Towards Robust Person Recognition On Handheld Devices Using Face and Speaker Identification Technologies

Timothy J. Hazen, Eugene Weinstein and Alex Park
MIT Computer Science and Artificial Intelligence Laboratory
200 Technology Square
Cambridge, MA, USA
{hazen,ecoder,malex}@mit.edu

ABSTRACT

Most face and speaker identification techniques are tested on data collected in controlled environments using high quality cameras and microphones. However, the use of these technologies in variable environments and with the help of the inexpensive sound and image capture hardware present in mobile devices presents an additional challenge. In this study, we investigate the application of existing face and speaker identification techniques to a person identification task on a handheld device. These techniques have proven to perform accurately on tightly constrained experiments where the lighting conditions, visual backgrounds, and audio environments are fixed and specifically adjusted for optimal data quality. When these techniques are applied on mobile devices where the visual and audio conditions are highly variable, degradations in performance can be expected. Under these circumstances, the combination of multiple biometric modalities can improve the robustness and accuracy of the person identification task. In this paper, we present our approach for combining face and speaker identification technologies and experimentally demonstrate a fused multi-biometric system which achieves a 50% reduction in equal error rate over the better of the two independent systems.

Categories and Subject Descriptors

I.5.1 [**Pattern Recognition**]: Applications

General Terms

Security, Human Factors

Keywords

Face identification, handheld devices, multi-biometric interfaces, speaker identification.

ICMI'03, November 5–7, 2003, Vancouver, British Columbia, Canada.

1. INTRODUCTION

In recent years, the availability of small laptop and handheld computers has allowed computation to become more mobile and pervasive. Even formerly specialized devices such as cellular telephones now offer a range of capabilities beyond simple voice transmission, such as the ability to take, transmit and display digital images. As these devices become more ubiquitous and their range of applications increases, the need for security also increases. To prevent impostors from gaining access to sensitive information stored either locally on a device or on the device's network, security measures should be incorporated into these devices. In this paper we examine the integration of two biometric techniques, voice and face identification, into handheld devices.

Handheld 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 that yield potentially low signal-to-noise ratios. Similarly, the images captured by the devices will likely have 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 conducted an initial study into the use of two existing biometric techniques, speaker identification and face identification, within a user verification "login" scenario. The two biometric techniques we utilize are capable of highly accurate person identification under tightly constrained conditions. This paper examines their performance when utilized in mobile environments, and the potential benefit of combining these techniques to improve the robustness of person identification.

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 mobile-device paradigm in which we are conducting our experiments and the methods of data collection 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. PERSON IDENTIFICATION

2.1 Speaker Identification

Speech is a reasonable biometric for person identification because physiological characteristics such as vocal tract size and vocal fold length manifest themselves as variations in the speech signal. However, the linguistic content of an utterance is another important source of variation in the signal. In systems where the linguistic content of the speech is unknown (as is the case for surveillance tasks), text-independent systems are generally used. However, in security applications where the user is a cooperative participant in the attempt to prove their identity, the linguistic content of the speech message is typically known and can be tightly constrained. In this case, a text-dependent system can be used. When the linguistic content of the message is known, text-dependent speaker recognition systems generally perform better than text-independent systems because they can model the characteristics of the specific phonetic-content contained in the speech signal.

A common technique used in speech-based person identification is to prompt the user with a randomly generated challenge phrase. During authentication, automatic speech recognition can be used to verify that the spoken utterance matches the prompted utterance. For this type of scenario, it is both reasonable and beneficial to use the automatic speech recognition (ASR) output to leverage the phonetic constraints that give text-dependent systems their advantage. In [5], two techniques were described that use the ASR output during the analysis of the phonetic content from the test utterance.

In our speaker adaptive ASR approach, the system uses speaker-dependent speech recognizers to model each speaker. During training, phonetically transcribed enrollment utterances are used to train context-dependent phonetic models for each speaker. During testing, a speaker-independent ASR component generates a phonetic transcription from the test utterance. This transcription is then used by the system to score each segment of speech against each speaker-dependent phonetic model. Modeling speakers at the phonetic level can be problematic because enrollment data sets are typically too small to build robust speaker-dependent models for every context-dependent phonetic model. To compensate for this difficulty, we use an adaptive scoring approach in which the speaker-dependent score is interpolated with a speaker-independent score.

Mathematically, if the word recognition hypothesis assigns each feature vector x from the utterance X to phonetic unit j, then the score for speaker S_i, $p(X|S_i)$, is given by

$$\frac{1}{|X|}\sum_{x\in X}\log\left(\frac{\lambda_{i,j}p_{SD}(x|M_j,S_i)+(1-\lambda_{i,j})p_{SI}(x|M_j)}{p_{SI}(x|M_j)}\right)$$

where M_j is the model for phonetic unit j and $\lambda_{i,j}$ is an interpolation factor given by

$$\lambda_{i,j}=\frac{n_{i,j}}{n_{i,j}+\tau}$$

In this equation, $n_{i,j}$ is the number of training examples of phonetic unit j observed for speaker S_i, and τ is a global tuning parameter that is set empirically using a separate development set. The log ratio in the equation generates positive scores when the input speech is a good match to a particular speaker's models and negative scores when the speech is a poor match.

This scoring strategy results in models that capture detailed phonetic-level characteristics for a speaker when sufficient training data is available, but relies more on speaker independent models for phonetic units with sparse training data. Thus, for cases with limited training data, the speaker independent model provides a more *neutral* score. In the limiting case, if no speakers have training data for any of the phones observed in a particular test utterance, then they will all receive the same neutral score of zero, which is an intuitively consistent result.

2.2 Face Identification

The face identification framework used in our work is described largely in [9]. A face detection algorithm based on a boosting cascade of small, efficient classifiers [8] is first applied to the image, to determine if there are is indeed a face in the image; and if so, where the face is located. The face detection process allows us to crop out a face from a larger image to eliminate background pixels. The cropped face is first normalized to improve contrast in poorly- or overly-illuminated images. Next, the image is sent to the face recognition algorithm.

For face recognition, we use an approach based on support vector machine (SVM) classifiers similar to the one described in [2, 3]. The image is scaled down to 40x40 pixels, and the gray values of the pixels are treated as a 1600-dimensional feature vector. For recognition, a one-vs-all SVM scheme is used, where one classifier is trained to distinguish each person in the database from all the others [7]. In the SVM training process, for each person's classifier, that person's training images are used as positive examples, and the others' images are used as negative examples. The SVM training process finds the optimal hyperplane in the feature space that separates the positive and negative data points. Since the training data may not be separable, a mapping function corresponding to a second-order polynomial SVM kernel function [7] is applied to the data before training.

The runtime recognition process consists of computing the SVM classifier output score for each person's SVM classifier [7]. The scores are zero-centered – that is, a score of zero means the data point lies directly on the decision hyperplane, and positive and negative scores mean the data point lies on the positive and negative example side of the decision hyperplane, respectively. The absolute value of the SVM output is a multiple of the distance from the decision hyperplane, and could be normalized to produce the distance. Thus, a highly positive score represents a large degree of certainty that the data point belongs to the person the SVM was trained for, and a highly negative score represents the opposite. The output scores from all SVM classifiers make up the n-best list that we treat as our face recognition result.

For our face identification task, we collected and tested frontal face image data only. Most state of the art face identification systems (e.g. [2]) attempt to account for rotations and/or occlusions, which would be present in a typical surveillance task. However, for the handheld face identification problem, the user will be cooperating with the identification process; and in general, the user certainly will be looking at the screen of the handheld device as he or she is using it. Thus, accounting for rotated faces is not important

in this project. Generally, rotations in the face images make the problem of identification more challenging; thus, our problem is easier in this respect. Nonetheless, the variable lighting and background conditions and inexpensive camera present an orthogonal challenge, to ensure the non-triviality of our problem.

2.3 Multi-Biometric Fusion

Past work on fusing face and speaker classifiers have generally used very simple combination strategies. Poh and Korczak used a logical AND rule on the results of their independent face and speaker systems [6]. This rule is most useful when the goal is to limit false acceptances, since both classifiers must accept the user in order to produce an acceptance by the fused-classifier. Kittler *et al* explored a variety of probabilistic combination operators, including sum, product, max and min, on the *a posteriori* probability scores from their independent recognizers [4]. These basic fusion rules can be suboptimal in the circumstance that the *a posteriori* scores from any of the independent recognizers are poorly estimated.

In our work, a linear weighted summation is employed for the classifier fusion where the weights for each classifier are trained discriminatively on a held-out development set using minimum classification error (MCE) training. The MCE training optimizes the equal error rate of false acceptances and false rejections under the user verification scenario. Because the final decision only requires the combination of two independent classifiers, only one additional parameter (the ratio of the weights of the classifiers) needs to be learned. A simple brute force sampling of the parameter space is used for this MCE training. More complicated techniques (such as gradient descent training) could be used in situations where more than two scores must be fused.

3. EXPERIMENTS

3.1 The Handheld Device

For our experiments we utilized a collection of iPAQ handheld computers. Speech data were collected utilizing the built-in microphone of the iPAQ. Two different models of iPAQs were used, with two different models of off-the-shelf, inexpensive electret condenser microphones. Face data were collected using a 640x480 CCD camera located on a custom-built expansion sleeve for the iPAQ. Because of the current computation and memory limitations of the iPAQ handhelds, the images and audio are captured by the handheld device, but then transmitted over a wireless network to servers which perform the operations of face detection, face identification, speech recognition, and speaker identification. In future work we hope to improve the computational efficiency and memory footprints of our systems so they can be deployed directly on small handheld devices.

3.2 The Login Scenario

Our experiments were conducted using a login scenario that combined face and speaker identification techniques to perform the multi biometric user verification process. When "logging on" to the handheld device, users snapped a frontal view of their face, spoke their name, and then spoke a prompted lock combination phrase consisting of three randomly selected two digit numbers (e.g. "*25-86-42*"). The system recognized the spoken name to obtain the "claimed identity". It then performed face verification on the face image and speaker verification on the prompted lock combination phrase. Users were "accepted" or "rejected" based on the combined scores of the two biometric techniques.

3.3 Data Collection

For our set of "enrolled" users, we collected face and voice data from 35 different people. Each person performed eight short enrollment sessions, four to collect image data and four to collect voice data. Each image collection session consisted of the user taking 25 frontal facial images in a variety of rooms in our lab with different lighting conditions. No specific controls were placed on the distribution of the locations and lighting conditions; users were allowed to self-select the locales and lighting conditions of the image. For voice collection, each user recited 16 prompted lock-combination phrases in each session. Each session was typically collected on a different day, with the time span between sessions often spanning several days and occasionally a week or more. Each enrollment session typically lasted less than 5 minutes with the total enrollment time taking approximately 30 minutes on average. In total this yielded 100 images and 64 speech samples for enrolled user for training. An additional set of four enrollment sessions of audio data (i.e., 64 additional utterances) from 17 of the training speakers was available for development evaluations and multi-biometric weight fusion training.

For our evaluation, we collected 16 sample login sessions from 25 of the 35 enrolled users. This yielded 400 unique utterance/face evaluation pairs from enrolled users. We also collected 10 impostor login sessions from 20 people not in the set of enrolled users for an additional 200 utterance/face evaluation pairs from unenrolled people.

We used the evaluation data to perform our user verification experiments. Each utterance/face pair from in-set speakers was used as a positive example of that user. This yielded a total of 400 positive examples for our evaluation. Each utterance/face pair from each in-set user could also be used as an impostor for the other 34 users in the enrolled set. This yielded 13600 impostor examples from in-set speakers. Each utterance/face pair collected from out-of-set impostors was also used to generate an impostor example for each of the 35 users in the enrolled set. This yielded 7000 impostor examples from users not in the enrollment set. In general, it is expected that impostors that have never been observed by the system will generate more classification errors than enrolled users who try to impersonate other enrolled users. This is because the models are trained to discriminate between users observed in the training data and thus may not generalize well to unseen users.

3.4 Training

The face and speaker systems were trained on the enrollment data for the 35 enrolled users. To train the fusion weights, one of the four face enrollment sessions was held out and a development face ID system was trained on the remaining three face sessions. Face identification scores from this held-out set were pairwise combined with speaker identification scores generated for utterances from the existing speaker identification development set. The true in-set examples and in-set impostor examples were provided to the MCE weight training algorithm previously described to generate the multi-biometric fusion weights.

Table 1: User verification results expressed as equal error rates (%) on three systems (face only, speaker only, and multi-biometric fusion) under two impostor conditions (known in-set impostors vs. unknown out-of-set impostors).

System	In-set impostors	Out-of-set Impostors
Face	5.85	7.30
Speaker	0.66	1.77
Fused	0.40	0.89

3.5 Experimental Results

Table 1 shows our user verification results for three systems (face ID only, speaker ID only, and our full multi-biometric system) under two different impostor conditions (using only known in-set impostors vs. using only unknown out-of-set impostors). Several observations should be made from these results. First, it is interesting to note the difference in performance when the system encounters impostors who are part of the enrollment set (i.e., enrolled users trying to impersonate other enrolled users) vs. the performance on impostors who are not known to the system. The fused system experienced a 120% increase in the equal error rate when unknown impostors were used instead of enrolled impostors. This shows the importance of evaluating the system using people that are not part of the training data.

Second, the speaker identification system is performing much better than the face identification system. This can partially be explained by the fact that we are using a fairly advanced speaker identification system, but we have implemented a relatively basic face identification system for these experiments. The primary reason for this decision was based on computational requirements. The more advanced face identification techniques that were potential candidates for our system had too high of a computational latency for use in a real time system. Because the construction of a real-time prototype system was one of our goals (which is an achievement we have reached), computational efficiency was a primary factor in our design decisions. In future work we hope to integrate a more state-of-the-art face identification algorithm into our system.

Finally, despite the gap in performance between the face and speaker systems, the fused system is still able to achieve a significant improvement over the better of the two systems. When examining the equal error rate on the experiment using out-of-set impostors, the fused system achieves an equal error rate which is 50% less than the equal error rate of the speech only system.

To examine the degradation that might be experienced when our speaker identification technique is utilized in a mobile environment, we compared the performance of closed-set speaker recognition on the mobile handheld data set against the performance of our system on the tightly constrained YOHO corpus, which uses the same lock combination phrase approach that we employed [1]. It is important to note that the YOHO corpus was collected using a single close-talking telephone handset in a quiet office, and thus does not suffer from the degradations that are present in our mobile devices due to the low quality far-field microphone and the variable background conditions. In [5], it was shown that our system's speaker recognition error rate was 0.31% over YOHO's closed-set of 138 speakers. Using our 400 utterance in-set speaker evaluation set, our system's speaker recognition error rate was 0.25% over our closed set of 35 enrolled speakers (i.e., only one misrecognition in 400 trials). Thus we have achieved roughly the same error rate as on YOHO, but only with a much smaller set of speakers.

4. SUMMARY AND FUTURE WORK

In summary, our initial study in biometric fusion for user verification has demonstrated the benefits of combining face and speaker identification even when one of the biometric techniques has superior performance to the other. A 50% reduction in user verification equal error rate was observed when our initial speaker identification system was fused with a face identification system. Though our initial study demonstrated the feasibility of our approach, our initial evaluation set is quite small. In future work we plan to expand the size of evaluation set and examine the specific types of errors the system makes. We also plan to investigate the performance of the system under the conditions where impostors are specifically selected based on resemblances of their voice or facial properties (i.e., same gender or ethnicity) to particular enrolled users.

5. ACKNOWLEDGMENTS

The authors wish to thank Dave Dopson and Ken Steele, who helped in the development of the application, and Bernd Heisele, who has provided the face identification algorithms and on-going assistance. This work was supported by an industrial consortium supporting the MIT Oxygen Alliance.

6. REFERENCES

[1] J. Campbell. Testing with the YOHO CD-ROM voice verification corpus. In *Proc. ICASSP*, pages 341–344, Detroit, MI, May 1995.

[2] B. Heisele, P. Ho, and T. Poggio. Face recognition with support vector machines: Global versus component-based approach. In *Proc. ICCV*, volume 2, pages 688–694, Vancouver, Canada, 2001.

[3] P. Ho. Rotation invariant real-time face detection and recognition system. Technical Report 2001-010, MIT Artificial Intelligence Lab., Cambridge, MA, 2001.

[4] J. Kittler, Y. Li, J. Matas, and M. Sanchez. Combining evidence in multimodal personal identity recognition systems. In *Proc. AVBPA*, pages 327–334, Crans-Montana, Switzerland, March 1997.

[5] A. Park and T. J. Hazen. ASR dependent techniques for speaker identification. In *Proc. of ICSLP*, pages 1337–1340, Denver, CO, September 2002.

[6] N. Poh and J. Korczak. Hybrid biometric person authentication using face and voice features. In *Proc. AVBPA*, pages 348–353, Halmstad, Sweden, June 2001.

[7] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Spinger-Verlag, Berlin, Germany, 1995.

[8] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. CVPR*, Kauai, HI, 2001.

[9] E. Weinstein, *et al.* Handheld face identification technology in a pervasive computing environment. In *Short Paper Proceedings, Pervasive 2002*, pages 48–54, Zurich, Switzerland, 2002.