Quality-Based Score-level Fusion for Secure and Robust Multimodal Biometrics-based Authentication on Consumer Mobile Devices

Mikhail Gofman, Sinjini Mitra, Kevin Cheng, and Nicholas Smith
California State University, Fullerton, California, USA
Email: {mgofman, smitra}@fullerton.edu and {kevincheng99, nicholastoddsmith}@csu.fullerton.edu

***Abstract*—Biometric authentication is a promising approach to access control in consumer mobile devices. Most current mobile biometric authentication techniques, however, authenticate people based on a single biometric modality (e.g., iPhone 6 uses only fingerprints), which limits resistance to trait spoofing attacks and ability to accurately identify users under uncontrolled conditions in which mobile devices operate. These challenges can be alleviated by *multimodal biometrics* or authentication based on multiple modalities. Therefore, we develop a proof-of-concept mobile biometric system which integrates information from face and voice using a novel score-level fusion scheme driven by the quality of the captured biometric samples. We implement our scheme on the Samsung Galaxy S5 smartphone. Preliminary evaluation shows that the approach increases accuracy by 4.14% and 7.86% compared to using face and voice recognition individually, respectively.**

Keywords–Multimodal biometrics; quality; score-level fusion; mobile

I. INTRODUCTION

Biometric authentication is the science of identifying people based on their physical and behavioral traits, such as face and voice. Recent advances in mobile technology have enabled such authentication in consumer mobile devices. Although generally regarded as more secure than passwords, most state-of-the-art mobile biometric authentication approaches are unimodal: they identify people based on a single trait. To bypass a unimodal system, an attacker only needs fabricate the single trait the system uses for identification [1]. Unimodal mobile biometric systems also have difficulty accurately recognizing users in conditions known to distort the quality of biometric images (e.g., poor lighting affecting the visibility of a face [2]).

We present the design, implementation, and performance evaluation of a proof-of-concept system to demonstrate that a promising approach to improve the security and robustness of mobile biometric authentication is *multimodal biometrics*, which uses multiple traits to identify people. Our contributions are as follows:

1) We study the effects of face and voice sample quality on recognition accuracy in mobile devices.
2) We develop a multimodal biometric system integrating information from face and voice through a novel *quality-based score-level fusion* scheme, which improves recognition accuracy by letting the modality with higher quality sample have a greater impact on the authentication outcome. Such a scheme lets the system adapt to varying background conditions, which affect the quality of biometric images.
3) We evaluate the system using our database of face and voice samples captured using a Galaxy S5 smartphone in a variety of background conditions. The results indicate that the approach achieves higher recognition accuracy than unimodal approaches based solely on face or voice.

The rest of the paper is organized as follows. Section II discusses the role of quality in mobile biometric authentication, Section III introduces multimodal biometric systems. We present our quality-based score-level fusion scheme in Section IV followed by the results in Section V. Finally, we conclude in Section VI.

II. ROLE OF QUALITY IN MOBILE BIOMETRICS

A low-quality biometric sample, such as low resolution face photograph or noisy voice recording, can cause a biometric algorithm to incorrectly identify an impostor as a legitimate user (*false acceptance*) or a legitimate user as an impostor (*false rejection*). Capturing high-quality samples on mobile devices is especially difficult because (i) people often operate mobile devices in insufficiently lit and noisy environments (e.g., malls and restaurants), choose less-than-optimal camera angles, and might have dirty fingers [2]; and (ii) biometric sensors in consumer mobile devices often trade sample quality for portability and lower costs, leaving them vulnerable to trait spoofing attacks [3]. We believe that these challenges can be addressed via multimodal biometrics.

III. MULTIMODAL BIOMETRICS ON MOBILE DEVICES

Multimodal biometrics require users to authenticate using multiple relatively-independent traits, which adds layers of security by forcing attackers to fabricate multiple traits. Also, identifying information from modalities with high-quality images can compensate for the missing/inaccurate identifying information in low-quality images from other modalities.

In multimodal biometric systems, information from different modalities can be consolidated (i.e., fused) at the decision-, match score-, or feature-level [6]. We integrate face and voice modalities using score-level fusion, because it is considered more effective than decision-level fusion and less complex and computationally expensive than feature-level fusion.

IV. QUALITY-BASED SCORE-LEVEL FUSION

Sample quality can drastically affect recognition accuracy; therefore we integrate it into our multimodal scheme. We assess the quality of face images based on luminosity, sharpness, and contrast [4] and use the signal-to-noise ratio (SNR) approach [5] to assess the quality of voice samples. Once assessed, the metrics are normalized using the z-score normalization method. This particular method was selected since it is a commonly used normalization method, is easy to implement, and is highly efficient [9].

For face recognition, we use *FisherFaces*, which works well when images are captured under varying conditions, as is the case with mobile devices [7]. The algorithm uses pixel intensities in the image as identifying features. For voice recognition, we use Mel-Frequency Cepstral Coefficients (MFCCs) as the identifying features in a Hidden Markov Model (HMM)-based identification method [8]. After training

 ISBN: 978-1-61208-438-1

these algorithms with samples from users, they are used to match samples supplied during authentication, and are the basis of the score-level fusion scheme, as described below.

Let t_1 and t_2 denote the quality scores for the face and voice samples in the training data, respectively. During authentication, we calculate the quality scores Q_1 and Q_2 of the two biometrics from the test data and determine their proximity to t_1 and t_2, respectively. We then compute the weights of the face and voice modalities w_1 and w_2, as $w_i = \frac{p_i}{p_1+p_2}$, so that w_1 + w_2 = 1, where p_1 and p_2 are percent proximities of Q_1 to t_1 and Q_2 to t_2, respectively. The closer Q_i is to t_i, the greater is the weight assigned to the corresponding modality, which ensures the effective integration of quality in the final authentication process. Next, the matching scores S_1 and S_2 are obtained from face and voice recognition algorithms. The overall match score is then computed using the *weighted sum rule*: $M = S_1w_1 + S_2w_2$. If $M \geq T$ (T: pre-selected threshold), the system accepts the person as authentic; otherwise, it declares the person an imposter.

While using the above scheme, it is important to exercise caution to ensure significant representation of both modalities in the fusion process. For example, if Q_2 differs greatly from t_2 but Q_1 is close to t_1, the face modality will dominate the authentication process, resulting in a nearly unimodal scheme based on the face biometric. Thus, a mandated benchmark is required for each quality score to ensure that the system denies access if the benchmarks for both scores are not met.

V. Results

A. The Dataset

Due to the unavailability of a diverse multimodal mobile biometric database, we created one with videos from 54 people of different genders and ethnicities. They held a phone camera in front of their face while saying a certain phrase. The videos were recorded in various real-world settings using a Samsung Galaxy S5 smartphone. The faces display the following variations: (1) four expressions: neutral, happy, sad, angry, and scared; (2) three poses: front and sideways (left and right); and (3) two illumination conditions: uniform and partial shadows. The voices in videos have different levels of background noise, from traffic noises to music and chatter, and voice distortions like raspiness. Twenty popular phrases were used (e.g., *unlock* and *football*). The database is still in development and will be made available to researchers upon completion.

B. Performance Results

We implemented our score-level fusion scheme on the Android-based Samsung Galaxy S5 device. Table I shows preliminary performance results. We measure recognition accuracy using equal error rate (EER), which is traditionally used in biometrics applications and is the value that produces the best possible combination of the False Acceptance Rate (FAR) and False Rejection Rate (FRR) (i.e., where FAR and FRR are equal) [6]. The final results were obtained by selecting a random set of five users from the database and training the face and voice algorithms with 40 face images and 40 voice samples of these users. Most samples were automatically extracted from one good-quality video and few samples were extracted from low-quality videos. For testing, we used 80 combinations of randomly selected face frames and voice samples from videos of varying quality. The experiment was repeated for 1000 different training/testing combinations of users, and the face, voice, and score-level fusion EERs and training and authentication execution times were average. According to the table, our quality-based score-level fusion approach improves accuracy by 4.14% and 7.86% compared unimodal face and voice recognition approaches, respectively.

Note: Good-quality training samples are important because their quality metrics provide a baseline for judging qualities of samples supplied during authentication. Adding a few noisy training samples also increases the chances of recognizing the user in similar noisy conditions. Also, although the training times are longer than authentication times (common in classification problems), training happens only once when the user registers his/her training data with the device. Authentication is real-time and should require less time, as is the case here.

TABLE I. QUALITY-BASED SCORE-LEVEL FUSION EER RESULTS.

Modality	EER	Training Time (sec)	Auth. Time (sec)
Face	18.70%	575.491	0.2133
Voice	22.42%	295.692	0.0728
Score-level Fusion	14.56%	871.183	0.2861

C. Analysis of Sample Quality

We briefly discuss the quality of face and voice samples in our database and its complex relationship to recognition accuracy. We find that face luminosity and contrast metrics exhibit bimodal distributions caused by different conditions, such as shadows. Voice SNR exhibits normal distribution for good-quality samples, sick voices, and voices with chatter in the background. These distributions can provide useful guidance for designing automatic sample quality enhancement mechanisms in mobile devices. We also observe that variations in luminosity and pose are greater challenges to minimizing the face recognition EER, compared to sharpness and contrast. However, there are important exceptions such as, images distorted by motion blur, matching poorly due to the differences in sharpness despite similar luminosity. These findings illustrate the complex sample-quality challenges facing mobile biometrics. They also lay the groundwork for devising a statistical framework for predicting optimal modality weights in our scheme and determining the quality thresholds for acceptable error rates.

VI. Conclusions and Future Work

The preliminary results show that multimodal biometrics can improve biometrics-based authentication in consumer mobile devices. Our next step is to refine the method to reduce EER more and incorporate other modalities (e.g., ears and fingerprints).

References

[1] D. Smith, A. Wiliem, and B. Lovell, "Face recognition on consumer devices: Reflections on replay attacks," IEEE Transactions on Information Forensics and Security, vol. 10, 2015, pp. 736–745.

[2] C. Bhagavatula, B. Ur, K. Iacovino, S. M. Kywe, L. F. Cranor, and M. Savvides, "Biometric authentication on iPhone and Android: Usability, perceptions, and influences on adoption," presented at Proceedings of the NDSS Workshop on Usable Security 2015 (USEC). San Diego, California, February 2015. URL: http://www.blaseur.com/papers/usec15_talk.pdf [accessed: 2015-25-8].

[3] "The trouble with Apples Touch ID fingerprint reader, Wired.com, December 3, 2013, URL: http://www.wired.com/2013/12/touch-id-issues-and-fixes/ [accessed: 2015-25-8].

 ISBN: 978-1-61208-438-1

[4] K. Nasrollahi and T. B. Moeslund, “Face Quality Assessment System in Video Sequences,” Biometrics and Identity Management: First European Workshop (BIOID), May 7-9, 2008 Roskilde, Denmark. Springer Berlin Heidelberg, 2008, pp. 10–18, URL: http://dx.doi.org/10.1007/978-3-540-89991-4_2 [accessed: 8-25-2015].

[5] M. Vondrášek and P. Pollak, “Methods for speech SNR estimation: Evaluation tool and analysis of VAD dependency,” Radioengineering, vol. 14, no. 1, 2005, pp. 6–11.

[6] A. K. Jain and A. Ross, “Multibiometric systems,” Communications of the ACM, vol. 47, no. 1, 2004, pp. 34–40.

[7] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, “Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, 1997, pp. 711–720.

[8] D. Hsu, S. M. Kakade, and T. Zhang, “A spectral algorithm for learning Hidden Markov Models,” Journal of Computer and System Sciences, vol. 78, no. 5, 2012, pp. 1460–1480.

[9] A.K. Jain, K. Nandakumar, and A. Ross, “Score normalization in multimodal biometric systems,” *Pattern Recognition*, 2005, Vol. 38, pp. 2270–2285.

 ISBN: 978-1-61208-438-1