USER VERIFICATION: MATCHING THE UPLOADERS OF VIDEOS ACROSS ACCOUNTS

Howard Lei[‡◇], Jaeyoung Choi[‡◇], Adam Janin[‡], and Gerald Friedland[‡]

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

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

ABSTRACT

This article presents an attempt to link the uploaders of videos based on the audio track of the videos. Using a subset of the MediaEval [10] Placing Task's Flickr video set, which is labeled with the uploader's name, we conducted an experiment with a similar setup as a typical NIST speaker recognition evaluation run. Based on the assumption that the audio might be matched in various ways (speaker, channel, environmental noise, etc.), we trained one of ICSI's simplified speaker recognition systems on the audio tracks of the Flickr videos. Note that since the selection of videos is essentially random, the audio track can contain any sounds. We obtain an equal error rate of 36.7 % on 312 videos with 11,550 trials. The result has implications for audio research, security applications, and raises privacy concerns.

Index Terms— User verification, security, privacy

1. INTRODUCTION

With more and more multimedia data being uploaded to the web, it has become increasingly interesting for researchers to build massive corpora out of "wild" videos, images, and audio files. While the quality of randomly downloaded content from 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 to increase the robustness of approaches on a never before seen scale. Moreover, new tasks might be tackled that couldn't even be attempted before. In the following article, we present the task of linking personas based on modeling of the audio tracks of random Flickr videos. The experiment we describe in the paper aims to answer the question: "Do these two videos belong to the same Flickr user?". The experiment is modeled after speaker recognition experiments, where two audio recordings are analyzed for a possible match of the speaker. While using only the audio tracks for persona linking ignores other cues, such as the video tracks, it allows for applications of existing, well-established, speaker recognition approaches, while greatly reducing the computational requirements. Hence, we are able to determine the cross-domain applicability of the existing speaker recognition approaches, and can determine ways to tailor such approaches to cross-domain applications. Moreover, by using only the audio tracks, we can gain intuition as to the extent to which video personas depend on the audio tracks.

While speaker recognition evaluations usually follow strict guidelines concerning the quality and the channel of the recording, the experiment described herein uses random videos, which contain audio track with a large variance in quality. Nevertheless, the results of the experiment were far from random. 66.3 % of the users could be matched. Not only does this result provide evidence for the potential utility of "wild" videos, the outcome also has interesting implications for security applications and raises privacy concerns.

The article is structured as follows: Section 2 presents related work, before Section 3 describes the publicly available dataset. Section 4 then describes the speaker recognition system used for the experiment followed by Section 5 describing the results. Section 6 presents a final discussion and outlook to future work.

2. RELATED WORK

Work on using heterogeneous video collections from the Internet is an emerging topic of research; prominent examples include [12], which uses a speaker recognition system to identify famous celebrities in YouTube videos. An audio-visual system for recognizing celebrities in broadcast TV is presented in [4]. Linking online personas is an often discussed topic in security and privacy conferences. So far, however, we know of no work has been presented that uses multimedia such as audio and video features to link personas across web sites. In [5], the authors present an experiment to find YouTube users that are currently on vacation based on the geo-tagging of videos. The experiments presented in [1] investigate how much information can be extracted about a user from posted text across different social networking sites, linking users by querying potential email addresses on a large scale. While [8] presents experiments on matching personas using public information in a persona's social networking profile, it again exclusively concentrates on textual information.

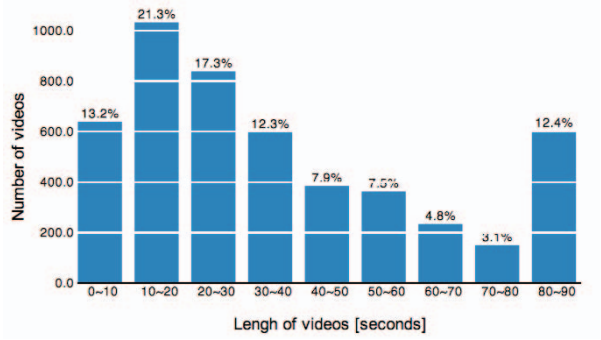

Fig. 1. A histogram visualizing the duration of the videos of the data set used in our experiments.

3. DATASET

3.1. Characteristics

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

Manual inspection of the data set lead us to conclude that most of visual/audio contents lack reasonable evidence to estimate the location [3]. For example, some videos were recorded indoors or in a private space such as a backyard of a house, which make 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.

The data set consists of 5125 Creative Commons licensed Flickr videos uploaded by Flickr users. 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). This policy generally ensures that each uploader's set of videos is "personal" in the sense that they were created by the same person and therefore likely have certain characteristic in common, such as editing style, recording device, or frequently recorded scenes/environments, etc.

From an examination of randomly sampled videos from the data set, we find that most of videos' audio tracks are quite "wild". We have observed 84 videos from 9 users. 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. 65.5 % of the videos had heavy ambient noises such as crowds chatting in the background, traffic noise, wind blowing into microphone, etc. 14.3 % of the videos contained music, either played in the background of the recorded scene, or inserted at the editing phase. About 50 % of the videos did not contain any form of human speech at all, and even for the ones that contain human speech, almost half were from multiple subjects and crowds in the background speaking to one another, often at the same time. 5 % of the videos were edited to contain changed scenes, fast-forwarding, muted audio, or inserted background music. Although we found that 7.2 % of videos contained audio of the person behind the camera, there is no guarantee that the owner of the voice is the actual uploader; it is possible that all videos from the same uploader were recorded by different people (such as family members).

We also sampled videos for similarity in the visual domain. If the videos were a series of scenes of a single event, it was fairly straightforward to identify them with a single uploader. For example, series of videos of President Obama's speech or the underwater footage of fish and coral reefs were easy to classify as from the same uploader. Of course, there are problematic examples of using this method — for example, one user uploaded a series of videos with different people at various locations around the world and with different recording devices, but conveying the same message. With an understanding of semantics, these are fairly easily identified by a human examiner, but nearly impossible for a machine.

The relatively short lengths of each audio track should be noted as can be seen in Figure 1. The length of Flickr videos are limited to 90 seconds. Moreover, around 70 % of videos in our data set have less than 50 seconds playtime, which is considerably shorter than for NIST evaluations.

3.2. Setup

All videos in the data set had been labeled by their uploader's username as specified in the video's metadata. For the experiment, we group videos by the uploaders who have uploaded more than 10 videos, and randomly select 20 % from each uploader's set (we sample randomly to reduce the running time of our system). Eventually, 312 videos from 83 users were selected for the experiment. We extract audios in PCM format from the selected videos and use these as the data set for the experiment.

4. TECHNICAL APPROACH

A generic 128-mixture GMM-UBM speaker recognition system [11] with relevance MAP adaptation and MFCC features C0-C12 (with 25 ms windows and 10 ms intervals) with deltas and double-deltas is used as our system to identify the uploader of the videos. We denote this system as our user ID system. The reason for using a GMM-UBM system is that it is commonly used as a speaker recognition baseline, and is easy to implement. For our system, Gaussian Mixture Models (GMMs) are used to model different users, and user-specific GMMs are adapted from a user-independent GMM (UBM) via relevance MAP adaptation [11]. Testing is done by com-

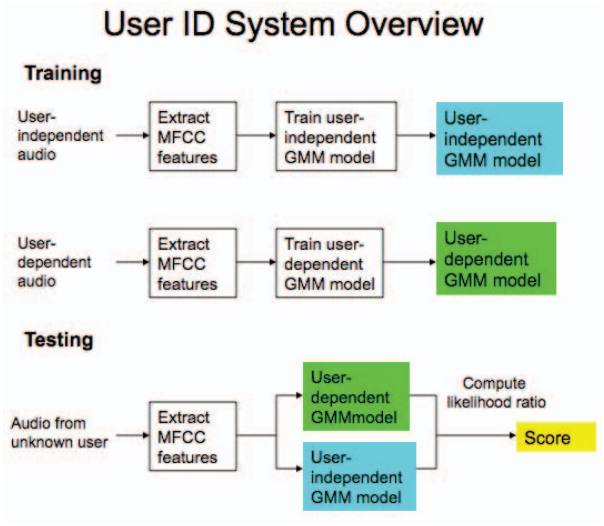

Fig. 2. An architectural overview of the user ID system as described in Section 4.

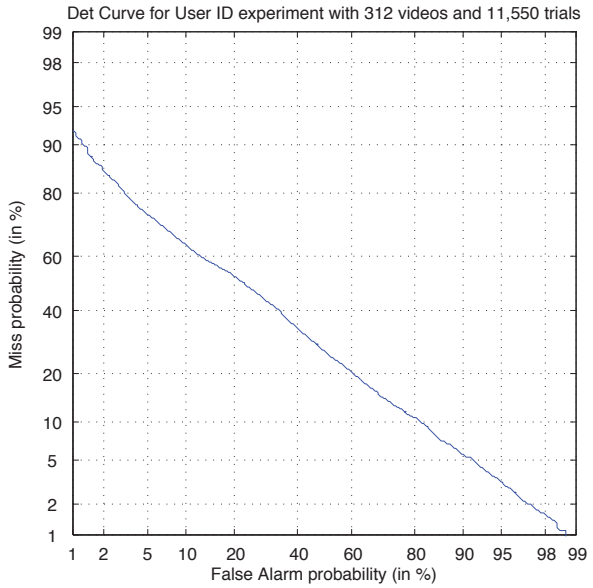

Fig. 3. The DET curve of the results described in Section 5.

puting the likelihood ratio of the features from an arbitrary piece of audio with one of the user models. Hence, each likelihood ratio computation produces a score representing the likelihood that the user in the arbitrary piece of audio is represented by the GMM user model, and is denoted as a trial. Trials for which the user corresponding to the audio matches the user for the GMM user model is denoted as a true speaker trial; other trials are denoted as impostor trials. The Equal Error Rate (EER) occurs at a scoring threshold where the rate at which impostor trials are misclassified as true speaker trials (false alarms) equals the rate at which true speaker trials are misclassfied as impostor trials (misses). The ALIZE speaker recognition system implementation is used [2], and the MFCC features are obtained via HTK [6].

While more advanced MAP speaker model adaptation techniques, such as eigenvoice and eigenchannel factor analysis, are available [9], the lack of sufficient high-quality training data prevent us from applying those techniques. Nevertheless, the approach we use has a benefit, in that it involves a robust well-establish speaker recognition algorithm that can be easily implemented and applied by anyone with an interest in speaker recognition. Moreover, our goal is currently not to establish the best speaker recognition results, but to demonstrate the possibility of applying speaker recognition techniques on large, universally accessible datasets.

5. EXPERIMENTS AND RESULTS

Experiments are run and user ID results are obtained using the aforementioned data and approach. The Shout [7] speech/non-speech detector is used to extract speech segments from each piece of audio, and the MFCC features corresponding to the speech regions are mean and variance normalized. Note that the random set of 312 videos used are ones where the speech/non-speech detector gave valid speech segmentations. The fact that the same set of videos used to train the UBM are used for training and testing is not a problem, because in the real world, we assume that people have access to all videos prior to user ID training and testing, and the user is free to train a user-independent model using all videos to improve user ID accuracy.

The GMM-UBM user ID system gives a 36.7 % EER on the 312 videos, according to the Detection Error Tradeoff (DET) curve shown in Figure 3. We can also obtain a measure of raw accuracy for our user ID system by setting the scoring threshold to the level for the EER, and simply tallying the number of videos whose user ID is correctly identified (i.e. videos whose impostor score falls below the scoring threshold, and whose true speaker score is above the threshold). 63.3 % (1391 out of 2196) of the true speaker trials are correctly classified; 63.3 % (5924 out of 9354) of the impostor trials are correctly classified.

In real world applications, it is arguably preferable to favor minimizing the number of false alarms at the expense of misses, since we want our hits to have a high likelihood of being correct. In other words, we would like to increase the accuracy of the impostor trials at the expense of the true speaker trials. If we increase the scoring threshold such that 90 % of the impostor trials are correctly classified (i.e. 10 % false alarm), we get that 36.5 % of true speaker trials are correctly classified. If we further increase the scoring threshold such that 99 % of the impostor trials are correctly classified (i.e. 1 % false alarm), we get that 7.9 % of the true speaker trials are correctly classified. Due to the enormous number of

online videos, 7.9 % still represents a significant number of videos that our system can correctly obtain the user ID of with relatively small risk (1 %) of being incorrect. While these results represent a first attempt at applying a generic speaker-recognition algorithm to the task of user ID of online videos, these results can be further improved as we seek ways to employ fancier speaker recognition approaches at our disposal.

6. CONCLUSION AND FUTURE WORK

This article describes the use of a simple speaker recognition system to match the uploaders of heterogeneous Flickr videos. The article shows that even with a very simple setup, videos can be matched with an accuracy of 63.3 % (with false alarms and misses equally balanced). The result is interesting for several reasons. When the false alarm rate is reduced to 1 % in favor of misses, the accuracy is at 7.9 %. In other words, a matched video is about 8 times as likely to be a match then a false alarm. This first shows that even highly tuned systems, like current speaker recognition systems, are generic enough to be "abused" for a different tasks. Second, it shows that random Internet data is not nearly as random as one might think, and therefore handleable by machine learning algorithms – supporting the current trend in the research community to work on this data. Third, and most importantly, the experiment has implications for security and privacy. A speaker recognition system can be used to link independent personas. In other words, it is not safe to use different user names to keep sets of videos distinct. Law enforcement might use the result to try to match videos of a criminal against a public video database in an attempt to identify the perpetrator.

In future work, the tuning of the speaker recognition system to this specific task would likely improve the accuracy. More importantly, we expect that the combination with other cues, such as text or video features, will improve results dramatically.

7. ACKNOWLEDGMENTS

This work was partially funded by an NGA NURI grant.

8. REFERENCES

[1] Balduzzi, M., Platzer, C., Holz, T., Kirda, E., Balzarotti, D., Kruegel, C., "Abusing Social Networks for Automated User Profiling", Lecture Notes in Computer Science, Vol. 6307/2010, pp. 422–441, 2010

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

[3] Choi, J., Janin, A., Friedland, G., "The 2010 ICSI Video Location Estimation System," to appear in Proceedings of MediaEval 2010, Pisa, Italy, October 2010.

[4] Everingham, M., Sivic, J., Zisserman, A., "Hello! my name is... Buffy – automatic naming of characters in TV video", Proceedings of the British Machine Vision Conference, vol. 2, 2006

[5] Friedland, G., Sommer, R., "Cybercasing the Joint: On the Privacy Implications of Geo-Tagging", Proceedings of the Fifth USENIX Workshop on Hot Topics in Security (HotSec 10), Washington, DC, 2010

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

[7] Huijbregts, M, "Segmentation, Diarization and Speech Transcription: Surprise Data Unraveled", PhD Thesis, University of Twente, 2008

[8] Irani, D., Webb, S., Li, K., Pu, C, "Large Online Social Footprints – An Emerging Threat", International Conference on Computational Science and Engineering, vol. 3, pp. 271–276, 2009 , 2009

[9] Kenny, P., Boulianne, G., Ouellet, P. and P. Dumouchel., "Joint factor analysis versus eigenchannels in speaker recognition", in IEEE Transactions on Audio, Speech and Language Processing 15 (4), pp. 1435-1447, May 2007

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

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

[12] Sargin, M.E., Aradhye, H., Moreno, P.J., Ming Zhao, "Audiovisual celebrity recognition in unconstrained web videos", in ICASSP, pp. 1977 – 1980 (2009)