Multi-Modal User Authentication from Video for Mobile or Variable-Environment Applications

Timothy J. Hazen and Daniel Schultz

MIT Computer Science and Artificial Intelligence Laboratory
Cambridge, Massachusetts, USA

Abstract

In this study, we apply a combination of face and speaker identification techniques to the task of multi-modal (i.e., multi-biometric) user authentication for mobile or variable-environment applications. Audio-visual data was collected using a web camera connected to a laptop computer in three different environments: a quiet indoor office, a busy indoor cafe, and near a noisy outdoor street intersection. Experiments demonstrated the benefits that may be obtained from using a multi-modal approach, even when both input modalities suffer from difficult environmental conditions or a poor match between training and testing conditions. Over twelve different training and testing conditions, user authentication equal error rates were reduced an average of 19% from the best individual biometric in each condition, and 36% from an audio-only system.
Index Terms: speaker identification, face identification, multi-biometric user authentication

1. Introduction

This paper investigates the integration of two biometric techniques, face and speaker identification, for mobile and/or variable-environment applications. As small, mobile computing devices, such as personal data assistants and handheld tablets, continue to become more pervasive, the need for security on them also increases. To prevent impostors from gaining access to sensitive information, stored on a device or on the device's network, security measures must be incorporated into these devices. Face and speaker verification are two techniques that can be used in place of, or in conjunction with, pre-existing security measures such as personal identification numbers or passwords. The computational power and video capture technology that could make audio-visual user authentication possible on small handheld devices will likely become generally available on these devices in the near future.

Mobile devices offer two distinct challenges for standard face and voice identification approaches. First, their mobility ensures that the environmental conditions the devices will experience will be highly variable. Specifically, the audio captured by these devices could contain highly variable background noises resulting in low signal-to-noise ratios. Similarly, the images captured by the devices can contain highly variable lighting and background conditions. Second, the quality of the video and audio capture devices is also a factor. Typical consumer products are constrained to use audio/visual components that are both small and inexpensive, resulting in a lower quality audio and video than is typically used in laboratory experiments.

To examine these issues we have previously developed a prototype system for incorporating two biometric techniques, speaker identification and face identification, onto a mobile device [1, 2]. In this previous study, we evaluated a combined face and speaker identification system within a user verification "login" scenario on an iPAQ handheld computer. The multi-biometric system was able to achieve reductions of up to 90% in the verification equal error rate (EER) over a system using only our speaker identification technology. In this previous study, all data was collected indoors in relatively quiet offices and hallways. Additionally, the face identification operated on only single still image of the user's face. In this paper, we extend our previous work by altering two significant experimental conditions: (1) the data is collected using full motion video collected by a web camera, and (2) the data is collected in three significantly different and potentially difficult environments.

The rest of the paper is organized as follows. We first present an overview of our two biometric techniques and the fusion technique for combining them. Next, we discuss the verification paradigm we are assuming and the methods of data collection we employed. We follow this with experimental results showing the performance of the two biometric techniques on the data we have collected, both individually and in combination. Finally, we summarize and discuss the results and present plans for future directions of our work.

2. Data Collection

For our experiments, subjects were recorded reading prompted digit strings, each ten digits long in length. Data was collected using a Logitech QuickCam Pro web-camera attached to a laptop. The audio was collected from the camera's far-field microphone at a sampling rate of 16 kHz. The video was recorded in 24-bit color at a resolution of 160×120 pixels. To improve computation time during video processing, the videos were down-sampled from 24-bit color resolution to 8-bit grayscale resolution. While the frame rates of the videos varied, they were typically between 25 and 30 frames per second. The length of the videos also varied but were usually between 4 and 8 seconds.

The portability of the laptop allowed for the collection of data in multiple environments. During each session, a subject was recorded in three separate locations. The first location was a quiet, well-lit office setting with generally very little acoustic noise and consistent lighting conditions between recording sessions. The second location was a busy cafe area in the lobby of an academic building. This setting contained a variety of noises associated with busy cafe/lobby areas (e.g., babble noise, footsteps, cafe machinery, etc.). The lighting conditions also contained a mix of natural and artificial lighting which varied across different sections of the cafe/lobby area. The third location was an outdoor setting near a busy street intersection containing heavy motor traffic. In some sessions, heavy wind noise is also present in the recordings. Lighting in the outdoor setting also varied the most of the three locations, with significant

differences in lighting based on the presence or absence of direct or strong indirect sunlight, as well as the positioning of the speaker relative to the sun (i.e., back-lit, side-lit, etc.)

For each recording session, the subject read the same nine strings in each of the three locations. Each utterance was recorded within a separate video file. The recording sessions always began in the office setting before next moving to the cafe and then the street intersection. The subject would start recording, read one utterance, and then stop recording. Once complete, each session contained 27 total recordings, nine in each of the three locations.

In total, 100 different subjects were recorded. Fifty subjects provided two recording sessions on different days; these subjects became the *enrolled* users in our experiments. The remaining 50 subjects provided only one recording session and were used as *impostors*. Of the 100 subjects, 41 were male, 59 were female. There were 86 native speakers of American English. Of the 14 non-native speakers, most were native speakers of Chinese, though there were also subjects whose native language was UK English, Dutch, and Gujarati (an Indian dialect).

3. User Identification

This work builds upon our previous work in multi-modal person identification [1, 2, 3]. Below we summarize the technologies used in our experiments. Readers are referred to our previous work for additional details.

3.1. Speaker Identification

One technique used in speaker verification is to prompt the user with a randomly generated challenge phrase. This lets the system avoid pre-recorded spoofing, and allows it to build phone-specific models for the items in its phrase repertoire. In our work, an automatic speech recognition system first identifies the phonetic content of the utterance, before analyzing the utterance using a text-dependent speaker verification system [4].

When processing the speech signal, a speech recognition engine is first used to verify that the prompted pass-phrase was spoken. In noisy conditions, it is possible that the speech recognition engine is unable to reliably detect the pass phrase in the audio signal. In this case, the system could elect to either reject the user outright, or in the case of a multi-modal system fall back onto fully trusting the visual modality.

After speech recognition is complete, the speaker verification system produces scores based on phone-dependent models for the specific phonetic content in the phrase. These scores represent zero-centered log likelihood ratios between the scores for the purported speaker and the scores from a background speech model trained from data from many speakers. In addition to the score for the purported speaker, the system also generates scores for a set of different *cohort* speaker models.

3.2. Face Identification

The face identification framework used in our work is similar to the one described in [5], but with some minor differences in the face detection methods, as well as the extension from still images to a sequence of images collected in a video. The system applies a face detection algorithm to locate and track the face throughout the video of a spoken utterance. From each visual frame, a feature vector describing the facial features of the user is extracted and the face identification system scores each individual image. The scores from all image frames with detected faces are then aggregated over the duration of the video.

Figure 1: Example images extracted for face identification. The top-row contains three different people in the office environment. The bottom row shows the same person in each of the office, cafe, and outdoor environments.

3.2.1. Face Detection

Each video was analyzed by face detection software built upon the Viola-Jones face detection algorithm [6]. Specifically, our system applies the face detection component contained in Intel's Open Computer Vision software package [7] to the individual frames of each video, and then smooths the results over multiple frames using a Kalman filter post-processing stage [8]. For poor quality videos (e.g. those with severe shadows, poor lighting, or blurriness) the face detection algorithm sometimes fails to detect a face for some, or sometimes all, of the frames within a video.

3.2.2. Feature Extraction

After face detection, a feature vector describing the face (or a portion of the face) can be extracted. For the experiments presented in this paper, the facial images are vertically cropped to a region spanning from just below the lower lip to just above the eyebrows. Horizontally, the images span roughly between the outer edges of the eyes. The detected region is resized into a 40x40 image. Figure 1 shows typical example images extracted from the full video frame images. Other potential image-based feature sets, including images of the eyes and nose only, images of the mouth only, and image differentials between successive mouth images, have also been explored in [3].

3.2.3. Face Recognition

For recognition, a one-vs-all support vector machine (SVM) classification scheme is used, where one classifier is trained to distinguish each person in the database from all others [9]. In the SVM training process for each person's classifier, the feature vectors corresponding to that person's training images are used as positive examples, and the feature vectors corresponding to images from all other enrolled users are used as negative examples. A second-order polynomial SVM kernel function [9] is applied to the data in our experiments. We used the SvmFu support vector machine package [10] for all SVM operations in our experiments.

During the face identification process the system outputs a score from the individual SVM classifier corresponding to the purported user, as well as classifier scores from a set of cohort users. The scores are zero-centered, i.e., a score of zero means the data point lies directly on the decision hyperplane, and positive and negative scores correspond to the positive and negative sides of the decision hyperplane respectively. The absolute value of the SVM output represents the distance from the decision hyperplane.

3.3. Multi-Modal Fusion

A variety of methods are available for fusing the scores of multiple output classifiers including probabilistic combination operators [11], Bayesian statistics methods [12], and linear discriminant functions [13]. In general the fusion scheme for combining the outputs of different classifiers must compensate for the variance and mean biases of the different classifiers as well as relative reliability of each classifier. In our past work, we have generally found a simple linear combination of the audio and visual output scores to be sufficient. However, this does require the training of these weights on development data.

In this work we investigate the application of test normalization (or *T-norm*) [14] in the fusion process. T-norm has been widely used for speaker verification tasks as a mechanism for normalizing output scores that may vary in range across different training and testing conditions. To perform the T-norm process, the score S_u for a purported user u is converted to a normalized score S'_u using the following expression:

$$S'_u = \frac{S_u - \mu_C}{\sigma_C} \quad (1)$$

Here, μ_C and σ_C represent the mean and standard deviation of the scores produced by a set of cohort speakers C.

One way to use T-norm in a multi-modal systems is to apply the technique to the final scores after the multi-modal score fusion has occured. We will refer to this as *post-fusion normalization*. Alternatively, T-norm can be applied to each modality's output results before score fusion. We will refer to this as *pre-fusion normalization*. The use of pre-fusion normalization helps compensate for score variance and mean bias discrepancies prior to score fusion. This allows any additional fusion weighting scheme to concentrate solely on the issue of the reliability of the different classifiers.

4. Experimental Results

4.1. Experimental Conditions

All of our experiments are performed on the task of person verification, i.e., a recorded utterance is presented to the system along with a purported identity and the system must determine if the purported identity is true or false. True user trials are drawn from the set of 50 enrolled users who contributed two recording sessions. The false impostor trials are all drawn from the impostor users who contributed only one recording session each. We utilize one session of each enrolled user exclusively for training, and the second recording session exclusively for testing. During testing, the set of cohort speaker models used for the T-norm process is simply the set of 49 speaker models other than the specific model set of the purported speaker.

In real-world applications, a deployed system may not be able to control the conditions under which a user either enrolls into a system or later tries to gain reentrance into the system. To simulate the wide range of possible training and testing cross-condition cases, our experiments examined twelve different training and testing scenarios, resulting from four different training conditions crossed against three different testing conditions. Tests were conducted in each of the three different environments: the *office*, the *cafe*, and near the *street*. The first three training conditions used only the 9 utterances from their specific environment for enrollment. The final training condition used all 27 utterances per speaker aggregated from all three environments. This final *mixed* training condition represents the desirable condition where a user is able to provide enrollment utterances from each of the environments that they might expect to encounter in the future. Because our focus in this work was on the multi-modal fusion of audio and visual information, we did not attempt to apply compensation techniques to reduce the potential deleterious effects of background noise or other poor environmental conditions present in the audio-visual data.

Testing Location	Detection Failures (%)		
	Audio	Video	Both
Office	0.00	2.00	0.00
Cafe	0.44	0.89	0.22
Street	3.56	1.89	0.11
Average	1.33	1.60	0.11

Table 1: Failure rates for detecting either the speech signal, the user's face, or both, for different testing conditions.

4.2. Speech and Face Detection

As discussed earlier, difficult environmental conditions can cause the audio or visual components to fail to detect or locate either the speech signal or the face of the user. Table 1 shows the detection failure rates over the test utterances in each environment for both speech detection and face detection. Overall, while the rate of detection failures is higher than 1% for speech and face individually. The percentage of utterances where both fail is only 0.11%. In our experiments, we treat a detection failure as a rejection of the purported user. In the multi-modal case, we rely on a single modality's score if the other modality fails and we only outright reject the utterance if both modalities have a detection failure.

4.3. Verification Results

Table 2 summarizes the full set of results from our experiments evaluated using the equal error rate (EER) point between false rejections of true users and false acceptances of imposters. There are several trends to note. First the audio component of the system exhibits a greater sensitivity to mismatches between training and testing conditions than the visual component. The equal error rates for the audio component ranged from 2.0% all the way up to 41.3%. The equal error rates of the visual system only varied from 10.5% to 28.2% across different conditions. Averaged over all conditions, the audio and visual systems achieved comparable performances of 19.4% and 19.8% respectively. Also shown in the table is a column giving the EER of the best modality for each condition. If the best modality could be pre-determined for any condition, the average EER performance of the best selected modality would be 15.5%.

Table 2 also shows three columns of results for multi-modal speaker verification. For the pre-fusion normalization and post-fusion normalization columns, each modality's score was given an equal weight of 0.5 in the linear fusion process. The *oracle weighting* column shows the results if the pre-determined optimal weighting for a given training/testing condition were applied to pre-fusion normalized scores.

The results show that pre-normalizing each modality with T-norm achieves results that are, on average, fairly close to the *oracle weighting* results. There is an average relative degradation of only 7% from the *oracle weighting* result to the pre-fusion T-norm result using equal weighting, and in only 3 of the 12 train/test conditions is the relative degradation from the oracle result more than 10%. Furthermore, the average EER of the pre-fusion T-norm system is 36% less than the audio-only

Location		Audio	Video	Best	Multi-Modal Fusion		
Train	Test	Only	Only	Mode	Post-Fusion T-norm	Pre-Fusion T-Norm	Oracle Weighting
Office	Office	3.6	13.6	3.6	5.3	2.7	2.6
	Cafe	28.4	23.2	23.2	17.4	15.1	14.3
	Street	33.3	16.3	16.3	14.1	17.7	13.7
Cafe	Office	36.7	27.6	27.6	24.7	23.1	22.7
	Cafe	11.8	21.4	11.8	12.5	9.8	9.4
	Street	23.8	22.6	22.6	19.7	16.7	16.7
Street	Office	41.3	28.2	28.2	26.2	28.0	26.0
	Cafe	27.0	25.9	25.9	20.8	18.0	17.6
	Street	11.9	21.3	11.9	13.7	9.2	9.2
Mixed	Office	2.0	13.1	2.0	5.0	1.8	1.1
	Cafe	6.0	13.6	6.0	5.6	3.8	3.6
	Street	6.9	10.5	6.9	3.6	4.3	3.5
Average		19.4	19.8	15.5	14.0	12.5	11.7

Table 2: User verification equal error rate (EER) results over variable training and testing conditions for the audio-only, video-only, and multi-modal systems.

system, 19% less than the best individual system for each condition, and 11% better than the post-fusion T-norm system with equal weighting. Improvements were observed for the multi-modal system even in cases where one biometric was considerably more accurate than the other.

5. Conclusions

In this paper we presented recent work in the area of multi-modal user authentication using audio-visual data collected from a web-camera in multiple environments. The study shows the difficulties of performing user authentication in actual real-world environments using an inexpensive commercially available audio-visual capture device (i.e. a web camera). Our approach, which fuses scores from speaker identification and face identification systems, does not use any pre-learned fusion weights or normalization parameters. Relying solely on independent pre-normalization of the speaker and face identification scores using T-norm, our system suffers only minor degradation from a system using oracle predetermined fusion weights. Overall our multi-modal system achieves an average relative reduction of 36% over using speaker verification by itself, and an average reduction of 19% over the single best biometric system in each training/testing condition. In future work we hope to incorporate knowledge of classifier reliability into out multi-modal fusion approach.

6. Acknowledgements

This research was supported in part by the Industrial Technology Research Institute and in part by the Intel Corporation. The authors wish to thank Kate Saenko and Michael Siracusa for help with the face detection software used in this study.

7. References

[1] T. Hazen, E. Weinstein, and A. Park, "Towards robust person recognition on handheld devices using face and speaker identification technologies," in *Proc. of Int. Conf. on Multimodal Interfaces*, Vancouver, November 2003.

[2] T. Hazen, *et al*, "Multi-modal face and speaker identification on a handheld device," in *Proc. of Workshop on Multimodal User Authentication*, Santa Barbara, December 2003.

[3] D. Schultz, "Robust audio-visual person verification using web-camera video," MEng. Thesis, MIT, Cambridge, September 2006.

[4] A. Park and T. Hazen, "A comparison of normalization and training approaches for ASR-dependent speaker identification," in *Proc. of Int. Conf. on Spoken Language Processing*, Jeju Island, Korea, October 2004.

[5] E. Weinstein, *et al*, "Handheld face identification technology in a pervasive computing environment," in *Short Paper Proceedings, Pervasive 2002*, Zurich, Switzerland, August 2002.

[6] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. of Conf. on Computer Vision and Pattern Recognition*, Kauai, Hawaii, December 2001.

[7] OpenCV, http://sourceforge.net/projects/opencvlibrary/.

[8] K. Saenko, *et al*, "Visual speech recognition with loosely synchronized feature streams," in *Proc. Int. Conf. on Computer Vision*, Beijing, October 2005.

[9] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, Berlin, Germany, 1995.

[10] SvmFu, http://fpn.mit.edu/SvmFu/.

[11] J. Kittler, Y. Li, J. Matas, and M. Sanchez, "Combining evidence in multimodal personal identity recognition systems," in *Proc. of Int. Conf. on Audio and Video-Based Biometric Person Authentication*, Crans-Montana, Switzerland, March 1997.

[12] E. Bigün, J. Bigün, B. Duc, and S. Fischer, "Expert conciliation for multi modal person authentication systems by Bayesian statistics," in *Proc. of Int. Conf. on Audio and Video-Based Biometric Person Authentication*, Crans-Montana, Switzerland, March 1997.

[13] A. Ross, A. Jain, and J. Qian, "Information fusion in biometrics," in *Proc. of Int. Conf. on Audio and Video-Based Biometric Person Authentication*, Halmstad, Sweden, June 2001.

[14] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, nos. 1-3, pp. 42-54, 2000.