.4
trn_s1	*trn_s2*	Trigram LM	Yes	30.9
trn_s1	*trn_s2*	GMM-UBM	Yes	25.3
trn_s1	*trn_s2*	GMM-UBM + Unigram LM	Yes	**21.8**

Table 2. Results of metadata-based approaches to city verification. The Uni- and Bigram LM metadata approaches were comparable in EER to the GMM-UBM audio-based approach. Combining the Unigram LM approach with the GMM-UBM approach gave a minimum EER of 21.8%.

from the recording device, and the user's preferred video-recording environment, contributed significantly to accuracy. Overall, the results demonstrated the feasibility of using the audio tracks of videos to identify their cities of origin.

The metadata experiments gave surprisingly similar results compared to the audio experiments. For training set *trn_all* and test set *tst*, the Unigram LM approach gave a 33.5% EER (within 3.7% of the GMM-UBM result). For training set *trn_s1* and test set *trn_s2* the Unigram LM approach gave a 23.9% EER (averaged over the 100 splits), which was a 5.5% relative EER improvement over the GMM-UBM approach. Combining the Unigram LM and GMM UBM approaches resulted in a 21.8% EER, an 8.8% relative EER improvement over the Unigram LM standalone. This result suggested that almost 80% of the test videos, when tested against its correct target city, were correctly identified as belonging to that city. Note that using lower-order language models in general resulted in lower EERs. This was likely because most metadata keywords had no lexical connections with other keywords, and the description metadata (where connection exist) was short. Overall, there were an average of only 6.3 usable lexical tags per video in our dataset, such that higher-order language models were sparse and would likely result in overtraining. Given the sparsity of metadata information, it was surprising that the metadata-based approach performed similarly to the audio-based approaches.

5. DISCUSSION AND ANALYSIS

Our audio-based results are interesting considering that after listening to a random sample of the videos across different cities, we did not get the sense that there were any clear, distinctive audio features for each city. For instance, there were no sounds that would clearly identify audio as belonging to the city of San Francisco. However, a close listening to the test videos with the high true trial scores indicated that speech may play a significant role in city verification. Test videos with the top three true trial scores were all from Rio and contain monologue speech from a family excursion, where the words "Rio De Janeiro" are spoken. One high-scoring test video from London contains speech with British accents, while one from Paris contains city-specific ambulance noise. Many high-scoring videos hence appeared to contain some kind of city-specific audio feature (i.e. speech or language/dialect marker, or other city-specific noise).

However, there were also high-scoring test videos without city-specific audio features - a video from Tokyo contains audio of a train arriving, one from San Francisco contains bagpipe music, and one from Paris contains loud engine noise. Because it would likely be

difficult for humans to correctly classify these videos, we think that the GMM-UBM and GMM-SVM machine learning approaches may well be better than humans at performing city verification of videos based on their audio.

It would appear easier for humans to perform the same task using metadata. An analysis of high-scoring test videos for the unigram language model experiments indicated that many high-scoring test videos contain location- or langauge-specific metadata keywords. Such videos include one from Barcelona, with Spanish metadata *bicicletas*, *policia*, *brigada*, along with the location-specific word *barcelona*. A video from London contains the metadata *london*, one from San Francisco contains *sanfrancisco*, and one from Beijing contains *asia*, *china*, *tibetan*, and *buddha*. However, some high-scoring test videos do not contain any location or language-specific metadata, which would make them difficult for users to classify.

Because high-scoring test videos from the audio-based approaches differed from those for the metadata approaches, the two approaches are complementary, resulting in an EER improvement in their combination. Furthermore, potential improvements in city verification could be obtained by combining other modalities, such as video and keyframe image data, as well as making better use of the audio and metadata.

6. CONCLUSION AND FUTURE WORK

According to our knowledge, this work is the first attempt at geolocating consumer produced, "from the wild" videos at the city scale, and demonstrates the applicability and adaptability of standard GMM-UBM and GMM-SVM approaches, as well as language model-based approaches. The approaches utilize the videos' audio and metadata information. Our work shows the feasibility of using implicit audio cues (as opposed to building explicit detectors for individual cues) for location estimation of the videos. Therefore, an EER of 32.3% for the audio-based approaches on a test set of 285 videos, with no common users in the training set, is a significant result, and is far from random (50% EER). For test sets with common users in the training set, we obtained an EER as low as 21.8% (combining the audio and metadata modalities), suggesting that almost 80% of the test videos, when tested against its correct target city, were identified as belonging to that city. The unigram language model approach for classifying videos based on metadata is close in performance to the audio-based approaches, which is surprising given the metadata sparsity (average of 6.3 usable tags per video).

The results also indicate that machine learning approaches are arguably better than human performance for city-verification, given that we have been able to correctly classify some videos that would appear difficult for humans. A conglomeration of factors, such as differences in music, language, loudness, and broad metadata usage (including non language- and location-specific words), may have been taken into account by the machine learning approaches. Future work may involve improving our systems to better handle the audio and metadata modalities, and incorporating other modalities to enhance performance.

7. ACKNOWLEDGEMENTS

This research is supported by NGA NURI grant number HM11582-10-1-0008, NSF EAGER grant IIS-1138599, and NSF Award CNS-1065240. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the sponsors.

8. REFERENCES

[1] Amazon Mechanical Turk, http://www.mturk.com

[2] Bonastre, J.F., Wils, F., and Meignier, S., "ALIZE, a free Toolkit for Speaker Recognition", in ICASSP, Vol. 1, pp. 737–740 (2005).

[3] Campbell, W., Sturim, D., and Reynolds, D.,"Support Vector Machines using GMM Supervectors for Speaker Verification", in IEEE Signal Processing Letters, Vol. 13, pp. 308 – 311, 2006.

[4] Choi, J., Lei, H., and Friedland, G., "The 2011 ICSI Video Location Estimation System," in Proc. of MediaEval Workshop, 2011.

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

[6] Hays, J., and Efros, A.,"IM2GPS: Estimating Geographic Information from a Single Image", in IEEE Conference on Computer Vision and Pattern Recognition, 2008.

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

[8] Jacobs, N., Satkin, S., Roman, N., Speyer, R., and Pless, R.,"Geolocation Static Cameras", in IEEE International Conference on Computer Vision, 2007.

[9] Joachims, T., "Making Large Scale SVM Learning Practical", in Advances in kernel methods - support vector learning, B. Schoelkopf, C. Burges, and A. Smola, Eds. MIT-press, 1999.

[10] Kender, J., Hill, M., Natsev, A., Smith, J., and Xie, L., "Video Genetics: A Case Study from YouTube", in ICASSP, Vol. 1, pp. 737–740 (2005).

[11] Lei, H., Choi, J., Janin, A., and Friedland, G., "User Verification: Matching the Uploaders of Videos across Accounts", in Proc. of ICASSP, 2011.

[12] Lippmann, R.P., Kukolich, L.C., Singer, E., "LNKnet: Neural Network, Machine Learning, and Statistical Software for Pattern Classification, in Lincoln Laboratory Journal, Vol. 6, pp 249- 268, 1993.

[13] MediaEval Web Site, http://www.multimediaeval.org

[14] Mertens, R., Lei, H., Gottlieb, L., Friedland, G., and Divakaran, A.,"Acoustic Super Models for Large Scale Video Event Detection", to appear in Proc. of ACM Multimedia Workshop on Social Media, 2011.

[15] Rae, A., Murdock, V., Serdyukov, P., and Kelm, P.,"Working Notes for the Placing Task at MediaEval 2011", in Proc. of MediaEval, 2011.

[16] Reynolds, D.A., Quatieri, T.F., and Dunn, R., "Speaker Verification using Adapted Gaussian Mixture Models", in Digital Signal Processing, Vol. 10, pp 19–41, 2000.

[17] Schindler, G., Brown, M., and Szeliski, R.,"City-scale Location Recognition", in IEEE Conference on Computer Vision and Pattern Recognition, 2007.

[18] Stolcke, A., "SRILM - An Extensible Language Modeling Toolkit", in Proc. Intl. Conf. Spoken Language Processing, Denver, Colorado, 2002.

[19] Zhang, W., and Kosecka, J., "Image based Localization in Urban Environments", in 3rd International Symposium on 3D Data Processing, Visualization, and Transmission, 2006.