

CASCADING MULTIMODAL VERIFICATION USING FACE, VOICE AND IRIS INFORMATION

Ping-Han Lee[†], Lu-Jong Chu[†] Yi-Ping Hung[‡] Sheng-Wen Shih[§] Chu-Song Chen[¶], Hsin-Min Wang[¶]

[†]Institution of Computer Science and Information Engineering, National Taiwan University, Taiwan

[‡]Graduate Institute of Networking and Multimedia, National Taiwan University, Taiwan

[§]Department of Computer Science and Information Engineering National, Chi Nan University, Taiwan

[¶]Institute of Information Science, Academia Sinica, Taiwan

ABSTRACT

In this paper we propose a novel fusion strategy which fuses information from multiple physical traits via a cascading verification process. In the proposed system users are verified by each individual modules sequentially in turns of face, voice and iris, and would be accepted once he/she is verified by one of the modules without performing the rest of the verifications. Through adjusting thresholds for each module, the proposed approach exhibits different behavior with respect to security and user convenience. We provide a criterion to select thresholds for different requirements and we also design an user interface which helps users find the personalized thresholds intuitively. The proposed approach is verified with experiments on our in-house face-voice-iris database. The experimental results indicate that besides the flexibility between security and convenience, the proposed system also achieves better accuracy than its most accurate module.

1. INTRODUCTION

With the advancement in biometrics, which verifies the identities of users via user's physical traits, it becomes a legitimate method for identity verification in recent years. One of the merit of biometrics is that since physical traits are intrinsic to each person, they are hard to be forged and missed. Physical traits applied in the literature of biometrics include fingerprints, iris, face, hand vein, signature, voice, etc. [5]. While most researchers focus on unimodal biometrics (i.e., utilizing only one physical trait), it has been reported that multimodal biometrics outperform the unimodal ones recently [7] [5].

One crucial problem in multimodal biometrics is how to combine information from multiple physical traits. Generally speaking [5], the strategy of fusion can be categorized into feature level [3], matching score level [1] and decision level. In [3], the feature vectors of palmprint and hand geometry are concatenated as a new feature vector. The authors utilize the

normalized correlation coefficient to compute the matching scores between two feature vectors. In [1], matching scores resulting from face and voice classifiers are combined as a new two-dimensional feature vector. The author train support vector machine based on this new feature vector to perform the classification. For decision level fusions, final decisions are typically obtained by voting on multiple classifiers.

Despite of the improvement in accuracy, there are two issues in the fusion strategies mentioned above. One issue is that when classifiers with high accuracy such as an iris classifier, the improvement through the fusion may be limited. In [8] the authors combine face and iris for identity verification in the matching score level. They point out when iris samples are well aligned, the accuracy of the system fusing face and iris is worse than a stand-alone iris verification module. The other issue is the multiple sample acquisition processes in multimodal systems is very intrusive to users, especially when iris verification module exists in systems.

To address the two issues, we propose a new fusion strategy that fuses multiple physical traits in a cascade structure, in which users are verified with individual modules sequentially in separate stages, each contains an unimodal module. Once the user is verified with one module, he/she is accepted and the verifications for the rest of the modules are avoided. The modules in the proposed cascade structure are in turns of face, voice and iris. Note that this sequence is in the order of increasing module accuracy, and it happens to be also in the order of increasing user intrusiveness. Through adjusting the thresholds in each module, the proposed system has different behaviors by levitating between accuracy and user intrusiveness.

2. SYSTEM OVERVIEW

The login process of the proposed system is a cascade structure illustrated in Fig. 1. In an environment with access control where there is a hallway and a door, an user is verified with the face verification module without his/her notice when approaching the door. If the face verification module accepts

This work is supported in part by National Science Council, Taiwan, under grant NSC 94-2213-E-002-026 and by MOEA Technology Development Program for Academia under grant 95I513.

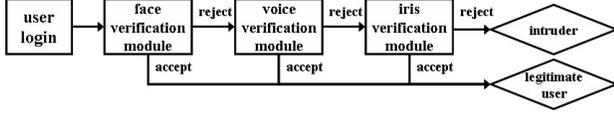


Fig. 1. System overview.

the user, the door will open automatically when he/she comes in front of it. Otherwise, the system begins to ask the user some questions and utilizes voice verification module to verify his/her identity. The door will open when the user passes the voice verification. If again the user fails the voice verification, then when he/she comes in front of the door, the system will request the user to perform the iris verification. An user will be regarded as an intruder if he/she fails all the three verification modules and the door will be blocked. On the other hand, an user is regarded as a valid user if he/she passes one of the modules in any stage.

The reason for this setup is twofold. Modules in the cascade structure in turns of face, voice and iris is in the increasing order of **user intrusiveness** and it happens to be also in the increasing order of **verification accuracy**. In this setup, users may have chance to be accepted by the face verification module, which is less intrusive, and avoid voice and iris verifications, which are more intrusive. On the other hand, since the module in the latter stage is more accurate than the module in the former stage, falsely rejected users may be correctly accepted in the latter stage. Note that although the falsely accepted users will not be further verified in the latter stage of the proposed cascading framework, we can set strict thresholds for face and voice verification modules to reduce the false accept rate. In the next section we will describe the face, voice and iris verification modules we applied in our system.

3. THREE VERIFICATION MODULES

3.1. Face Verification Module

We apply Eigenface [6] as our face verification module. We select the reduced dimension in Principle Component Analysis in the sense that it retains about 95% reconstruction accuracy.

3.2. Speech Verification Module

We follow [4] to build our speech verification module. We use Gaussian Mixture Models (GMMs) to model digits numbers 0 to 9. The verification is based on the log-likelihood ratio (LLR) [4], which helps minimize the non-speaker related variations in the test utterance scores, allowing stable decision thresholds to be set. Please refer to [4] for more details.

3.3. Iris Verification Module

Recently an iris recognition approach using multi-scale edge-type matching is proposed [2]. The authors utilize a new kind of feature which encodes the step/ridge edges in human iris. Two types of edges are detected via derivative of Gaussian and Laplacian of Gaussian, respectively. The approach is reported to have competing accuracy with Gabor filters based approaches but has lower computational cost. In our work we apply [2] as our iris verification module.

4. CASCADING MULTIMODAL VERIFICATION

4.1. Threshold-Performance Entry (TPE)

We have three individual modules in our system. Each module outputs a matching score in the interval of $[0, 1]$. Thus the thresholds in the proposed system are triplets (T_f, T_v, T_i) , where T_f , T_v and T_i are thresholds for face, voice and iris verification modules, respectively. The Receiver Operator Characteristic (ROC) analysis is done by sampling **threshold triplets** in three-dimensional space (T_f, T_v, T_i) , which is different from unimodal systems with one-dimensional threshold space. Under each threshold triplet, we perform the verification process described in section 2 for all the testing samples (each consists of one single face, voice and iris data) and obtain the false acceptance rate (FAR) and false rejection rate (FRR) with respect to the whole system (denoted as FAR_s and FRR_s , respectively). Besides the FAR_s and FRR_s , the $FRRs$ for the face verification module (FRR_f) and the voice verification module (FRR_v) are also important. FRR_f indicates the probability a legitimate user will be requested to performed a voice verification. Small FRR_f is desired when we want to login successfully without voice verifications as much as possible. Likewise, with small FRR_v , a legitimate user will have better chance to login successfully in the voice verification without an additional iris test. Note that in the proposed system, although the three modules have their own $FARs$, only the FAR of the system (FAR_s) is important. Since once an intruder is falsely accepted by any module, the system regards this user as a legitimate one and do not ask he/she for further verifications, $FARs$ for individual modules does not affect the verification process as FRR_f and FRR_v . Hence we consider only the FAR_s for the security issue. We define the **threshold-performance entry** (TPE) as a tuple $\{T_f, T_v, T_i, FRR_f, FRR_v, FRR_s, FAR_s\}$.

4.2. Threshold Triplet Selection

In the environment that requires high level of **security**, threshold triplets resulting in small FAR_s are desirable. On the other hand, if users care more about **convenience** than security, small FRR_f , FRR_v and FRR_s are preferred. To meet a specific requirement toward security and convenience, we

sample plentiful threshold triplets and obtain the corresponding TPEs. Then we select the best TPE using the criteria below:

$$\arg \min_{T_f, T_v, T_i} (w_{sa} \cdot FAR_s + w_{sr} \cdot FRR_s + w_f \cdot FRR_f + w_v \cdot FRR_v), \quad (1)$$

where w_{sa} , w_{sr} , w_f and w_v are corresponding weights. We further enforce $w_{sa} + w_{sr} + w_f + w_v = 1$.

With TPEs selected through different **weights settings**, the proposed system exhibits different behaviors towards security and convenience. In Table 1 we list three weights settings. In **Security Mode**, we emphasize low FAR_s and we also penalize high FRR_s moderately. Since we neglect FRR_f and FRR_v totally, the two values in the selected TPE may be large. Thus the users will be asked to perform all the three verification tests frequently. In the **Convenience Mode**, better user convenience is gained through sacrificing some system security. We increase weights for all FRRs and give a small weight for FAR_s . We also list a mode that favors neither of the four **attributes** (FRR_f , FRR_v , FRR_s and FAR_s), which is **Normal Mode**.

Table 1. Three different weights settings.

Mode	w_{sa}	w_{sr}	w_f	w_v
Security	0.6	0.4	0	0
Convenience	0.1	0.2	0.4	0.3
Normal	0.25	0.25	0.25	0.25

4.3. User Interface

To help users select the threshold triplets that best satisfy their needs, we design an user interface illustrated in Fig. 2. To provide an intuitive interface to users, we define four bars which are **Intruder Rejection**, **Face Login Rate**, **Voice Login Rate** and **Iris Login Rate** as $(1 - FAR_s)$, $(1 - FRR_s)$, $(1 - FRR_f)$ and $(1 - FRR_v)$, respectively. Users can use these four bars on the left portion of the panel to adjust the four main attributes hence satisfy their needs, with the aid of the four buttons of different weights setting in the lower right region of the panel. The upper right region of the panel shows the detailed thresholds, FARs and FRRs for the face, voice and iris modules. The resulting thresholds will be applied in the proposed cascading framework and thus the system behaviors will change as users requested.

Each time an user set the value of certain attribute to x , TPEs with the corresponding values of attributes y which satisfy $x - \varepsilon < y < x + \varepsilon$ are picked as a subset of all TPEs, where ε is the tolerance. Then we use (1) together with the selected weights setting to select one best TPE in this subset. The suggested threshold triplet is returned on the upper right region of the panel.

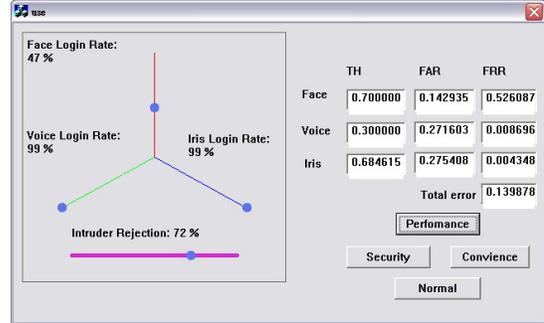


Fig. 2. The user interface.



Fig. 3. Some samples from our in-house face-voice-iris database. The first row shows face samples and the second row presents iris samples.

5. EXPERIMENTAL RESULTS

We conduct experiments on our in-house multimodal biometrics database, which consists of face, voice and iris data. We have two sessions of data in our database, each session contains 10 triplets of samples. Data of two sessions are taken between a week. We have 19 registered users with complete data of two sessions in this database, which are composed of 14 males and 5 females. There is an additional single session of 5 people who can serve as intruders. The face samples include some variations in lighting, pose and expression. We manually crop faces and normalize these faces to the size of 50x50 pixels, with zero mean and unit variance. Iris samples are cropped automatically and resized to 256x128 pixels. Fig. 3 illustrates some face and iris samples. Voice data are recorded in a controlled environment, we also subtract background noise.

In sampling threshold triplets, 10 thresholds for each modules are equally sampled in the interval of $[0, 1]$ for face and voice verification modules. We sample 20 thresholds for iris verification module since the iris verification module is more sensitive to different thresholds. In our experiment, we use data in session 1 for training. We use data in session 2 to obtain total 2000 TPEs under the sampled threshold triplets. To compare the accuracy of the proposed system with the three individual modules, we select threshold triplets that result in small **total error** ($FAR_s + FRR_s$) and plot them in Fig. 4 together with the Receiver Operating Characteristic (ROC) curves of the three modules. Note that each point of the proposed system in this figure results from a TPE. We

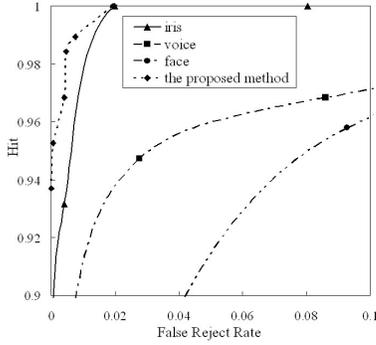


Fig. 4. Comparison of ROC curves.

can see from this figure that the proposed multimodal system achieve better accuracy than its most accurate unimodal module, which is the iris module.

The detailed verification results of the proposed system are shown in Table 2, where **VTPH/ITPH** stands for voice/iris test required per hundred logins, and **HTE** is the half total error. In this Table, the rows **face**, **voice** and **iris** shows the results of the three stand-alone modules. The rest of the rows are results of the proposed system with different configurations of threshold triplets. Note that we use the user interface described in section 4.3 with different weights settings described in Table 1 to select these threshold triplets. We can see from this table that for individual modules, the iris verification module is most accurate, secondly the voice verification module, thirdly the face verification module. As to the proposed system, under the **Security Mode** we have least FAR_s , but VTPH and ITPH are both relatively large. Under the **Convenience Mode**, the VTPH and ITPH are substantially small compared with the security mode. However, FAR_s is large in this mode. The behavior of the **Normal mode** is between the security and the convenience mode. Note the accuracy of the proposed system outperforms that of the iris verification module. The reason is some samples that will be falsely rejected by the individual iris verification module may be correctly accepted in the face or voice verification module in the proposed system. Thus under the same threshold the proposed system may sometimes have smaller FRR than the iris verification module.

Table 2. Comparison of the three individual modules and the proposed system with different configurations.

	VTPH	ITPH	FRR_s	FAR_s	HTE
face	-	-	9.27%	4.21%	6.74%
voice	-	-	2.75%	5.26%	4.01%
iris	-	-	1.99%	0%	0.99%
Security	72.63	12.62	1.05%	0.76%	0.91%
Normal	36.84	4.31	1.05%	1.84%	1.45%
Convenience	15.26	0.72	0%	3.92%	1.96%

6. CONCLUSION REMARKS AND FUTURE WORKS

We propose a cascading multimodal verification system in which users are verified by each individual modules sequentially in turns of face, voice and iris. The behavior of the proposed system is determined by threshold performance entries (TPEs) of different threshold triplets. We propose a criterion to select the best threshold triplet that satisfy different needs toward security and convenience. An user interface is designed to help users find their personalized threshold settings intuitively. Besides its flexibility, the proposed system also outperforms its most accurate individual module, which is the iris verification module. The effectiveness of the proposed system is validated through our in-house database consists of face, voice and iris traits of human.

The current framework provides much convenience for users, since once an user is accepted by any individual module, he/she is accepted. The cascading framework can be implemented in another way that an user must be accepted by all the modules sequentially. Once rejected, he/she will not further be verified by other modules. We will study this different framework which stress on the security issue in our future work.

7. REFERENCES

- [1] H. T. Cheng, Y. H. Chao, S. L. Yeh, C. S. Chen, H. M. Wang, and Y. P. Hung. An efficient approach to multimodal person identity verification by fusing face and voice information. In *ICME*, pages 542–545, 2005.
- [2] C. T. Chou, S. W. Shih, W. S. Chen, and V. Cheng. Iris recognition with multi-scale edge-type matching. In *ICPR*, volume 4, pages 545–548, 2006.
- [3] A. Kumar, D. C. M. Wong, H. C. Shen, and A. K. Jain. Personal verification using palmprint and hand geometry biometric. In *AVBPA*, pages 668–678, 2003.
- [4] D. A. Reynolds. Speaker identification and verification using gaussian mixture speaker models. In *Speech Communication*, number 17, pages 179–192, 1995.
- [5] A. Ross and A. K. Jain. Multimodal biometrics: An overview. In *EUSIPCO*, pages 1221–1224, 2004.
- [6] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991.
- [7] P. Verlinde, G. Chollet, and M. Acheroy. Multi-modal identity verification using expert fusion. *Information Fusion*, 1:17–33, 2000.
- [8] Y. Wang, T. Tan, and A. K. Jain. Combining face and iris biometrics for identity verification. In *AVBPA*, pages 805–813, 2003.