NOWHERE TO HIDE: EXPLORING USER-VERIFICATION ACROSS FLICKR ACCOUNTS

Howard Lei[‡], Jaeyoung Choi[‡◇], 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 work presents improved audio-based user-verification analysis and results on Flickr videos, using a subset of the MediaEval 2011 [1] data set. User-verification is a new task, where the goal is to determine if two pieces of media are uploaded by the same user. Our best results, with a 19.7% Equal Error Rate, and a 53.9% Miss Rate at 1% False Positive, are obtained using an i-vector [2] system. A frequency-matching system that requires 96% less computation time than the other systems is also explored, and may be better suited for processing large datasets from Flickr and other social networks. The results have significant privacy implications as they present a framework for exploiting users' tendencies to assume that different accounts remain as separate realms.

***Index Terms*—** User-verification, i-vectors, social media, security, privacy

1. INTRODUCTION

With more and more multimedia data uploaded to the web, it has become increasingly interesting for researchers to build massive corpora out of 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 systems on a never-before-seen-scale. Moreover, new tasks might be tackled that couldn't even be attempted before.

We present the task of user-verification based on the audio tracks of random Flickr videos, which attempts to answer the question of whether the audio tracks of two Flickr videos came from the same user who uploaded the videos. This task is related to the task of speaker-verification, where the goal is to determine if two audio recordings contain the same speaker. However, the task of user-verification presents additional audio-processing challenges in terms of dealing with uncontrolled audio conditions, and where the majority of audio are essentially "wild", with high variance in audio bandwidth, quality, environmental noise, and context. Furthermore, there is no guarantee that the user who uploaded a Flickr video is the same as the one who produced the video, or the one who may have spoken in the video. User-verification also raises privacy and security concerns for social network usage. Oftentimes, social network users create two separate social network accounts to assume two distinct identities, with the assumption that each account is a separate realm in and of itself. The user-verification task, however, can invalidate such an assumption, as it attempts to link users across accounts.

While there are certainly other modalities that can be exploited, such as video, image, and textual metadata, audio is an easy-to-obtain modality that can be efficiently processed with many existing state-of-the-art algorithms, and do not require much in terms of storage. We recognize that there could be many benefits in terms of performance and understanding gained via the use of the other modalities. However, in this paper, we limit our user-verification experiments to audio, to illustrate the many implications that can result from even a simpler treatment of the task. Various points of analysis, improvements to the task, and understanding gained since our preliminary attempt at user-verification in [3], will be discussed. While our user-verification results in [3] demonstrate potential privacy and security implications, our approaches and results (of higher accuracy) presented in this work presents implications that are far more serious.

The systems used in our experiments are the i-vector system [2], the GMM-UBM system [4], and a new system based on frequency spectrum comparisons between audio file pairs. Both the i-vector and GMM-UBM systems are off-the-shelf approaches used in speaker-verification, with the i-vector system being a state-of-the-art approach. In this work, we demonstrate the potency that even such standardized approaches can have in user-verification. The frequency-matching system is a simplistic approach that can nevertheless be effective in user-verification. Although the user-verification task using "wild" audio presents many challenges, our i-vector system was able to achieve less than a 20% Equal Error Rate (EER) under certain audio conditions, with only a 54% Miss Rate at 1% False Positive (FP), and a 80% miss rate at 0.1% FP. This work presents our algorithmic approaches, results, analysis, along with the practical concerns amongst social network users associated with the results.

The rest of this paper is structured as follows: Section 2 presents related work; section 3 describes the publicly available dataset; section 4 describes the systems used for our experiments; section 5 describes the experiments and results for various systems and audio conditions; section 6 discusses the practical implications of our results, and section 7 presents the conclusion and future work.

2. PRIOR AND RELATED WORK

Work on using heterogeneous video collections from the Internet is an emerging topic of research; prominent examples include [5], which uses a speaker ID system to identify famous celebrities in YouTube videos. An audio-visual system for recognizing celebrities in broadcast TV is presented in [6]. In [7], 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 [8] 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 [9] and [10] present experiments on matching personas using public in-

978-1-4799-0356-6/13/$31.00 ©2013 IEEE

formation in a persona's social networking profile, they exclusively concentrate on textual information.

There is also related work on audio-based social media classification. The works [11] and [12] discuss audio-based city-verification and acoustic event detection using Flickr videos. We first attempted the task of linking Flickr users based on the audio-tracks of user-uploaded videos in [3], where we presented some preliminary, yet encouraging results, using only the GMM-UBM system. Aside from social-media classification, there is also significant amounts of work in the field of speaker verification [4] [2] [13], where the goal is to determine if two audio recordings contain the same speaker. The i-vector and GMM-UBM systems used in this work have been initially developed for speaker verification.

3. DATASET

3.1. Characteristics

The audio tracks for the experiment are extracted from the videos that have been distributed as training and test data sets for the Placing Task of MediaEval 2011 [1], a multimedia benchmark evaluation. The Placing Task involves automatically estimating the location of each test video using one or more of: metadata (e.g. textual description, tags), visual/audio contents, and social information. The videos are not pre-filtered or pre-selected in any way to make the data set more relevant to the user-verification task, and are therefore likely representative of videos selected at random.

A total of 10,857 Creative Commons licensed Flickr videos, uploaded by 2,943 Flickr users, were used in our experiments. 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 123 short-duration videos that were randomly selected from the data set, we found that most of videos' audio tracks are quite "wild". 59.3% of the videos are home-video style with ambient noises. 47.2% of the videos had heavy ambient noises such as crowds chatting in the background, traffic noise, and wind blowing into microphone. 25.2% of the videos contained music, either played in the background of the recorded scene, or inserted at the editing phase. 59.3% of the videos did not contain any form of human speech at all, and even for the ones that contained human speech, 64% were from multiple subjects and crowds in the background speaking to one another, often at the same time. Only 1 video was edited to give a slow motion effect. Although we found that 10.5% 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). All videos are limited to 90 seconds. 71.8% of videos have less than 50 seconds of playtime, while 50% have less than 30 seconds of playtime.

4. TECHNICAL APPROACHES

This sections describes the i-vector, GMM-UBM, and frequency-matching systems. The i-vector system is described in [2]. It involves training a total variability matrix T to model the variability (both user- and channel-related) of the acoustic features of all audio tracks, and using the matrix to obtain a low-dimensional vector characterizing the "user print" of each audio track. Specifically, for each audio file, a vector of first-order statistics M - of the acoustic feature vectors of the audio centered around the means of a GMM world model - is first obtained, and can be decomposed as follows:

$$M = m + T\omega \tag{1}$$

where m is GMM world model mean vector, and ω are low-dimensional vectors, known as the identity vectors or i-vectors.

The i-vector system then involves performing Probabilistic Linear Discriminant Analysis (pLDA) [14] and Within-Class Covariance Normalization (WCCN) [15] on the i-vectors. pLDA linearly projects the i-vectors ω onto a set of dimensions to maximize the ratio of between-user scatter to within-user scatter of the i-vectors, producing a new set of vectors. WCCN then whitens the pLDA-projected vectors via a second linear projection, such that the resulting vectors have an identity covariance matrix. For our user-verification system, 1,024 mixtures are used for the GMM world model, and a rank of 400 is used for the total variability matrix T, such that the i-vectors ω have 400 dimensions. pLDA projects the i-vectors onto a set of 200 dimensions. The cosine distance is used to obtain the user-similarity score of a pair of i-vectors ω between two audio tracks of user-uploaded videos [2]:

$$score(\omega_1, \omega_2) = \frac{(A^T\omega_1)^T W^{-1}(A^T\omega_2)}{\sqrt{(A^T\omega_1)^T W^{-1}(A^T\omega_1)}\sqrt{(A^T\omega_2)^T W^{-1}(A^T\omega_2)}} \tag{2}$$

where A and W are the LDA and WCCN projection matrices respectively, and ω_1 and ω_2 are i-vectors from the two audio tracks being compared against. The acoustic features consist of MFCC C0-C19+Δ+$\Delta\Delta$ coefficients of 60 dimensions, computed using 25 ms windows and 10 ms shifts, across 60 to 16,000 Hz. Note that the Brno University of Technology's (BUT's) Joint Factor Analysis Matlab demo [16] is used to assist in the i-vector system development. The open-source ALIZE toolkit [17] is used to train the GMM world model.

The GMM-UBM system [4] uses the same set of MFCC features as used in the i-vector system, and involves training user-specific GMM models via Maximum a-Posteriori adaptations of a GMM world model (i.e. the UBM) for each audio track based on MFCC features of the audio. For the GMM-UBM system, 128 mixtures are used for all GMM models. The user-similarity score between two audio tracks is obtained by the computation of the log-likelihood of the MFCC features from one of the files with the GMM model trained from the features of the other file. The GMM-UBM system was used in our prior attempt at linking Flickr users based on the audio-tracks of user-uploaded videos [3].

The system based on frequency-matching involves first obtaining the frequency envelope of each entire audio track through 1,024 critical band integrations. All bands are rectangular and evenly-spaced in frequency - from 0 Hz to half the sampling frequency (16,000 Hz) - and have the same height. The resulting value from each critical band integration is stored in a frequency bin, and the values from all frequency bins comprise the smoothed frequency envelope. The results do not appear to change by varying the number of critical bands from 512 to 2,048, so 1,024 is used. The Manhattan distance between the frequency bins of pairs audio tracks is used to get the user-similarity score. Note that in our experiments, the Manhattan distance performed better than a host of other distance metrics. While the system based on frequency-matching is simplistic, its results in Section 5 demonstrate improved speeds by which

System	EER	Miss Rate at 1% FP	Miss Rate at 0.1% FP
i-vector	27.3%	68.3%	88.3%
GMM-UBM	31.6%	74.4%	91.8%
Frequency-Matching	36.2%	83.9%	94.4%

Table 1. Results for the i-vector, GMM-UBM and frequency-matching systems for user-verification using a set of 6,108 audio tracks of user-uploaded Flickr videos. Similarity scores were computed on 6 million audio track pairs, with a total of 1,239 training users and 2,784 test users.

user-verification can be performed, while maintaining a viable accuracy.

System	EER	Miss Rate at 1% FP	Miss Rate at 0.1% FP	User Set
i-vector	19.7%	53.9%	80.3%	Open
GMM-UBM	19.7%	61.8%	86.8%	Open
Frequency-Matching	26.3%	71.1%	85.5%	Open
i-vector	19.7%	56.6%	80.3%	Closed
GMM-UBM	21.0%	67.1%	88.2%	Closed
Frequency-Matching	27.3%	71.1%	81.6%	Closed

Table 2. Results for the i-vector, GMM-UBM, and frequency-matching systems for user-verification using audio tracks of 10 or less seconds in duration. Open- and closed-set experiments were performed for a set of 47 common users.

5. EXPERIMENTS AND RESULTS

All experiments used the MediaEval 2011 corpora, consisting of user-uploaded Flickr videos, from which the audio tracks are used. A set of 1,239 Flickr users in the corpora were designated as training users, and 2,784 were designated as test users, with 1,226 users in common with the training users. There are a total of 1,239 audio tracks associated with the 1,239 training users, and 4,869 audio tracks associated with the 2,784 test users. Overall, a set of 6,108 audio tracks were used for training and testing. A separate set of 146 users with 4,605 audio tracks were used to train the T-matrix, and LDA and WCCN matrices of the i-vector system. 2,200 audio tracks from the 146 users were used to train the GMM world model for the i-vector and GMM-UBM systems. A total of 6 million similarity scores were computed between the audio tracks from the training and test users, with 3,287 of the scores coming from audio track pairs with the same Flickr user. Table 1 shows the EER and Miss Rates at 1% and 0.1% FP on the 6 million scores for the i-vector, GMM-UBM, and frequency-matching systems.

Results in Table 1 indicate that the i-vector system, which has a 27.3% EER, a 68.3% Miss Rate at 1% FP, and a 88.3% MIss Rate at 0.1% FP, outperforms both the GMM-UBM and frequency-matching systems. This system also outperforms the GMM-UBM system, which our previous results in [3] were based on. However, even though the frequency-matching system has the worst results amongst the three systems (36.2% EER, a 83.9% Miss Rate at 1% FP, and a 94.4% Miss Rate at 0.1% FP), the results are nevertheless respectable given the simplicity of the system and the difficulties presented in the dataset. The frequency-matching system requires no development nor pre-training, as each user-similarity score can be computed given only the waveforms of the pair of audio tracks.

One reason the i-vector system outperforms the GMM-UBM system is that its WCCN and pLDA components account for both the within-user and between-user i-vector scatter of the data. The WCCN component, which applies a linear transformation to the i-vectors such that the within-user i-vector covariances of each user would be closer to unity, aims to compensate for distortions in the covariances due to within-user variability. Such variability can include same-user videos captured in different acoustic environments. The pLDA applies a linear projection on the i-vectors such that the between-user scatter to within-user scatter of the i-vectors would be maximized. The GMM-UBM system, in contrast, only uses generative Gaussian models to model the acoustic features of the users, and does not account for within-user and between-user variability.

The end-to-end system runtime for processing the 6 million scores using the frequency-matching system is 2.5 hours on a single CPU, which represents 1.5 milliseconds per score. In contrast, the end-to-end runtime is 74.2 hours for the i-vector system, and 72.3 hours for the GMM-UBM system on a single CPU (note that these systems require MFCC feature extraction, and training of the GMM world models and various matrices). Given the massive amounts of media found on the web from the large numbers of social network users, the use of simple systems that require less processing time ought to be preferred.

Experiments were also performed on different splits of the 6,108 audio tracks to determine how likely users can be matched given the characteristics of the data they upload. One of the characteristics of data that can potentially lead to better user-matching is the length of the audio tracks. Audio tracks of shorter duration would contain more homogenous sounds of less acoustic variability than tracks with longer durations, which was confirmed via listening experiments. Audio with less acoustic variability would allow systems to process and characterize fewer acoustic characteristics, perhaps resulting in acoustic models that can more faithfully capture the entirety of the audio. Furthermore, videos of shorter duration would more likely contain instances of exciting events. Different users may have different preferences for which exciting events to capture on video, resulting in potentially greater user discriminability and system performance of shorter-duration audio tracks.

Hence, the three systems were also run on a subset of the 6,108 audio tracks of 10 or less seconds in duration to determine if shorter files uploaded by Flickr users might be more useful for user-matching. Note that the 10 second cutoff was arbitrary chosen. 532 of the 6,108 audio tracks are 10 or less seconds in duration, with 121 training users and 329 test users. There are 47 common users between the training and test audio tracks. Results are shown in table 2 for both the open-set (where not all test users correspond to a training user) and closed-set (where each test user corresponds to a training user) experiments. The open-set experiments use all 121 training users and 329 test users, with 49,731 user-similarity scores, among which 76 scores have matching users; the closed-set experiment use only 123 total audio tracks, with 3,496 similarity scores and the same 76 scores with matching users.

Results indicate that using audio tracks of 10 or less seconds in duration leads to better performances for all three user-matching systems, for both the open-set and closed-set experiments. For the open-set experiments, the i-vector system again outperforms the GMM-UBM and frequency-matching systems with a 19.7% EER, a 53.9% Miss Rate at 1% FP, and a 80.3% Miss Rate at 0.1% FP. Even though the frequency-matching system lags in performance compared to the other systems, it nevertheless demonstrates a better EER (26.3%) and a better Miss Rate at 0.1% FP (85.5%) than all three systems using the larger set of 6,108 audio tracks (in Table 1). It's Miss Rate at 1% FP (71.1%) is also better than two of the three systems for the experiments using the larger dataset. Overall, the results using only the short-duration audio tracks significantly outperform the results using the larger set of audio tracks for all three systems. The following section discusses the results of the closed-set experiments, and its implications on user privacy and security.

6. DISCUSSION

The results shown in Table 2 for the task of user-verification has implications for social network user privacy and security. For users with multiple Flickr accounts, or with multiple accounts across different social networks, the results suggest that it is possible to link the accounts of a single user based on his or her uploaded videos. Users who upload videos of 10 or less seconds in duration face increased risks of having their accounts linked. Amongst the 47 common users, the 56.6% Miss Rate at 1% FP for the i-vector system suggests that almost half of the audio track pairs with matching users as indicated by the system actually do have matching users, with only a 1% chance of being a false positive. The 81.6% Miss Rate at 0.1% FP for the simple frequency-matching system also has significant implications. It suggests that amongst the 47 common users, roughly 1-in-5 of the audio track pairs with matching users as indicated by the system actually have matching users, with only a 0.1% chance of being a false positive.

Given that there are 49,731 total audio pairs with only 76 matched-user pairs for the short-audio experiments, however, we acknowledge that even at the 1% and 0.1% FP rates, the number of false positives significantly outnumber the true positives. A 81.6% Miss Rate at 0.1% FP for the frequency-matching system suggests that there are roughly 14 true positives with 50 false positives – about 4 times the number of true positives. Hence, in order to correctly identify matching-user pairs in the data, either additional classifiers are needed to complement this audio-based approach, or the amount of overall data should be reduced, so as to avoid having to search through the many false positive to find the true positives. Nevertheless, the results provided by the system represents one step towards audio-based user matching in social media, and the system can allow matched-user pairs to be identified given scenarios where the numbers audio pairs are limited. Even for the case of 49,731 audio pairs, with 14 true positives and 50 false positives, it is possible to manually search through the 64 total positives for the 14 true positives that have matching users. The fact that these numbers are obtained using the simple frequency-matching system is significant as well, suggesting that a complicated approach is not necessarily needed to perform user-matching. Furthermore, the use of additional modalities such as video and text can enhance user-matching performance.

It is also interesting to examine why the short-audio experiments using a subset of the data produced superior results in comparison to the experiments using the larger dataset. A listening of the randomly-selected 123 short audio tracks from the 47 common users indicates that the audio tracks for 35 of the 47 users contain the same acoustic environment, which commonly consists of music concerts, crowd noise, engine noise, wind noise, and traffic noise. Hence, a major factor contributing to user-verification performance of the systems is the acoustic environments in which the users capture video, where the same user tends to upload videos captured in similar acoustic environments. This implies that a user can perhaps evade being matched by ensuring greater variability in the acoustic environments of his or her uploaded videos, such as with longer video recordings in a greater diversity of environments.

7. CONCLUSION AND FUTURE WORK

This work demonstrates the feasibility of linking users across Flickr accounts based on the audio tracks of user-uploaded videos. We used the i-vector and GMM-UBM systems, along with a system based on frequency-matching, and showed that the i-vector system achieves the best user-matching performance on short-duration (10 seconds or less) videos, with a 19.7% EER, 53.9% Miss Rate at 1% FP, and 80.3% Miss Rate at 0.1% FP. While the frequency-matching system achieves the worst results, it is a simple approach that computes the user-similarity scores using only the frequency spectra of audio files, and requires no development nor pre-training. The frequency-matching system is able to generate 6 million user-similarity scores in 2.5 hours, and is 96.5% and 96.6% faster than the GMM-UBM and i-vector systems respectively. Given the massive volumes of social network media that can be used for user-linking, the frequency-matching system seems desirable.

This work suggests that it is the acoustic characteristics of videos that enable audio-based user matching. Videos of shorter duration are more useful for user-matching, due to the lack of acoustic variability within short videos, and the potentially greater user-discriminative power of the short acoustic instances captured in the videos. The respectable user-matching performances is likely due to the fact that different users tend to capture videos from different acoustic environments, and videos uploaded by the same user tend to have similar acoustic environments. Future work could involve exploring additional modalities such as video, and textual metadata, and performing more detailed analysis to determine other factors in the audio tracks that impact user-matching performance. As future work, we also propose a larger discussion among the signal processing community with regards to the privacy and security implications associated with the user-verification task.

8. ACKNOWLEDGEMENTS

This research is supported by 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.

9. REFERENCES

[1] "Mediaeval web site," http://www.multimediaeval.org.

[2] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *Proceedings of Interspeech*, Brighton, UK, 2009.

[3] H. Lei, J. Choi, A. Janin, and G. Friedland, "User verification: Matching uploaders of videos accross accounts," in *Proceedings of ICASSP*, May 2011, pp. 2404–2407.

[4] D. A. Reynolds, T. F. Quatieri, and R. Dunn, "Speaker Verification using Adapted Gaussian Mixture Models," in *Digital Signal Processing*, 2000.

[5] M. E. Sargin, H. Aradhye, P. J. Moreno, and M. Zhao, "Audio-visual celebrity recognition in unconstrained web videos," in *Proceedings of ICASSP*, Taipei, Taiwan, 2009, pp. 1977–1980.

[6] M. Everingham, J. Sivic, and A. Zisserman, "Hello! my name is... buffy – automatic naming of characters in tv video," in *Proceedings of BMVC*, 2006, vol. 2.

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

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

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

[10] O. Goga, H. Lei, S. Parthasarathi, G. Friedland, R. Sommer, and R. Teixeira, "Exploiting Innocuous Activity for Correlating Users Across Sites," in *accepted to the 23rd International World-Wide Web Conference (WWW 2013)*, May 2013.

[11] H. Lei, J. Choi, and G. Friedland, "Multimodal City-Verification on Flickr Videos using Acoustic and Textual Features," in *Proceedings of ICASSP*, 2012.

[12] R. Mertens, H. Lei, L. Gottlieb, G. Friedland, and A. Divakarian, "Acoustic Super Models for Large Scale Video Event Detection," in *Proceedings of the Joint ACM Workshop on Modeling and Representing Events*, 2011.

[13] L. Burget, P. Oldřich, C. Sandro, Oldřej G., Pavel M., and N. Brümmer, "Discriminantly trained probabilistic linear discriminant analysis for speaker verification," in *Proceedings of ICASSP*, Brno, Czech Republic, 2011.

[14] S. Ioffe, "Probabilistic linear discriminant analysis," in *Proceedings of ECCV*, 2006, pp. 531–542.

[15] A. O. Hatch, "Generalized linear kernels for one-versus-all classification: Application to speaker recognition," in *Proceedings of ICASSP*, Toulouse, France, 2006.

[16] "Joint factor analysis matlab demo," http://speech.fit.vutbr.cz/software/joint-factor-analysis-matlab-demo/.

[17] J.F. Bonastre, F. Wils, and S. Meignier, "Alize, a free toolkit for speaker recognition," in *Proceedings of ICASSP*, 2005, vol. 1, pp. 737–740.