An Exploration of the Impacts of Three Factors in Multimodal Biometric Score Fusion: Score Modality, Recognition Method, and Fusion Process

YUFENG ZHENG
ERIK BLASCH

Operational applications for human identification require high credibility in order to determine or verify a person's identity to a desired confidence level. Multimodal biometric score fusion (MBSF) can significantly improve detection, recognition, and identification performance of humans. The goals of this research are to explore the impact of each factor in a MBSF process and to determine the most important (key) factor. The following are three main factors that will be investigated and discussed in this paper: score modality, recognition method, and fusion process. Specifically, score modality is defined as imaging device (hardware) for biometric data acquisition. Recognition method is defined as matching algorithm (software) for biometric score calculation. A fusion process such as arithmetic fusion, classifier-based fusion, or density-based fusion, is used to combine biometric scores. The hidden Markov model (HMM) is also applied to the MBSF process as a baseline comparison. The accuracy of human identification is measured with a verification rate. A new metric, *relative rate increase* (RRI), is proposed to evaluate the performance improvement using score fusion. Several recognition methods (two to four matchers) and four fusion processes (mean, linear discriminant analysis, k-nearest neighbors, and HMM) are compared over four multimodal databases in our experiments. The experimental results show that the score modality is the dominant factor in biometric score fusion. The fusion process becomes more important in a single modality fusion. Adding more recognition methods into the fusion process has the least impact on fusion improvement.

Manuscript received December 11, 2013; revised June 10, 2014 and October 1, 2014; released for publication October 2, 2014.

Refereeing of this contribution was handled by Ramona Georgescu.

This work was supported by the Department of Defense Research and Education Program (under Grant No. W911NF-12-1-0065) administered by the Army Research Office and the Office of Naval Research.

Authors' addresses: Y. Zheng, Alcorn State University, 1000 ASU Drive, Alcorn State, MS 39096, USA (e-mail: yzheng@alcorn.edu). E. Blasch, US Air Force Research Laboratory (AFRL), 525 Brooks Rd, Rome, NY 13441, USA (e-mail: erik.blasch@us.af.mil).

1557-6418/14/$17.00

1 INTRODUCTION

Data fusion can be performed at different levels, e.g., pixel, feature, score, and decision. Accordingly, the corresponding data preprocessing is also different for each level. For example, pixel-level and feature-level image fusion usually require registration and normalization (to align multiple images); while score-level fusion only requires normalization. Decision-level fusion such as a majority voting probably has the least requirements for preprocessing as the results are compiled from scores. The scope of fusion discussed in this paper is focused on multimodal biometric score fusion (MBSF). The source scores may originate from different types of devices, called *modality* (e.g., fingerprints, face images), and/or from variant analysis software, called *matcher* or *recognition* (e.g., linear discriminant analysis algorithm, Elastic Bunch Graphing Method (EBGM) algorithm) for face recognition. Score-level fusion usually involves score normalization, score fusion, and decision fusion. Score normalization (refer to Section 2.1) and decision fusion (refer to Section 3.2.2) may have some effects on the results of score fusion; however, the impacts of score modality (related with hardware), recognition method (software), and fusion process (post-processing in a hybrid human identification system) will be emphasized and investigated in this paper.

There are several types of score fusion methods: arithmetic combination of fusion scores, classifier-based fusion, and density-based fusion. In arithmetic fusion, the final score is a value of predefined function, f, with the input of normalized scores, $(s_1, s_2, \ldots)$. The output of such a fusion process, S_F, is computed by

$$S_F = f(s_1, s_2, \ldots, s_n), \tag{1}$$

where f stands for a fusion function or a set of fusion rules. f may be implemented by a simple arithmetic operation [15] such as taking the summation, average, product, minimum, maximum, median, majority vote, or by exploiting a Naive Bayes model [16]. In classifier-based fusion (referred to as *classifier fusion*), a classifier is first trained with the labeled score data, and then tested with unlabeled scores [4], [9]. The choices of classifiers include linear discriminant analysis (LDA) [8], k-nearest neighbors (KNN), artificial neural network (ANN) [14], and/or a support vector machine (SVM) [6]. In density-based fusion, a multi-dimensional density function is estimated with the score dataset, and then it can predict the probability of any given score vector [23], [28]. Nandakumar et al. [21] proposed a density-based fusion method where the likelihood ratio was estimated by Gaussian mixture model (GMM). Their experimental results [21] showed that the likelihood ratio fusion outperformed any single matcher and other fusion processes (like sum rule with min-max). A *Hidden Markov model* (HMM) was recently proposed for MBSF (referred to as *HMM fusion* [31]), which can

flexibly combine multiple scores from different modalities and/or from variant matchers. The early experimental results [31] showed that the HMM fusion was the most accurate and credible method in comparison to mean fusion and KNN fusion.

The security applications of a human identification system require achieving greater accuracy, efficiency, and credibility to robustly determine a person's identity (ID). It is clear that the MBSF process can significantly improve human identification performance [13], [21], [26], [27], [30], [31]. The set of literature focused on the advances of specific score fusion methods and its performance improvement typically evaluate the complete system. For example, Toh et al. [27] introduced a reduced multivariate polynomial model for multimodal biometric decision fusion (using three scores from fingerprint, speech and hand geometry), and they found that local learning and global decision did better than just fusing all three results at once. Ross and Jain [26] conducted a set of experiments in combining multimodal biometric scores (from face, fingerprint, and hand geometry), and their results indicated that the sum rule performed better than the decision tree and linear discriminant classifiers. Our early work [31] also focused on the discussion of performance improvement with the HMM fusion method. To the authors' knowledge, there are few published works that explore the key parameters that influence score fusion. Part of the reason may be lack of multimodal score databases and no effective metrics for fusion improvement evaluations across different methods and databases. Recognizing the key factor of score fusion will help design an accurate and credible human ID system to meet the critical needs of security applications. For instance, assuming that a human ID system permits a fusion with only two scores, should two modalities (one matcher per modality, e.g., fingerprint and face), or two matchers on one modality (e.g., fingerprint) be used? What is the impact of various fusion processes (e.g., HMM versus KNN) over different scenarios?

The main purpose of this research is to discover the key factor of multimodal biometric score fusion. Four fusion methods, mean, LDA, KNN, and HMM, are tested and compared using four biometric score datasets, wherein the HMM fusion is specifically configured for score fusion. Additionally, a new metric (called *relative rate increase*) is introduced for fusion improvement measurement. Our experiments reveal that score modality is the key factor in a score fusion scenario, which is meaningful to integrate and configure a multimodal biometric system. The rest of this paper is arranged as follows. The score normalization and fusion evaluation are depicted in Section 2. The score fusion processes including HMM fusion are described in Section 3. Experimental results, comparisons, and discussions are presented in Section 4. Finally, conclusions are drawn in Section 5.

2 SCORE NORMALIZATION AND FUSION EVALUATION

Multimodal biometric scores are computed with different modalities and algorithms, which may be similarity values (e.g., confidence values, probabilities, or logarithm probabilities), or distance measures (e.g., Euclidian distance, Hamming distance, or Mahalanobis distance). The variant source scores may contrast in a variety of ranges. Score normalization is required before score fusion. Meanwhile, fusion evaluation is needed to compare the performance of different fusion processes. To evaluate fusion performance, it is required that all original scores are either similarity scores or distance scores (but not the mix of similarity and distance). Converting a similarity score to a distance score is straightforward because of their *reciprocal* relationship.

2.1 Score Normalization

Prior to score fusion, score normalization is expected since the multimodal scores are heterogeneous and thus have varying dynamic ranges. The large variances of multimodal scores are caused either by different matching algorithms or by different natures of biometrical data. There are many normalization methods proposed in literature. Jain et al. [13] reported that min-max, z-score, and tanh normalization techniques, followed by a simple sum of scores fusion method, resulted in a superior GAR (genuine accept rate). It was also shown that both min-max and z-score methods are sensitive to outliers; whereas the tanh normalization method is both robust and efficient. The score data used in our experiments were obtained in the controlled lab environment (with less noise), thus a standard z-score *normalization* procedure is applied to all biometric scores,

$$\mathbf{S}_{\mathrm{N}} = (\mathbf{S}_0 - \boldsymbol{\mu}_0)/\boldsymbol{\sigma}_0, \tag{2}$$

where $\mathbf{S}_N$ is the normalized score vector, $\mathbf{S}_0$ is the original score vector, and $\boldsymbol{\mu}_0$ and $\boldsymbol{\sigma}_0$ denote the mean and standard deviation of original scores, respectively.

2.2 Fusion Evaluation

2.2.1 Verification Rate.

Genuine score is the matching score resulting from two samples of one user; while *impostor score* is the matching score of two samples originating from different users. *Genuine accept rate* (GAR) is the fraction of genuine scores exceeding the threshold; whereas *false accept rate* (FAR) is the fraction of impostor scores exceeding the threshold. A receiver operating characteristic (ROC) curve is computed from the FAR and true positive rate (TPR). On an open dataset (the query user may not be contained in the database), GAR/FAR/ROC area can be computed by choosing a threshold. On a closed dataset (the query user is surely included in the database), the identification performance can be measured by a *verification rate* (also called identification

Mt2 □ job no. 2175 □ jaif □ Journal of Advances in Information Fusion □ 2175d05 [2] □ (XXX) □ 01-06-15 □ 12:22 PM

rate or recognition rate), denoted as R_V, the percentage of correctly identified users over the total number of users. In our experiments, verification rate (VR) is used to evaluate the fusion performance since all users (i.e., subjects) are guaranteed in the database. Of course, the VR value may vary with a preset threshold. In a single-matcher evaluation, top-1 matching (e.g., the shortest distance) is used, while in a score fusion evaluation, the default threshold of each classifier is used. Finally, keep in mind that it is necessary to convert all multimodal scores to either similarity scores or distance scores before score fusion.

2.2.2 Relative Rate Increase.

The performance improvement using score fusion cannot be properly measured by using the absolute difference of two verification rates. For example, improving R_V from 80% to 90% seems to be more difficult than the improvement from 98% to 99%. Generally speaking, we know that the improvement of R_V via score fusion becomes more and more difficult when the original rate is approaching 100%. We propose to use a *relative rate increase* (denoted as RRI) to evaluate the fusion improvement.

$$\mathrm{RRI} = \frac{\mathrm{ARI}}{1-\overline{R_V}} = \frac{R_F - \overline{R_V}}{1-\overline{R_V}}, \tag{3}$$

where R_F is the verification rate via score fusion; $\overline{R_V}$ is the mean of original verification rates from individual modalities or matchers. $\mathrm{ARI} = R_F - \overline{R_V}$ is the *absolute rate increase* (ARI), which may not precisely measure the performance improvement as stated above. $\mathrm{RRI} \in (0,1]$; the higher, the better. According to the RRI definition, two fusion improvements, from 80% to 90% and from 98% to 99%, are equivalent, and their RRI = 0.50. It may be understood that the two improvements are "equivalent" in the sense of their difficulty levels and/or of the extent of their effort.

Many metrics could be devised, wherein the RRI metric seeks to measure the actual improvement against the total amount of possible improvement. With future large databases, the RRI metric would help in the quality of the fusion performance over the entire dataset (versus an assumed recognition performance with a small data set).

3 SCORE FUSION PROCESSES

In this section, arithmetic fusion and classifier fusion are briefly reviewed, and then HMM models are introduced for biometric score fusion.

3.1 Arithmetic Fusion and Classifier Fusion

Arithmetic fusion means to combine multiple scores by taking the summation, average (mean), product (called geometric mean), minimum, maximum, and median [15]. Majority vote is actually a kind of decision-level fusion, which requires the number of decision makers to be an odd number to avoid a possible tie. The mean fusion is selected in our experiments because it has the best performance of all aforementioned arithmetic fusion processes.

In *classifier fusion*, four frequently-used classification methods are discussed. These methods include linear discriminant analysis (LDA), k-nearest neighbor (KNN), artificial neural network (ANN), and support vector machine (SVM). The fusion results of LDA and KNN methods will be presented in our experiments due to their better performance on average [33], and thus these two methods are briefly described as follows, where the reader can find descriptions of ANN and SVM in the literature. The purpose of LDA is to predict group membership based on a linear combination of a set of predictor variables (i.e., a feature vector) [8]. The end result of the LDA procedure is a model (i.e., linear discriminant function, LDF) that allows prediction of group membership when only the predictor variables are known. The KNN method is usually deployed with a clustering technique. Fuzzy C-means (FCM) [3] is a data clustering technique wherein each data point belongs to a cluster to some degree that is specified by a membership grade. FCM starts with an initial guess of data membership and iteratively moves the cluster centers to the correct location within a data set. Once a certain number of clusters are formed by the FCM algorithm, the k-nearest neighbors can be found from those clusters using a Euclidean distance (between a testing feature vector and the clustered feature vectors). The probability of a given feature vector (multimodality scores) can be calculated with the labeled clusters.

To sufficiently use the sample data in classification evaluation, a *cross validation* method is applied to split original data into two groups for training and testing purposes. K-fold cross validation [25] is ideal for small databases. Notice that the divisions of k subsets ($k = 10$ used in our experiments) are based upon the users. If one user is grouped into Subset 1, then all scores of that user (including all his/her genuine and impostor scores) belong to Subset 1.

3.2 Hidden Markov Model for Multimodal Score Fusion

3.2.1 Basics on Hidden Markov Models.

In the past two decades, HMM models have emerged as a powerful tool for modeling stochastic processes and pattern sequences. Originally, HMMs have been applied to the domain of speech recognition and have become the dominating technology [24]. In recent years, they have attracted growing interest in computational molecular biology, bioinformatics, mine detection [12], handwritten character/word recognition [19], face and gesture recognition, shape recognition, image database retrieval, and other computer vision applications [5]. Generally speaking, an HMM is a model of a stochastic process that produces a sequence of random observa-

tion vectors at discrete times according to an underlying *Markov chain*. At each observation time, the Markov chain may be in one of N states $\{s_1,\ldots,s_N\}$ (hidden from the observation) and, given that the chain is in a certain state, there are probabilities of moving to other states, called the transition probabilities. An HMM is characterized by three sets of probability density functions: the *state transition probabilities* (**A**), the *observation symbol probabilities* (**B**), and the *initial state probabilities* ($\boldsymbol{\pi}$).

Let T be the length of the observation sequence (i.e., number of time steps; $t = 1,\ldots,T$), $\mathbf{O} = \{\mathbf{O}_1,\ldots,\mathbf{O}_T\}$ be the observation sequence, and $\mathbf{Q} = \{\mathbf{q}_1,\ldots,\mathbf{q}_T\}$ be the state sequence. The compact notation,

$$\lambda = (\mathbf{A},\mathbf{B},\boldsymbol{\pi}) \tag{4}$$

is generally used to indicate the complete parameter set of the HMM model, λ. In the above, $\mathbf{A} = \{a_{ij}\}$ is the state transition probability matrix, where $a_{ij} = P(q_t = s_j \mid q_{t-1} = s_i)$ for $i,j = 1,\ldots,N$; $\boldsymbol{\pi} = \{\pi_i\}$, where $\pi_i = P(q_1 = s_i)$, are the initial state probabilities. In the case of the *discrete HMM*, the observation vectors are commonly quantized into a finite set of symbols, $\{\mathbf{v}_1,\ldots,\mathbf{v}_M\}$ called the *codebook*. Each state is represented by a discrete probability density function and each symbol has a probability of occurring given that the system is in a given state. The observation symbol probability distribution $\mathbf{B} = \{b_i(\mathbf{O}_t)\}$ becomes a simple set of fixed probabilities for each class, i.e., $b_i(\mathbf{O}_t) = b_i(k) = P(\mathbf{v}_k \mid q_t = s_i)$, where $\mathbf{v}_k$ is the symbol of the nearest codebook of $\mathbf{O}_t$.

Three key problems [24] must be solved for the model defined in Eq. (4) to be useful in real world applications: the classification (*testing*) problem, the problem of finding an optimal state sequence (*tuning*), and the problem of estimating the model parameters (*training*). The classification problem involves computing the probability of an observation sequence $\mathbf{O} = \{\mathbf{O}_1,\ldots,\mathbf{O}_T\}$ given a model λ, i.e., $P(\mathbf{O} \mid \lambda)$. The Viterbi algorithm [20] is an efficient and formal technique for finding this maximum (optimal) state sequence and associate probability. The third problem is the training problem, i.e., how does one estimate the parameters of the model? First, all the states themselves must be estimated. Then the model parameters need to be estimated. In the discrete HMM, the codebook is first determined, usually using clustering techniques such as K-means [7] or fuzzy C-mean clustering algorithms [3]. The probability distribution **B** may be estimated either by fuzzy memberships [3] in a discrete HMM model, or by Gaussian mixture model (GMM) [10], [21] in a continuous HMM model. Then the parameters $(\mathbf{A},\mathbf{B},\boldsymbol{\pi})$ are estimated iteratively using the Baum-Welch algorithm [2].

3.2.2 HMMs for Multimodal Score Fusion.

The HMM fusion is a type of classifier fusion, but it significantly differs in data preparation and classification process. In the context of this paper, we need to distinguish two terms, multimodal scores and multi-matcher scores. *Multimodal* biometric scores (also referred to as *inter-modality* scores) result from different modalities (such as different hardware devices for imaging face and fingerprint); while *multi-matcher* scores (also referred to as *intra-modality* scores) result from different software algorithms but use the same modality (e.g., three face scores generated from three face recognition algorithms, respectively).

For HMM training, a large database with known users (labeled with user IDs) are expected, and thus a k-fold cross validation is utilized to satisfy this need. All scores are normalized and then organized as the inputs of HMM models using k-fold cross validation. The HMM model is adapted to a MBSF process and initialized with parameters like HMM(m,n,g), or denoted as $m \times n \times g$ HMM. Where m is the number of intra-modality scores (from m matchers upon one modality data) representing an *observation vector* in HMM, and n is the number of modalities corresponding to n hidden states, respectively. By placing n pieces of m-dimension observation vectors together, an observation sequence (over time, t) is formed. g is the number Gaussian components per state in a Gaussian mixture model (GMM). The GMM is applied to estimate the state probability density functions of each hidden state in a continuous HMM model.

Two HMM models are derived using genuine scores and impostor scores (in the training dataset), respectively. Given an observation sequence formed with multiple scores (of dimension $m \times n$) in the testing dataset, the two trained HMM models can compute the probabilities of being a genuine user and an impostor user, respectively. The user is identified as genuine if the probability given by the genuine HMM is higher. The details of the HMM model [31] and its adaption to a MBSF process are described as follows.

3.2.3 HMM Adaption to Multimodal Score Fusion.

The HMM models have wide applications in different fields and require proper data initialization for a specific application. In HMM score fusion, the observation vector $\mathbf{O}_t$ can be the m-dimensional intra-modality scores from m matchers. The observation sequence $\mathbf{O}(t,s)$ can be formed by combining n pieces of $\mathbf{O}_t$ from n modalities: $\mathbf{O}(t,s) = \{\mathbf{S}_{\mathrm{mn}}\}$. For example, there are 2 biometric modalities ($n = 2$; e.g., face, finger) and 2 matching algorithms (matchers) for each modality ($m = 2$). Thus, the length of $\mathbf{O}(t,s)$ is 4 (refer to NIST-Face-Fingerer database in Table 1a). The elements of **B** can be initialized with GMM, where the number of Gaussian models (g) in each state are usually fixed (e.g., $g = 3$) or automatically decided [10]. Notice that two HMM models, λ_{Gen} and λ_{Imp}, are actually trained using genuine scores and impostor scores, respectively; where their parameters can be estimated

Mt2 □ job no. 2175 □ jaif □ Journal of Advances in Information Fusion □ 2175d05 [4] □ (XXX) □ 01-06-15 □ 12:22 PM

TABLE 1a
Summary of four multimodal biometric databases.

Database	No. of Modalities	No. of Matchers	No. of Users	No. of Images	No. of Scores
NIST-Face-Finger	2: Finger & Face	4	517	3,102	1,069,156
NIST-Finger-Finger	1: Finger	2	6,000	24,000	72,000,000
NIST-Face	1: Face	2	3,000	6,000	36,000,000
ASUMS-Face-Face	2: Face (IR & DC)	6	96	576	110,592

TABLE 1b
Details of the trimmed databases (Sim. = Similarity; Dist. = Distance).

Database	Genuine : Impostor	HMM Models	Matchers	Score Type	Plot
NIST-Face-Finger_M1 (Face)	1,034 : 2,068	$2 \times 1 \times 3$	2 Face matchers	Sim., Sim.	Fig. 2a
NIST-Face-Finger_M2 (Finger)	1,034 : 2,068	$2 \times 1 \times 3$	Left, Right Finger	Sim., Sim.	Fig. 2b
NIST-Finger-Finger	12,000 : 24,000	$2 \times 1 \times 3$	Left, Right Finger	Sim., Sim.	Fig. 2c
NIST-Face	6,000 : 12,000	$2 \times 1 \times 5$	2 Face matchers	Sim., Sim.	Fig. 2d
ASUMS-Face-Face_M1 (IR)	576 : 1,152	$3 \times 1 \times 2$	FPB, LDA, EBGM	Dist., Dist., Sim.	Fig. 3a
ASUMS-Face-Face_M2 (DC)	576 : 1,152	$3 \times 1 \times 2$	FPB, LDA, EBGM	Dist., Dist., Sim.	Fig. 3b

Fig. 1. Sample faces from the ASUMS-Face-Face database: Notice that the two images (DC/visible, IR/thermal) shown at two neighboring columns were acquired from the same subject. The images are the aligned faces (320×320 pixels).

using the Baum-Welch algorithm [2]. An unlabeled biometric score sequence, $\mathbf{O}$, will be classified as a "genuine user" if $P_{\text{Gen}}(\mathbf{O} \mid \lambda_{\text{Gen}}) > P_{\text{Imp}}(\mathbf{O} \mid \lambda_{\text{Imp}}) + \eta$ (a simple decision rule); otherwise, $\mathbf{O}$ will be an "impostor user," where η is a small positive number empirically decided by experiments.

$$\mathbf{O} = \begin{cases} \text{Genuine User} & \text{if } P_{\text{Gen}}(\mathbf{O} \mid \lambda_{\text{Gen}}) > P_{\text{Imp}}(\mathbf{O} \mid \lambda_{\text{Imp}}) + \eta \\ \text{Impostor User} & \text{Otherwise} \end{cases} \tag{5}$$

In general, $m \geq 1$, $n \geq 1$, and $m \times n \geq 2$ are expected. In other words, at least two scores are required for HMM fusion. If the number of biometric modality is one ($n = 1$), then the number of matching scores from that modality must be two or greater (produced from different matching algorithms, e.g., LDA and EBGM [29] for face recognition). If there are two or more modalities ($n \geq 2$), in order to properly initialize and train the HMM models, the numbers of intra-modality scores ($m \geq 1$) derived from each modality must be same. There are usually more impostor scores than genuine scores in a biometric score dataset. To prevent a HMM model from being biased by the excessive impostor scores, the number of impostor scores used in training λ_{Imp} should be equivalent to the number of genuine scores used in training λ_{Gen}.

4 EXPERIMENTAL RESULTS AND DISCUSSIONS

The MBSF experiments were conducted on four biometric score databases and evaluated by reporting the verification rates (R_{V} and R_{F}) and the values of relative rate increase (RRI). Four fusion processes were selected to be reported in our experiments because of their better performance on average. The four fusion processes include one arithmetic fusion (mean fusion), two classifier fusions (LDA fusion and KNN fusion), and HMM fusion [31]. In the context, "modality" represents a biometric device (fingerprint, face); "matcher" is the software implementation of a "recognition method"; and "fusion method" means how to combine multiple scores (e.g., KNN fusion, HMM fusion). In the following discussion, Row 1 (or Column 1) referring to a table means the 1st row (or column) after the header row (or column).

4.1 Multimodal Scores and Experimental Design

Four biometric score databases (see Table 1a) were used in our experiments; three of which were from NIST-BSSR1 (Biometric Scores Set Release 1, from National Institute of Standards and Technology) [22], [31], and one of which was the face scores generated in our lab. Specifically, as shown in Table 1a, the *NIST-Face-Finger* database consists of a total of 1,069,156 biometric scores that were computed with 3,102 images from 517 users (individuals). Two face images, two left index fingerprints (images), and two right index fingerprints were acquired from each user; and then two face matching systems and one fingerprint matching system were applied to those images, respectively. So

TABLE 2a
The verification rates (%) of four fusion processes (R_F) across four databases.

Database	Single Matcher (R_V)	Mean Fusion	LDA Fusion	KNN Fusion	HMM Fusion (m,n,g)
NIST-Face	77.50, *81.02*	81.88	92.28	96.82	97.01 (2,1,5)
NIST-Finger-Finger	80.52, *87.88*	93.98	97.60	92.29	98.16 (2,1,3)
NIST-Face-Finger	89.17, 84.33 86.46, *92.65*	99.61	99.10	99.55	99.68 (2,2,5)
ASUMS-Face-Face	91.67, 93.75, 96.88 90.63, 93.75, *97.92*	100.00	99.48	98.48	99.83 (3,2,2)

TABLE 2b
The relative rate increase (RRI) of four fusion processes across four databases. $\overline{R_V}$ is the averaged R_V of all matchers.

Database	$\overline{R_V}$	Mean Fusion	LDA Fusion	KNN Fusion	HMM Fusion	(μ_{RRI},σ_{RRI})
NIST-Face	79.26	0.1263	0.6278	0.8467	0.8558	0.6142, 0.3419
NIST-Finger-Finger	84.20	0.6190	0.8481	0.5120	0.8835	0.7157, 0.1794
NIST-Face-Finger	88.15	0.9671	0.9240	0.9620	0.9730	0.9565, 0.0221
ASUMS-Face-Face	94.10	1.0000	0.9119	0.7424	0.9712	0.9064, 0.1153
(μ_{RRI},σ_{RRI})	(NA)	0.6781, 0.4062	0.8279, 0.1375	0.7658, 0.1915	0.9209, 0.0602	(NA)

there are 4 genuine scores for each user, two scores from two face matching systems, two scores from one fingerprint system but running on two fingerprints (left and right). There are two modalities (finger and face) and a total of four matchers in the NIST-Face-Finger database, and thus two $2 \times 2 \times 5$ HMM models [31] (for genuine and impostor, respectively) were initialized ($g = 5$ gave the best performance when varying g from 2 to 7[1]). The *NIST-Finger-Finger* database contains the scores from one fingerprint system running on two fingerprints (left and right); and then two $2 \times 1 \times 3$ HMM models were established. The *NIST-Face* database is comprised of the scores from two face matching systems; and two $2 \times 1 \times 5$ HMM models were created.

The *ASUMS-Face-Face* (Alcorn State University [ASU] MultiSpectral) database (Row 4 in Table 1a) includes the scores from three face recognition algorithms and from two modalities ASUIR (ASU long-wave Infrared) face images and ASUDC (ASU Digital Camera) face images (see Fig. 1). Three face recognition algorithms are linear discriminant analysis (LDA) [18], elastic bunch graph matching (EBGM) [29], and face pattern byte (FPB) [32]. The corresponding HMM models were configured as $3 \times 2 \times g$ (refer to Table 2a). The *ASUIR-Face* subset [32] includes thermal (long-wave infrared, IR) face images, whereas the *ASUDC-Face* subset consists of visible (digital camera, DC) face images from the same group of users. In these two sub-datasets, 3 face images were acquired from each user, where one randomly-selected image was used as *probe face* (i.e., a face image from a live camera) and the other two as *gallery faces* (i.e., face images from a database). Table 1b shows the comparative relations over the trimmed datasets (of reduced impostor scores) between numbers of genuine to impostor scores, parameters of HMM models ($m \times n \times g$), matchers, score type, and plots (also refer to Tables 2a, 3a, 4a).

The total number of scores is massive in that it mainly contains impostor scores. For example, ASUMS-Face-Face consists of 1,152 genuine scores and 109,440 impostor scores (for all 3 matchers and 2 modalities). All scores are normalized by using Eq. (2). An unbalanced training with the excessive impostor scores may result a biased or over-trained model. To avoid possible bias in model training as mentioned in Section 3, two impostor scores per matcher per user were randomly selected for training. All genuine scores plus reduced impostor scores are called "trimmed database." Arithmetic fusion used all scores (full databases), whereas HMM fusion and classifier fusion used *trimmed databases* (refer to Table 1b). The distributions of normalized scores of four trimmed databases are presented as scatter plots in Figs. 2–3, where the x-axis denotes Score 1 and the y-axis represents other scores. The distributions of 3 scores shown in Fig. 3 indicate low correlation of three scores. Figs. 2–3 also show that three NIST databases contain similarity scores (genuine scores are large); while the ASUMS database includes both similarity scores and distance scores (genuine scores are small). Notice that less impostor markers shown in Fig. 2c is because most impostor markers are behind (thus blocked by) the genuine markers.

Four fusion processes were tested across the four score databases (refer to Table 1a). The fusion results of mean fusion, LDA (with quadratic kernel), KNN, and HMM were reported in Table 2a. The HMM models were implemented and adapted upon the "Hidden Markov Model (HMM) Toolbox for Matlab" [20]. All HMM models were tested by varying the number of Gaussian components (g) from 2 to 7, the best results of HMM fusions (together with initialization parameters) are shown in Table 2a. The verification rates of original scores are presented in Column "Single Matcher" in

[1] g was determined empirically in the experiment from which the differences of g had a marginal impact on the results.

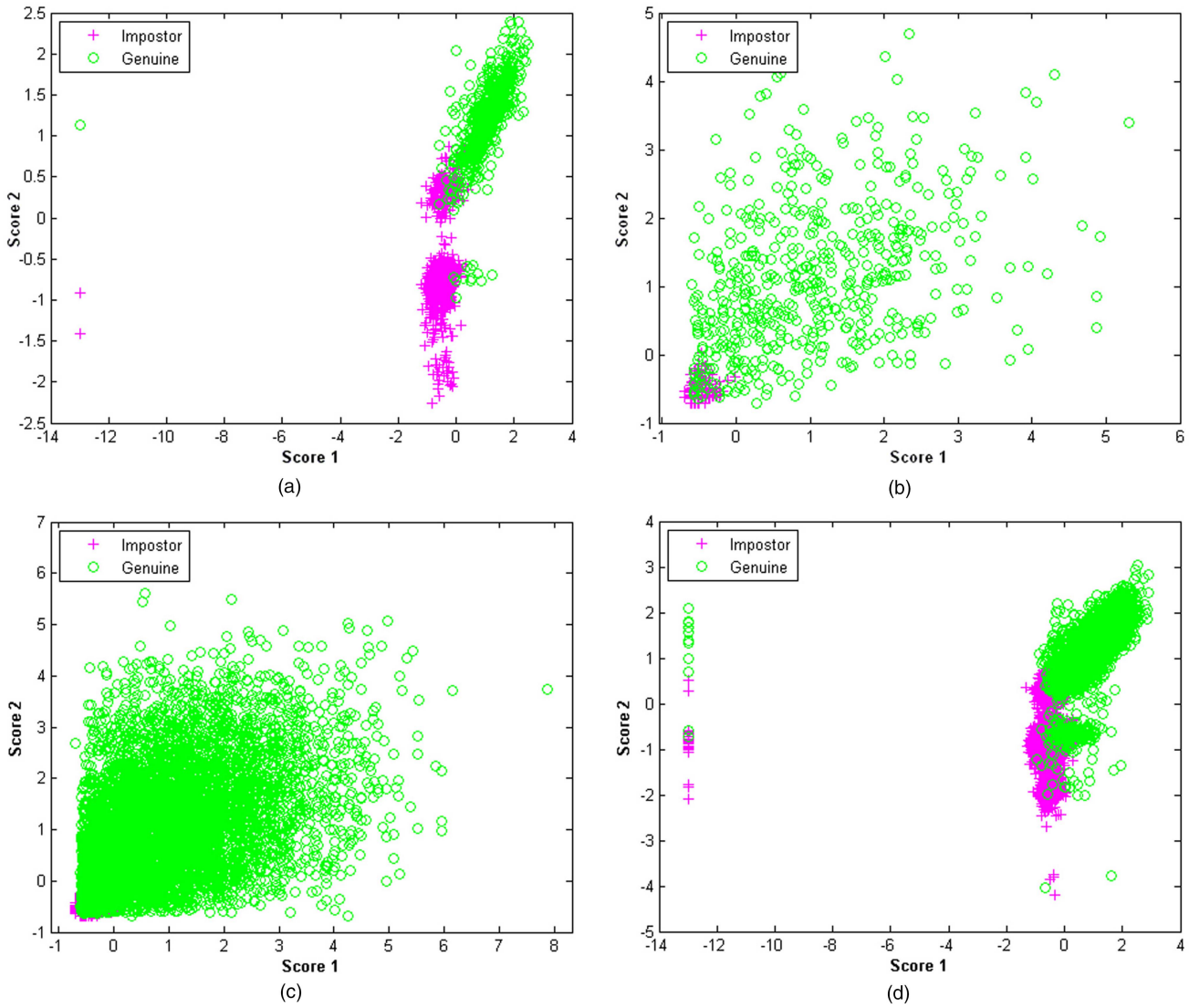

Fig. 2. Scatter plots (score distributions) of normalized multimodal biometric scores from three NIST databases (trimmed), where the x-axis is Score 1 and the y-axis is Score 2: (a) NIST-Face-Finger_M1 (Face); (b) NIST-Face-Finger_M2 (Finger); (c) NIST-Finger-Finger; (d) NIST-Face. Notice that all genuine scores and the two randomly-selected impostor scores per matcher per user are presented.

Table 2a, where the rightmost number (in *italic* font) is the single best performance.

4.2 Results and Discussions

The performance of individual matcher (R_V) and four fusion processes (R_F) on four databases are presented in Table 2a. It is clear that all four fusion approaches yield improvements compared to the corresponding single best matcher (SBM) on each database. Overall, the HMM fusion performs the best. It seems that the *mean fusion* performs very well on the multimodal databases (99.61% on NIST-Face-Finger and 100% on ASUMS-Face-Face). The possible reason might be that the genuine scores and the impostor scores on these two databases are well separated (refer to the score distributions shown in Figs. 2–3), which makes a linear separation (like mean fusion) ideal. Surprisingly, in another independent research [30], the weighted-sum score fusion reached the highest rate of 99% (SBM = 97%) when two weights were equal, which turned out to be a mean fusion (but the score distributions were not presented). The level of improvement will be analyzed using the values of *relative rate increase* (RRI).

The RRI values of four fusion processes are given in Table 2b. Table 2a and Table 2b are corresponding cell-by-cell except for the last row and the last column. Let us examine the rationality of RRI, which is proposed to measure the improvement of score fusion. The RRI value of mean fusion on NIST-Face is 0.1263 (the smallest value in Table 2b), which corresponds an absolute rate increase (ARI = 2.62%) from 79.26% ($\overline{R_V}$) to 81.88%. The RRI value of HMM fusion on NIST-Face-Finger is 0.9730 (the second largest value in Table 2b), which corresponds ARI = 11.53% (from 88.15% to 99.68%). There is a special case, RRI = 1.0000, for the mean fusion on ASUMS-Face-Face, which represents a

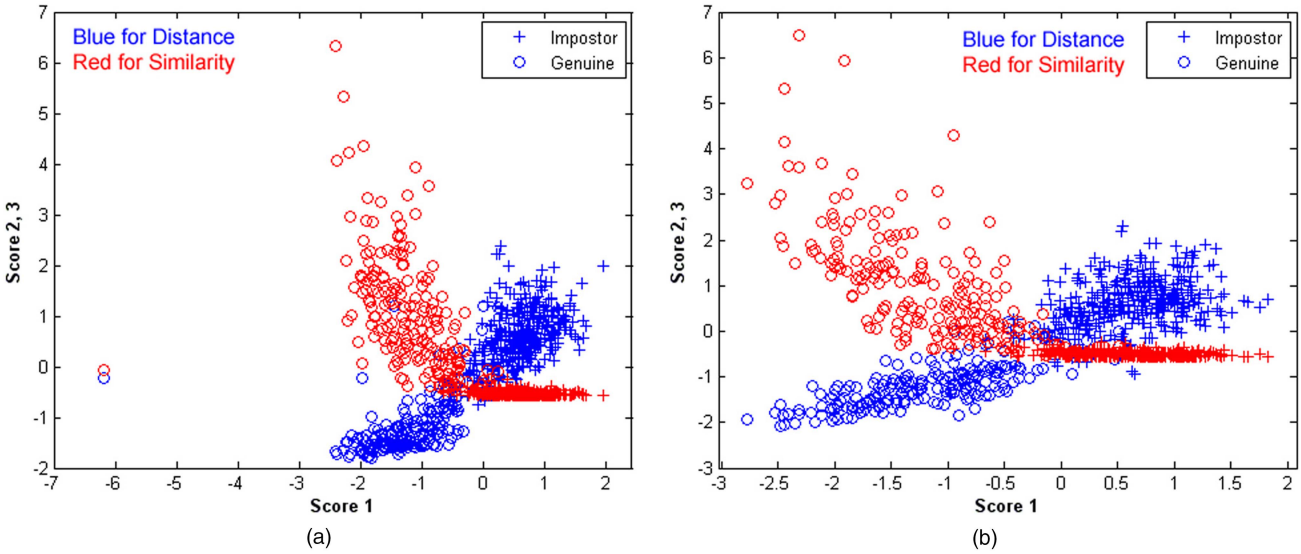

Fig. 3. Scatter plots (score distributions) of normalized face scores from two ASUMS datasets (trimmed), where the x-axis shows Score 1 from FPB (distance score), and the y-axis represents Score 2 from LDA (distance score shown in blue) and Score 3 from EBGM (similarity score shown in red): (a) Scores from ASUMS-Face-Face_M1 (IR); (b) Scores from ASUMS-Face-Face_M2 (DC). Notice that all genuine scores and the two randomly-selected impostor scores per matcher per subject are presented.

verification rate increase from 94.10% to 100%. Since $R_F = 100\%$ means a perfect fusion (i.e., a perfect human identification system), it is reasonable for RRI to take its maximum value, 1.0. On a large database (e.g., millions of users), RRI rarely reaches 1.0. According to the definition of RRI in Eq. (3), improving R_F from 90% to 100% (ARI = 10%) and from 99.9% to 100% (ARI = 0.1%), both will have RRI = 1.0, which makes sense in terms of difficulty or effort. In other words, the level of difficulty or the amount of effort for both cases may be equivalent.

In Table 2b, the means and standard deviations of RRI, denoted as $(\mu_{RRI}, \sigma_{RRI})$, in each row and in each column are listed in the last column and in the last row, respectively, where "NA" means not applicable. As shown in the bottom row of Table 2b, when averaging across four biometric databases, the HMM fusion has the highest μ_{RRI} and also the least σ_{RRI}. We may conclude that the *HMM fusion* is the best for MBSF in terms of accuracy (high improvement) and credibility (low variance). The LDA fusion is the second best. According to the rightmost column in Table 2b, when averaging across four fusion processes, the NIST-Face-Finger database gives the highest μ_{RRI} (0.9565) with the least σ_{RRI}. The ASUMS-Face-Face database is the second best ($\mu_{RRI} = 0.9064$). It is clear that *multimodal fusion* (NIST-Face-Finger and ASUMS-Face-Face, their averaged $\mu_{RRI} = 0.9314$) is superior to single-modal fusion (NIST-Finger-Finger and NIST-Face, their averaged $\mu_{RRI} = 0.6649$). It also makes sense that NIST-Face-Finger produces a better improvement than ASUMS-Face-Face since NIST-Face-Finger consists of truly diverse modalities (face and finger), whereas ASUMS-Face-Face contains two bands of face images (thermal and visible).

In Table 2b, μ_{RRI}(NIST-Face-Finger) = 0.9565 represents a *modality fusion* with multimodal scores when averaging fusions; μ_{RRI}(HMM_Fusion) = 0.9209 is from the best *fusion process* when mixing modalities and recognitions; and μ_{RRI}(NIST-Face) = 0.6142 is considered as a *recognition fusion* result with single-modal (face) scores when averaging fusions. It reveals that the importance of fusion factors from the highest to the lowest are as follows: score modality, fusion process, and recognition method. These three factors may interact with one another; however, we do not have sufficient data (power) to conduct an analysis of variance (ANOVA).

To investigate and verify key factor that influences the score fusion (i.e., sensitivity test), we need to separate three fusion factors: modality, recognition, and fusion. Thus two multimodal databases, NIST-Face-Finger and ASUMS-Face-Face, are selected, and divided into modality subsets (e.g., NIST-Face-Finger_M1 and NIST-Face-Finger_M2; refer to Table 1b) and recognition subsets (e.g., NIST-Face-Finger_R1 and NIST-Face-Finger_R2). For example, on NIST-Face-Finger_M1 (face), the mean fusion is performed by averaging the scores from two matchers (i.e., two recognition methods), which is used to research the impact of the recognition method. While on ASUMS-Face-Face_R2 (EBGM), the mean fusion is achieved by averaging the two EBGM scores from two modalities (i.e., IR and DC; see Fig. 1) and used to study the impact of score modality. The performance of individual matcher (R_V) and four fusion processes (R_F) on subsets are listed in Table 3a and Table 4a, and the relative rate increase (RRI) of four fusion processes on subsets and their $(\mu_{RRI}, \sigma_{RRI})$ are given in Table 3b and Table 4b. The following discussions are based on the results of RRI

Mt2 □ job no. 2175 □ jaif □ Journal of Advances in Information Fusion □ 2175d05 [8] □ (XXX) □ 01-06-15 □ 12:22 PM

TABLE 3a
The verification rates (%) of four fusion processes (R_F) on four subsets derived from the NIST-Face-Finger database.

Database	Single Matcher (R_V)	Mean Fusion	LDA Fusion	KNN Fusion	HMM Fusion (m,n,g)
NIST-Face-Finger_M1 (Face)	84.33, *89.17*	90.52	94.45	96.97	97.42 (2,1,3)
NIST-Face-Finger_M2 (Finger)	86.46, *92.65*	94.78	97.23	97.48	98.06 (2,1,3)
NIST-Face-Finger_R1 (Matcher 1)	84.33, 86.46	94.78	97.16	99.22	99.22 (1,2,2)
NIST-Face-Finger_R2 (Matcher 2)	89.17, 92.65	96.71	99.16	99.42	99.42 (1,2,3)

TABLE 3b
The relative rate increase (RRI) of four fusion processes on four subsets derived from the NIST-Face-Finger database.

Database	$\overline{R_V}$	Mean Fusion	LDA Fusion	KNN Fusion	HMM Fusion	(μ_{RRI},σ_{RRI})
NIST-Face-Finger_M1 (Finger)	86.75	0.2845	0.5811	0.7713	0.8053	0.6106, 0.2387
NIST-Face-Finger_M2 (Face)	89.56	0.5002	0.7348	0.7587	0.8143	0.7020, 0.1386
NIST-Face-Finger_R1 (Matcher 1)	85.39	0.6426	0.8055	0.9466	0.9466	0.8353, 0.1447
NIST-Face-Finger_R2 (Matcher 2)	90.91	0.6381	0.9076	0.9362	0.9362	0.8545, 0.1449
(μ_{RRI},σ_{RRI})	(NA)	0.5164, 0.1681	0.7573, 0.1372	0.8532, 0.1020	0.8756, 0.0762	(NA)

or μ_{RRI} since they can more properly evaluate the improvement than R_F.

As shown in Table 3b, μ_{RRI}(NIST-Face-Finger_R1) = 0.8353 and μ_{RRI}(NIST-Face-Finger_R2) = 0.8545 are significantly higher than μ_{RRI}(NIST-Face-Finger_M1) = 0.6106 and μ_{RRI}(NIST-Face-Finger_M2) = 0.7020, respectively. Further averaging the RRI values of two recognition subsets (Rows 3–4), we have μ_{RRI}(NIST-Face-Finger_Rn) = 0.8449, which is much higher than μ_{RRI}(NIST-Face-Finger_Mn) = 0.6563, where $n = 1,2$. The comparisons above indicate that *score modality* is a more important factor than *recognition method* (matcher) when averaging (or mixing) fusion processes. This statement complies with the conclusion from Table 2b, *multimodal fusion* is superior to single-modal fusion. Table 4b shows the same fact, where μ_{RRI}(ASUMS-Face-Face_Rn) = 0.6814 ($n = 1,2,3$) and μ_{RRI}(ASUMS-Face-Face_Mn) = 0.6574 ($n = 1,2$), although the difference is small as expected (due to less diversity in score modalities).

To further explore the impacts of fusion factors within one score database, the (μ_{RRI},σ_{RRI}) values of the combined modality subsets (Rows 1–2 in Table 3b and Rows 1–2 in Table 4b) and the (μ_{RRI},σ_{RRI}) values of the combined recognition subsets (Rows 3–4 in Table 3b and Rows 3–5 in Table 4b) are shown in Table 5. As seen before, the fusion of modalities is superior to the fusion of recognitions (matchers) with one exception (the LDA fusion on ASUMS-Face-Face database). The mean fusion results are not used in the following discussions due to their high variances (i.e., low credibility).

We shall make quantitative comparisons on NIST-Face-Finger database (Rows 1–2 in Table 5). When the fusion process is selected (fixed), for instance, with HMM fusion, the difference of μ_{RRI} values between the fusion of modalities and the fusion of recognitions is 0.1316, denoted as $\Delta\mu_{RRI}$(Modality, Recognition | HMM) = 0.9414 − 0.8098 = 0.1316. This big difference shows the fusion of modalities is much better than the fusion of recognitions. When the modalities are selected and the matchers (i.e., recognitions) are fixed (refer to Row 2 in Table 5), no big difference between different fusion processes is observed, for example, $\Delta\mu_{RRI}$(HMM, KNN | Recognition) = 0, and $\Delta\mu_{RRI}$(HMM, LDA | Recognition) = 0.0848. These comparisons show that the fusion of different modalities is a *dominant* factor, which makes the different fusion processes have less impact on fusion improvement. When the matchers are chosen and the modality is fixed (refer to Row 1 in Table 5), we have μ_{RRI}(HMM, KNN | Modality) = 0.0448, and μ_{RRI}(HMM, LDA | Modality) = 0.1518. These results show that the *fusion process* plays an important role when fusing multi-matcher scores from a single modality (i.e., without the dominant factor of modality). Note that in Table 5 the two identical entries at Row 2, Column 3 and 4 are just coincident.

Similar quantitative analyses on ASUMS-Face-Face database (Rows 3–4 in Table 5) are given as follows. $\Delta\mu_{RRI}$(Modality, Recognition | HMM) = 0.7231 − 0.6788 = 0.0443 reveals that the fusion of different modalities (thermal face and visible face) is slightly better than the fusion of recognitions but no longer a dominant factor. $\Delta\mu_{RRI}$(HMM, KNN | Recognition) = 0.1135 and μ_{RRI}(HMM, KNN | Modality) = 0.0876 indicate that the different fusion processes become a more important factor when the modality is not a dominant factor.

How to apply these findings to guide a MBSF development and application is discussed below. *Modality* is the key and dominant factor in score fusion, but adding more matcher scores to the fusion will improve the performance further. In fact, R_F(NIST-Face-Finger, HMM_Fusion) = 99.68% (4-score fusion shown in Table 2a) is higher than R_F(NIST-Face-Finger_R2, HMM_Fusion) =

TABLE 4a
The verification rates (%) of four fusion processes (R_F) on five subsets derived from the ASUMS-Face-Face database.

Database	Single Matcher (R_V)	Mean Fusion	LDA Fusion	KNN Fusion	HMM Fusion (m,n,g)
ASUMS-Face-Face_M1 (IR)	91.67, 93.75, *96.88*	96.88	98.97	97.61	98.45 (3,1,2)
ASUMS-Face-Face_M2 (DC)	90.63, 93.75, *97.92*	98.96	97.94	97.26	97.76 (3,1,2)
ASUMS-Face-Face_R1 (LDA)	91.67, 90.63	96.88	96.39	96.90	97.95 (1,2,3)
ASUMS-Face-Face_R2 (EBGM)	93.75, 93.75	97.92	98.28	99.49	99.14 (1,2,2)
ASUMS-Face-Face_R3 (FPB)	96.88, 97.92	100	98.62	98.28	98.80 (1,2,2)

TABLE 4b
The relative rate increase (RRI) of four fusion processes on five subsets derived from the ASUMS-Face-Face database.

Database	$\overline{R_V}$	Mean Fusion	LDA Fusion	KNN Fusion	HMM Fusion	(μ_{RRI},σ_{RRI})
ASUMS-Face-Face_M1 (IR)	94.10	0.4712	0.8254	0.5949	0.7373	0.6572, 0.1562
ASUMS-Face-Face_M2 (DC)	94.10	0.8237	0.6508	0.5356	0.6203	0.6576, 0.1210
ASUMS-Face-Face_R1 (LDA)	91.15	0.6475	0.5921	0.6497	0.7684	0.6644, 0.0743
ASUMS-Face-Face_R2 (EBGM)	93.75	0.6672	0.7248	0.9184	0.8624	0.7932, 0.1169
ASUMS-Face-Face_R3 (FPB)	97.40	1.0000	0.4692	0.3385	0.5385	0.5865, 0.2878
(μ_{RRI},σ_{RRI})	(NA)	0.7219, 0.1994	0.6525, 0.1345	0.6074, 0.2099	0.7054, 0.1272	(NA)

TABLE 5
The (μ_{RRI},σ_{RRI}) values of the combined modality subsets (Rows 1–2 in Table 3b and in Table 4b, respectively) and the (μ_{RRI},σ_{RRI}) values of the combined recognition subsets (the rest rows in Table 3b and in Table 4b, respectively).

Database	Mean Fusion	LDA Fusion	KNN Fusion	HMM Fusion	Fusion of What
NIST-Face-Finger_M1-M2	0.3924, 0.1525	0.6580, 0.1087	0.7650, 0.0089	0.8098, 0.0064	Recognitions/Matchers
NIST-Face-Finger_R1-R2	0.6403, 0.0032	0.8566, 0.0722	0.9414, 0.0074	0.9414, 0.0074	Modalities
ASUMS-Face-Face_M1-M2	0.6475, 0.2493	0.7381, 0.1234	0.5653, 0.0419	0.6788, 0.0827	Recognitions/Matchers
ASUMS-Face-Face_R1-R3	0.7716, 0.1981	0.5954, 0.1278	0.6355, 0.2902	0.7231, 0.1667	Modalities

99.42% (2-score fusion shown in Table 3a). The fusion process becomes very important when the score modalities are fixed, for instance, the fusion of multiple matchers of single modality. For example, imagine a human identification system of two modalities (face and finger) and of two matchers per modality that has $R_F = 99.68\%$ using HMM fusion, how can you further improve the system performance? According to the findings of this research, the recommended solution is first to add one more modality (e.g., voice or iris), then to develop a better fusion process than HMM, and/or to add more recognition methods (like Local Gabor Binary Patterns (LGBP) [30] for face recognition). Of course, using a high-performance matcher is always preferred. The implication hereby is that developing a better fusion process (e.g., better than HMM) will have a higher impact on fusion improvement (i.e., a larger RRI) than adding a third matcher into each modality.

A recent face recognition research [34] explored the performance improvement with the stereo fusion at three levels: image, feature, and score. The primary fusions investigated in that paper are *stereo fusion* with the stereo images captured from two identical cameras. Experimental results show that any level stereo fusion can improve the recognition performance. It seems that stereo image fusion and stereo feature fusion is better than stereo score fusion. However, the processes for the fusions at image level and feature level are more complicated (such as image registration). On the other hand, score fusion can be implemented without the knowledge of what images and what features, and can be performed flexibly by using variant score combinations from different cameras, modalities, and/or matchers. In addition, score fusion is faster than image fusion or feature fusion.

In the future we will sufficiently investigate and verify the current findings by developing more recognition methods and more fusion processes and by using more biometric modalities (like voice, iris, and palm geometry). A statistical analysis (e.g., ANOVA, ROCs [1]) will be conducted to study the interactions and significance of those fusion factors. We will also research the impacts of normalization procedures, decision rules, and image fusion techniques [17] on the MBSF process.

5 CONCLUSIONS

A set of experiments regarding multimodal biometric score fusion (MBSF) has been conducted in this research. A hidden Markov model (HMM) is tested for multimodal biometrics score fusion, which is the most accurate, reliable, and credible fusion process compared

Mt2 □ job no. 2175 □ jaif □ Journal of Advances in Information Fusion □ 2175d05 [10] □ (XXX) □ 01-06-15 □ 12:22 PM

to other three methods (mean, LDA, KNN). To evaluate and compare the improvement of variant fusion processes, a new metric, called *relative rate increase* (RRI), is proposed upon the concept of verification rate. The RRI metric has proved to be reasonably accurate in measuring the performance improvement resulting from MBSF. Based on the experimental results from four multimodal biometric databases, the findings can be summarized as follows. The *score modality* is the most important (key) factor in biometric score fusion which dominates the fusion result. When the number of score modalities is fixed, the fusion process becomes the next important factor to score fusion. Adding more recognition matchers has the least impact on fusion improvement. Another finding is that, different bands of face images (thermal and visible) are less diverse modalities than face and finger, which makes the score modality (of thermal faces and visible faces) no longer a dominant factor.

ACKNOWLEDGMENT

This research was supported by the Department of Defense Research and Education Program (under Grant No. W911NF-12-1-0065) administered by the Army Research Office and the Office of Naval Research.

REFERENCES

[1] S. Alsing, E. P. Blasch, and R. Bauer
Three-Dimensional Receiver Operating Characteristic (ROC) Trajectory Concepts for the Evaluation of Target Recognition Algorithms Faced with the Unknown Target Detection Problem,
Proc. SPIE. 3718, 1999.

[2] L. E. Baum, T. Petrie
Statistical Inference for Probability Functions of Finite State Markov Chains,
Ann. Math. Stat., 37:1554–1563, 1966.

[3] J. C. Bezdec
Pattern Recognition with Fuzzy Objective Function Algorithms,
Plenum Press, New York, 1981.

[4] R. Brunelli, D. Falavigna
Person Identification Using Multiple Cues,
IEEE Trans. Pattern Anal. Mach. Intell., 17(10): 955–966, 1995.

[5] H. Bunke, T. Caelli
Hidden Markov Models: Applications in Computer Vision,
World Scientific, River Edge, New Jersey, 2001.

[6] C. Burges
A Tutorial on Support Vector Machines for Pattern Recognition,
Data Mining and Knowledge Discovery, 2:121–167, 1998.

[7] H. P. Chan, K. Doi, S. Galhotra, C. J. Vyborny, H. MacMahon, P. M. Jokich
Image Feature Analysis and Computer-aided Diagnosis in Digital Radiography. I. Automated Detection of Microcalcifications in Mammography,
Med. Physics, 14:538–548, 1987.

[8] R. O. Duda, P. E. Hart
Pattern Classification and Scene Analysis,
Wiley, New York, 1973.

[9] J. Fierrez-Aguilar, J. Ortega-Garcia, J. Gonzalez-Rodriguez, J. Bigun
Discriminative Multimodal Biometric Authentication based on Quality Measures,
Pattern Recognition, 38(5):777–779, 2005.

[10] M. Figueiredo, A. K. Jain
Unsupervised Learning of Finite Mixture Models,
IEEE Trans. Pattern Anal. Mach. Intell. 24 (3), 381–396, 2002.

[11] G. D. Forney
The Viterbi Algorithm,
Proc. IEEE 61, 268–278, 1973.

[12] P. Gader, M. Mystkowski, Y. Zhao
Landmine Detection with Ground Penetrating Radar using Hidden Markov Models,
IEEE Trans. Geosci. Remote Sens., 39:1231–1244, 2001.

[13] A. Jain, K. Nandakumar, A. Ross
Score Normalization in Multimodal Biometric Systems,
Pattern Recognition, 38(12):2270–2285, 2005.

[14] R. M. Kil, I. Koo
Optimization of a Network with Gaussian Kernel Functions Based on the Estimation of Error Confidence Intervals,
Proc. of IJCNN 2001, 3:1762–176, 2001.

[15] L. I. Kuncheva
A Theoretical Study on Six Classifier Fusion Strategies,
IEEE Trans. Pattern Anal. Mach. Intell., 24(2):281–286, 2002.

[16] L. I. Kuncheva
Switching between Selection and Fusion in Combining Classifiers: an Experiment,
IEEE Trans. on Systems, Man, and Cybernetics, Part B, 32(2):146–156, 2002.

[17] Z. Liu, E. Blasch, Z. Xue, R. Langaniere, and W. Wu
Objective Assessment of Multiresolution Image Fusion Algorithms for Context Enhancement in Night Vision: A Comparative Survey,
IEEE Trans. Pattern Anal. Mach. Intell., 34(1):94–109, 2012.

[18] J. Lu, K. N. Plataniotis, A. N. Venetsanopoulos
Face Recognition Using LDA-Based Algorithms,
IEEE Trans. on Neural Networks, 14(1):195–200, 2003.

[19] M. Mohamed, P. Gader
Generalized Hidden Markov Models part 2: Applications to Handwritten Word Recognition,
IEEE Trans. Fuzzy Systems, 8:186–194, 2000.

[20] K. Murphy
Machine Learning: a Probabilistic Perspective,
MIT Press, Cambridge, Massachusetts, 2012.

[21] K. Nandakumar, Y. Chen, S. C. Dass, A. K. Jain
Likelihood ratio-based biometric score fusion,
IEEE Trans. Pattern Anal. Mach. Intell., 30(2): 342–347, 2008.

[22] National Institute of Standards and Technology
NIST Biometric Scores Set—release 1,
http://www.itl.nist.gov/iad/894.03/biometricscores, 2004.

[23] S. Prabhakar, A. K. Jain
Decision-level Fusion in Fingerprint Verification,
Pattern Recognition, 35(4):861–874, 2002.

[24] L. Rabiner
Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition,
Proc. IEEE 77, 257–286, 1989.

[25] S. Raudys
Statistical and Neural Classifiers: an Integrated Approach to Design,
Springer, London, 2001.

[26] A. Ross and A. Jain
Information Fusion in Biometrics,
Pattern Recognition Letters, 24:2115–2125, 2003.

[27] K.-A. Toh, X. Jiang, and W.-Y. Yau
Exploiting Global and Local Decisions for Multimodal Biometrics Verification,
IEEE Trans. on Signal Processing, 52(10):3059–3072, 2004.

[28] B. Ulery, A. R. Hicklin, C. Watson, W. Fellner, P. Hallinan
Studies of Biometric Fusion,
NIST Interagency Report, 2006.

[29] L. Wiskott, J. M. Fellous, N. Krüger, C. von der Malsburg
Face Recognition by Elastic Bunch Graph Matching,
IEEE Trans. Pattern Anal. Mach. Intell., 19 (7):775–779, 1997.

[30] S. Xie, S. Shan, X. Chen, and J. Chen
Fusing Local Patterns of Gabor Magnitude and Phase for Face Recognition,
IEEE Trans. Image Process., 19(5):1349–1361, 2010.

[31] Y. Zheng
A Hidden Markov Model for Multimodal Biometrics Score Fusion,
Proc. SPIE 8064, 80640D, 2011.

[32] Y. Zheng
A Novel Orientation Code for Face Recognition,
Proc. SPIE 8056, 805606, 2011.

[33] Y. Zheng and E. Blasch
Score Fusion and Decision Fusion for the Performance Improvement of Face Recognition,
Int'l Conf. on Info Fusion, 2013.

[34] Y. Zheng and E. Blasch
The Advantages of Stereo Vision in a Face Recognition System,
Proc. SPIE 9091, 2014.

Yufeng Zheng received his Ph.D. degree in Optical Engineering/Image Processing from the Tianjin University (Tianjin, China) in 1997. He is presently with the Alcorn State University (Mississippi, USA) as an associate professor. Dr. Zheng serves as a program director of the Computer Networking and Information Technology Program, and a director of the Pattern Recognition and Image Analysis Lab. He is the principle investigator of three federal research grants in night vision enhancement, and in multispectral face recognition. So far Dr. Zheng holds two patents in glaucoma classification and face recognition, and has published one book, six book chapters, and more than 70 peer-reviewed papers. His research interests include pattern recognition, biologically inspired image analysis, biometrics, information fusion, and computer-aided diagnosis. Dr. Zheng is a Cisco Certified Network Professional (CCNP), a senior member of SPIE, a member of IEEE, Computer Society & Signal Processing, as well as a technical reviewer.

Erik Blasch received his B.S. in mechanical engineering from the Massachusetts Institute of Technology in 1992 and M.S. degrees in mechanical engineering ('94), health science ('95), and industrial engineering (human factors) ('95) from Georgia Tech and attended the University of Wisconsin for a M.D./Ph.D. in mechanical engineering/neurosciences until being called to active duty in 1996 to the United States Air Force. He completed an M.B.A. ('98), M.S.E.E. ('98), M.S. econ ('99), M.S./Ph.D. psychology (ABD), and a Ph.D. in electrical engineering from Wright State University and is a graduate of Air War College. From 2000–2010, Dr. Blasch was the information fusion evaluation tech lead for the Air Force Research Laboratory (AFRL) Sensors Directorate—COMprehensive Performance Assessment of Sensor Exploitation (COMPASE) Center, adjunct professor with Wright State University, and a reserve officer with Air Office of Scientific Research. From 2010–2012. Dr. Blasch was an exchange scientist to Defence R&D Canada at Valcartier, Quebec in the Future Command and Control (C2) Concepts group. He is currently with the AFRL Information Directorate. He compiled over 30 top ten finishes as part of robotic teams in international contests, received the 2009 IEEE Russ Bioengineering Award, and the 2014 *Joseph Mignogna Data Fusion Award* from the U.S. Department of Defense Joint Directors of Laboratories Data Fusion Group. He is a past President of the International Society of Information Fusion (ISIF), a member of the IEEE Aerospace and Electronics Systems Society (AESS) Board of Governors, AIAA Associate Fellow, and a SPIE Fellow. His research interests include target tracking, information/sensor/image fusion, pattern recognition, and biologically-inspired applications.

Mt2 ☐ job no. 2175 ☐ jaif ☐ Journal of Advances in Information Fusion ☐ 2175d05 [12] ☐ (XXX) ☐ 01-06-15 ☐ 12:22 PM