Score Fusion and Decision Fusion for the Performance Improvement of Face Recognition

Yufeng Zheng
Alcorn State University
Lorman, MS, USA
yufeng.zheng@r2image.com

Erik Blasch
US Air Force Research Laboratory
Rome, NY, USA
erik.blasch@rl.af mil

Abstract — *To improve the performance of a face recognition system, we propose a fusion solution consisting of score fusion of multispectral images and decision fusion of stereo images. Score fusion combines several scores from multiple matchers and/or multiple modalities, which can increase the accuracy of face recognition. In a face recognition system, low false accept rate (FAR) is as important as high accuracy rate. The FAR can be reduced by using decision fusion of stereo images. The stereo face images are taken with two identical cameras aiming at a subject, where each camera is built in two spectral bands, visible and thermal. Specifically, the score fusion combines the face scores from three matchers (Circular Gaussian Filter, Face Pattern Byte, Linear Discriminant Analysis) and from two-spectral bands (visible and thermal). The decision fusion combines the score-fusion results (genuine or impostor) from left faces and right faces in stereo imaging. We present three score-fusion results using k-Nearest Neighbor fusion, binomial logistic regression, and Hidden Markov Model fusion, meanwhile two decision-fusion results using logical AND and OR. Our experiments are conducted with the Alcon State Univ. MultiSpectral Stereo face dataset that currently consists of the stereo face images of two spectral bands from 105 subjects. The experimental results show that score fusion can significantly improve the accuracy, whereas decision fusion (with AND rule) can reduce the FAR with a slight decrease in accuracy.*

Keywords: Circular Gaussian Filter (CGF); decision fusion; face pattern byte (FPB); multispectral face recognition; score fusion; stereo face imaging.

I. Introduction

Face recognition has relative low accuracy compared to fingerprint recognition and iris recognition. To improve face recognition, multispectral imagery is suggested as a viable solution to address this challenge. Chang *et al.* [1] demonstrated image quality enhancement of fusing multispectral face images (e.g., visible and thermal) but did not report any recognition performance. Bendada *et al.* [2] compared several face recognition algorithms (Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), etc.) using the Local Binary Pattern (LBP) and the local ternary pattern (LTP) extracted from four-band face images (i.e., Visible, Short-, Medium-, Long-Wave Infrared [SWIR, MWIR, LWIR]), and found that LDA algorithm performed the best on the Visible dataset (92%). No fusion results were presented in their work. In contrast with image fusion (with the prerequisite of image registration) and feature fusion (with the process of high dimensional data) [3], [4], score-level fusion is more efficient due to utilization of low dimensional data. Nandakumar *et al.* [5] showed that a density-based fusion estimated by Gaussian Mixture Model (GMM) outperformed any single matcher and other fusion methods (like sum rule with min-max). Poh *et al*. [24] reported that the more biometric scores used in score fusion, the higher fusion recognition performance achieved. Zheng *et al.* [6] recently had a brief survey on the score fusion methods working with multiple matchers and multispectral faces, and found that fusing multimodal data was more effective and significant than combing variant matchers and using different fusion methods. Their experiments also showed that the Hidden Markov Model (HMM) was the best fusion method.

A real face recognition system expects a low *False accept Rate* (FAR) as well as high recognition *accuracy*. Thus, we propose to improve the performance of a face recognition system using score fusion of multispectral images and decision fusion of stereo images. As we know, multimodal score fusion can improve the accuracy meanwhile lowering the FAR. This paper will investigate whether decision fusion [26] using stereo images can further reduce FAR without sacrificing much accuracy. In addition, a new face pattern extraction using a set of Circular Gaussian filters (CGF) will be introduced. The proposed Face Score and Decision Fusion (FSADF) solution is illustrated in Fig. 1, where the face recognition system consists of two stereo imaging cameras (Left and Right). Each side has two spectral bands, visible and thermal. The face scores from multiple matchers are fused. The final decision is made by ANDing (ORing) the two fusion outcomes from Left side and Right side, respectively.

The remainder of this paper is organized as follows. Three face recognition algorithms (CGF, Face Pattern Byte [FPB], LDA) are briefly described in Section 2. The score combination methods (mean, k-Nearest Neighbor [KNN], Binomial Logistic Recognition [BLR], HMM) are summarized in Section 3, which are used, when optimal, for score fusion. The decision fusion and performance evaluation are depicted in Section 4. The experiments are conducted on the *Alcorn State Univ. Multispectral Stereo* (ASUMSS) face dataset [7], and the experimental results and discussion are presented in

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **JUL 2013**	2. REPORT TYPE	3. DATES COVERED **00-00-2013 to 00-00-2013**

4. TITLE AND SUBTITLE **Score Fusion and Decision Fusion for the Performance Improvement of Face Recognition**	5a. CONTRACT NUMBER
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Alcorn State University,Lorman,MS,39096**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR’S ACRONYM(S)
	11. SPONSOR/MONITOR’S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

Presented at the 16th International Conference on Information Fusion held in Istanbul, Turkey on 9-12 July 2013. Sponsored in part by Office of Naval Research Global.

14. ABSTRACT

To improve the performance of a face recognition system, we propose a fusion solution consisting of score fusion of multispectral images and decision fusion of stereo images. Score fusion combines several scores from multiple matchers and/or multiple modalities, which can increase the accuracy of face recognition. In a face recognition system, low false accept rate (FAR) is as important as high accuracy rate. The FAR can be reduced by using decision fusion of stereo images. The stereo face images are taken with two identical cameras aiming at a subject, where each camera is built in two spectral bands, visible and thermal. Specifically, the score fusion combines the face scores from three matchers (Circular Gaussian Filter, Face Pattern Byte, Linear Discriminant Analysis) and from two-spectral bands (visible and thermal). The decision fusion combines the score-fusion results (genuine or impostor) from left faces and right faces in stereo imaging. We present three score-fusion results using k-Nearest Neighbor fusion, binomial logistic regression, and Hidden Markov Model fusion, meanwhile two decision-fusion results using logical AND and OR. Our experiments are conducted with the Alcon State Univ. MultiSpectral Stereo face dataset that currently consists of the stereo face images of two spectral bands from 105 subjects. The experimental results show that score fusion can significantly improve the accuracy, whereas decision fusion (with AND rule) can reduce the FAR with a slight decrease in accuracy.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a REPORT **unclassified**	b ABSTRACT **unclassified**	c THIS PAGE **unclassified**	**Same as Report (SAR)**	**8**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

Section 5. Finally, conclusions are drawn in Section 6.

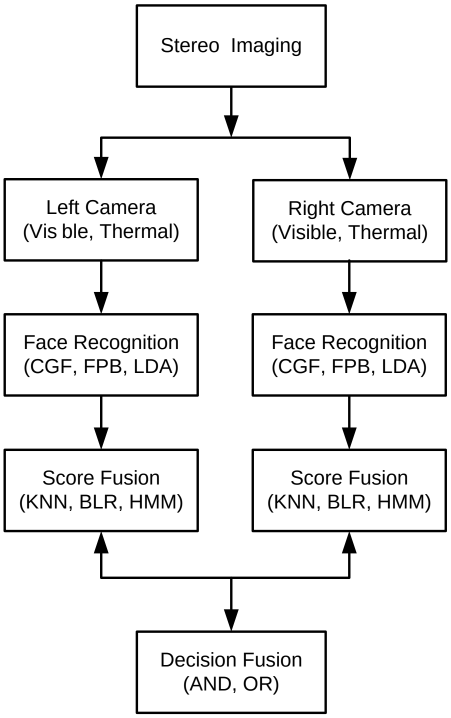

Fig. 1: Diagram of the proposed FSADF fusion solution for a face recognition system: The system consists of two stereo imaging cameras (Left and Right). Each side has two spectral bands, visible and thermal. The face scores from multiple matchers are fused. The final decision is made by ANDing (ORing) the two fusion outcomes from Left side and Right side, respectively.

II. FACE RECOGNITION METHODS

A given face image is first preprocessed with normalization, face detection and face alignment. Face recognition algorithms are performed with the preprocessed images. Three face recognition methods (matchers) are selected for score fusion, which are Circular Gaussian Filter (CGF), Face Pattern Byte (FPB), and Linear Discriminant Analysis (LDA).

A. Circular Gaussian Filter

We propose to extract facial features using a set of band-pass filters formed by rotating a 1-D Gaussian filter (off center) in frequency space, termed as "Circular Gaussian Filter" (refer to Eq. (1) and Fig. 2). The CGF convolutes a Gaussian function with an image. A CGF can be uniquely characterized by specifying a central frequency (*f*) and a frequency band (σ).[1]

In Fourier frequency domain, a *Circular Gaussian Filter* (CGF) is defined as follows:

$$CGF(u,v) = \frac{1}{2\pi\sigma^2} e^{-\frac{u_1^2+v_1^2}{2\sigma^2}}, \qquad (1)$$

where

$$u_1 = (u - f\cos\theta)\cos\theta - (v - f\sin\theta)\sin\theta, \qquad (2a)$$

$$v_1 = (u - f\cos\theta)\sin\theta + (v - f\sin\theta)\cos\theta, \qquad (2b)$$

where f specifies a central frequency, σ defines a frequency band, and $\theta \in [0, 2\pi]$.

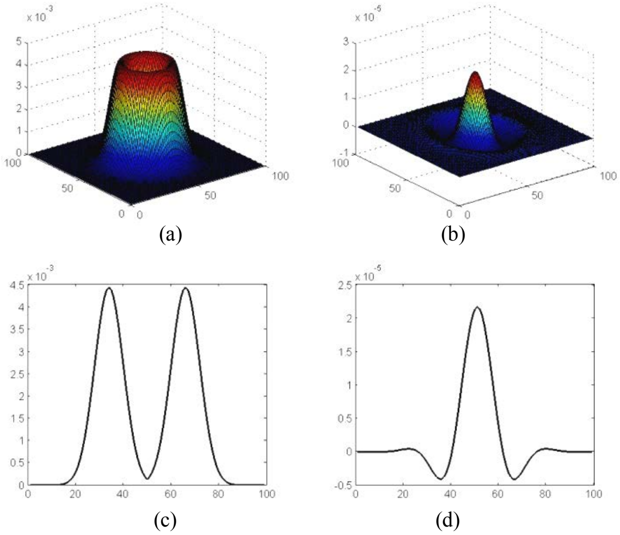

Fig. 2: Illustration of CGF with $f = 16$ and $\sigma = 6$ (Only the central part of CGF is presented): (a) CGF in frequency domain; (b) CGF in spatial domain; (c) A central slice of (a); (d) A central slice of (b).

Face patterns are formed pixel by pixel with the CGF-filtered images, termed as FP-CGF. Typically, the CGF filters at four bands (i.e., $f \in \{8, 16, 32, 64\}$ and $\sigma \in \{3, 6, 12, 24\}$) produce promising results with face images of 320×320-pixel resolution. Four phase images and four magnitude images are derived from four CGF filters, respectively. The FP-CGF has the same resolution as a face image and the depth of 8 bits per pixel. In each FP-CGF pixel, its most significant 4 bits are from the four binarized phase images, while its least 4 bits are from the four binarized magnitude images. In our experiments, the threshold for phase images is a range (i.e., 1 if *phase* > -0.15π and *phase* < 0.15π); whereas the threshold for the normalized magnitude images is 0.1 (i.e., 1 if *magnitude* > 0.1).

A Hamming distance (HD) [7] is calculated with the FP-CGF to measure the similarities among faces. The matched face has the shortest HD from the probe (query) face. In this paper, FP-CGF is also shorted as CGF to simplify the notation.

B. Face Pattern Byte

Face pattern bytes are comprised of the binary bit code of maximal orientational responses from a set of Gabor Wavelet Transforms (GWT). Multiple-band orientational codes are then put into a *face pattern byte* (FPB) pixel-by-pixel. A HD is calculated with the FPB to measure the similarities among faces, and recognition is made with the shortest HD.

[1] We chose not to use a Gabor Filter as that requires various scales and rotations; whereas we are interested in the band selection

A 8×16 GWT (with 8 bands by 16 orientations) produces 256 coefficients (128 magnitudes plus 128 phases) per pixel. We only use the magnitudes of GWT coefficients to create an 8-bit FPB for each pixel on a face image. First, at each pixel, an M-dimensional vector, $\mathbf{V}_m$, is created to store the orientation code representing the orientational magnitude strength (O_{mn}) at each frequency (refer to Eq. (3a)). At each frequency band (from 1 to 8), the index (0-15) of the *maximal* magnitude among 16 orientations is coded with 4 bits in order to maximize the orientation difference (among subjects in the database). The 4-bit orientational code is first put into an 8-dimensional vector ($\mathbf{V}_m$) by the frequency order, i.e., the lowest (highest) frequency corresponds to the lowest (highest) index of $\mathbf{V}_m$. A FPB is then formed with the most frequent orientations (called the *mode*) among some bands (refer to Eqs. (3b-c)). Specifically, the high half-byte (the most significant 4 bits) in a FPB, FPB_{HHB}, is the mode of high 4 bands (m = 5~8); whereas FPB_{LHB} is the mode of low 4 bands (m = 1~4). A factor, B_{FM}, will avoid choosing any orientation mode of a pattern in FPB if its frequency is only 1. This could make FPB immunized from noise. B_{FM} = 1 only when the frequency of the mode, f_{M}, is greater than or equal to 2, otherwise B_{FM} = 0 (see Eq. (3d)). It is clear that a FPB can code up to 16 orientations ($N \leq 16$) but there is no limit to the number of bands (M). 8 or 16 orientations are good for face recognition.

$$\mathbf{V}_m = O^{\mathrm{BC}}[Index(Max(O_{mn}))], \tag{3a}$$

$$FPB_{\mathrm{LHB}} = Mode(\mathbf{V}_m|_{m=1\sim M/2}) \cdot B_{\mathrm{FM}}, \tag{3b}$$

$$FPB_{\mathrm{HHB}} = Mode(\mathbf{V}_m|_{m=M/2+1\sim M}) \cdot B_{\mathrm{FM}}, \tag{3c}$$

$$B_{\mathrm{FM}} = \begin{cases} 1, \text{if } f_{\mathrm{M}} \geq 2 \\ 0, \text{otherwise} \end{cases} \tag{3d}$$

In Eqs. (3a-d), O_{mn} is the GWT magnitude at Frequency m and Orientation n; $m \in [1, M]$ and $n \in [1, N]$. The FPBs are stored as the feature vectors of the gallery faces in database, which will be compared with that of the probe face during the recognition process. A FPB (8 bits per pixel for 8×16 GWT) is usually stored in a byte.

Because the HD is computed by checking the bitwise difference between two face patterns (e.g., FPBs), the orientational bit code should favor the HD calculation. The FPBs are designed to reflect the orientational significance (or strength) along with frequency scales (locations). Therefore, the closer (neighboring) orientations should have less bitwise difference, whereas the further (orthogonal) orientations should have more bitwise difference. A set of optimal bit codes should minimize the HD of the neighboring orientations, as well as maximize the HD of orthogonal orientations [7]. One "optimized coding" solution of N = 16 (orientations) is as follows: {1110, 1100, 1000, 1010, 0010, 0110, 0100, 0000, 0001, 0101, 0111, 0011, 1011, 1001, 1101, 1111} (used in our experiments).

C. *Linear Discriminant Analysis*

There are many algorithms developed in face recognition domain in past two decades. Three classical algorithms (according to National Institute of Standards and Technology) are Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) [8-9], and Elastic Bunch Graph Matching (EBGM) [10-12]. However, PCA is not selected for score fusion due to its poor recognition performance [6] [7]. EBGM is not used for score fusion because it is time consuming and similar with FPB in technology and performance [6] [7]. On the other hand, LDA is chosen for score fusion since it has a better performance than PCA and runs faster than EBGM.

LDA [8] is designed to find an efficient way to represent the faces using the face class information. The *Fisherface* algorithm [9] is derived from the Fisher Linear Discriminant, which uses class specific information. By defining different classes with different statistics, the images in the training set are divided into the corresponding classes. Then, techniques similar to PCA algorithm are applied. The Fisherface algorithm results in a higher accuracy rate in recognizing faces when compared with Eigenface algorithm [9].

III. Score fusion

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 calculated by a simple arithmetic operation [13] such as taking the summation, average (mean), product, minimum, maximum, median, or majority vote. 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 [14], [15]. The choices of classifiers include linear discriminant analysis (LDA) [16], binomial logistic regression (BLR) [23], [24], k-nearest neighbors (KNN), artificial neural network (ANN), support vector machine (SVM), etc. 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 [17], [18].

Based on our previous experience [6], four score-fusion methods: mean fusion, KNN fusion, BLR fusion, and HMM fusion, are selected in our study.

A. *Score normalization and cross validation*

Score normalization is expected before fusing multiple scores since the multimodal scores from various modalities are heterogeneous and thus have differing dynamic ranges. To fuse the variant scores, it is required that all original scores are either similarity scores or distance scores (but not the mix of similarity and distance).

The source scores may originate from a different type of device, called *modality* (e.g., digital camera or thermal camera), and/or from variant analysis software, called *matcher* (e.g., CGF, FPB, or LDA algorithm). The large variances of multimodal scores were caused either by different matching algorithms or by different natures of biometrical data. In our experiments, a standard *z*-score *normalization* procedure is applied to all biometric scores,

$$\mathbf{S}_{\mathrm{N}} = (\mathbf{S}_0 - \boldsymbol{\mu}_0) / \boldsymbol{\sigma}_0, \tag{4}$$

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

To sufficiently use the sample data in recognition evaluation, *k*-fold cross validation is applied to split the original data into training and testing subsets. In a 10-fold cross validation (i.e., $k = 10$), for example, all scores are divided into 10 subsets. In each of 10 runs, one subset is held out for testing, and the remaining 9 subsets are used for training. In the next run, a different subset is used for testing until all 10 subsets are used for tests. The recognition performance is the averaged result of 10 runs. For fair comparisons, the same group of *k* folds will be used for all selected fusion methods.

B. Binomial logistic regression

Logistic regression (LR) [23] is a regression analysis to predict the outcome of a categorical dependent variable based on a set of predictor variables. The probability describing the possible outcome of a single trial, $P(\mathbf{x})$, is modeled as a function of explanatory variables (predictors), using a logistic function as defined below.

$$P(\mathbf{x}) = \frac{1}{1 + \exp[-g(\mathbf{x})]}, \tag{5}$$

where

$$g(\mathbf{x}) = \sum_{j=1}^{M} \beta_j x_j + \beta_0, \tag{6}$$

where x_j are the elements in $\mathbf{x}$. The weight parameters β_j are optimized to maximize the likelihood of the training data given the LR model [23]. $P(\mathbf{x})$ always takes on values between zero and one.

Logistic regression can be multinomial or binomial. Multinomial logistic regression (MLR) refers to cases where the outcome can have three or more possible types. *Binomial logistic regression* (BLR) refers to the instance in which the observed outcome can have only two possible types. For example, a face score fusion only outputs two types of results, impostor or genuine, which makes BLR [24] suitable for face scores. Generally, the outcome is coded as 0 (impostor) and 1 (genuine) in BLR as it leads to the most straightforward interpretation.

C. Hidden Markov Model for score fusion

The HMM fusion is a hybrid classifier-based and density-based 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 imaging devices, visible and thermal cameras); while multi-matcher scores (also referred to as *intra-modality* scores) result from different software algorithms using the same modality (e.g., three face scores generated from three face recognition algorithms).

For HMM training, a large database with known users (labeled with subject classifications) is expected, and thus a *k*-fold cross validation is utilized to satisfy this need. All scores are normalized and organized as the inputs of HMM models using *k*-fold cross validation. The HMM model is adapted to multimodal score fusion 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. By placing n pieces of m-dimensional observation vectors together, an observation sequence (over time, t) is formed. g is the number of 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. The details of the HMM model and its adaption to biometric score fusion are described elsewhere [6].

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 ($n = 2$; visible, thermal) and 3 matching algorithms (matchers) for each modality ($m = 3$). Thus, the length of $\mathbf{O}(t,s)$ is 6 (refer to the cells at the rightmost column in Table 2). The *observation symbol probabilities* matrix can be initialized with the GMM, where the number of Gaussian components (g) in each state are usually fixed (e.g., $g = 3$) or automatically decided [19]. Notice that two HMM models, λ_{Gen} and λ_{Imp}, are actually trained using genuine scores and impostor scores, respectively; their parameters can be estimated using the Baum-Welch algorithm [20]. An unlabeled biometric score sequence, $\mathbf{O}$, will be classified as a "genuine user" if $\mathrm{P}_{\mathrm{Gen}}(\mathbf{O}|\lambda_{\mathrm{Gen}}) > \mathrm{P}_{\mathrm{Imp}}(\mathbf{O}|\lambda_{\mathrm{Imp}}) + \eta$ (a simple decision rule); otherwise, $\mathbf{O}$ will be an "impostor user", where η is a small positive number empirically decided by experiments. 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 contain two or more (produced from different matching algorithms, e.g., CGF, FPB, and LDA). 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 the same.

IV. Decision fusion and performance evaluation

Score fusion combines multiple scores (either a distance or similarity score of a probe image); whereas decision fusion manipulates multiple decisions (either genuine or impostor for a user). The performance evaluation of a face recognition system includes accuracy, false accept rate, and false rejection rate.

A. *Decision fusion*

A commonly used decision fusion method is majority voting [13], which requires the number of decision makers to be an odd number (3 or more) to avoid possible tie. In addition, the decision makers in the voting group should have a similar performance; however, a weighted sum may be considered if decision makers have differing performances.

Our face dataset (ASUMSS) contains two-band stereo face images (refer to Section 5.1). The decision fusion is more suitable to be applied to the stereo images rather than two-band images (visible and thermal) by considering their equivalent performance. Since each pair of stereo images consists of two images that are taken by two (left and right) cameras. Thus, majority vote is not applicable to the stereo images. Instead, we propose a decision fusion using two logical rules: AND, OR. For example, when using the AND rule, a user is *genuine* only when both score-fusion results from two stereo cameras are genuine. Given an example of KNN fusion on a visible dataset (e.g., ASUDC), the final decision is made by the AND value of two KNN results (genuine or impostor) from left and right cameras, respectively (refer to Fig. 1). The KNN results come from the score fusion of three matchers: CGF, FPB, and LDA.

B. *Performance evaluation*

The *genuine score* is the matching score resulting from two samples of a single user; while the *impostor score* is the matching score of two samples originating from different users. On a *closed dataset* (i.e., all query users are included in the database), the recognition performance can be measured by a *verification rate* (VR), the percentage of the number of correctly recognized users (i.e., genuine users are recognized as genuine) over the total number of users. A higher VR means a better recognition algorithm.[2]

On an *open dataset* or for a commercial biometric system, a query user (i.e., probe face) may not be contained in the database. To evaluate the system performance, we define the following terms based on a confusion matrix [26] (see Fig. 3):

$$\text{AC} = (\text{TN}+\text{TP}) / (\text{TN} + \text{FP} + \text{FN} + \text{TP}), \tag{7a}$$

$$\text{FAR} = \text{FPR} = \text{FP} / (\text{TN} + \text{FP}), \tag{7b}$$

$$\text{FRR} = \text{FNR} = \text{FN} / (\text{FN} + \text{TP}), \tag{7c}$$

$$\text{GAR} = \text{TPR} = \text{TP} / (\text{FN} + \text{TP}), \tag{7d}$$

$$\text{IRR} = \text{TNR} = \text{TN} / (\text{TN} + \text{FP}), \tag{7e}$$

where AC denotes *accuracy* that means genuine (or impostor) users recognized as genuine (or impostor) users; FAR stands for *false accept rate* (also called false positive rate), i.e., impostor users are recognized as genuine users (falsely accepted by the system); FRR is for *false rejection rate* (also called false negative rate), i.e., genuine users are recognized as impostor users (falsely rejected by the system). Similarly, we can also define *genuine accept rate* (GAR, or true positive rate), and *impostor rejection rate* (IRR, true negative rate) in Eqs. (7d) and (7e).

Actual \ Predicted	Predicted Negative	Predicted Positive
Actual Negative	True Negative (TN)	False Positive (FP)
Actual Positive	False Negative (FN)	True Positive (TP)

Fig. 3: Definition of *confusion matrix*: the 2-by-2 elements within the thick-line rectangle.

It is usually acceptable to evaluate a biometric system by reporting AC, FAR and FRR. Of course, an ideal biometric system expects a high AC, a low FAR and a low FRR. If there is no negative cases (i.e., no intruders), then TN + FP = 0 in Eq. (7a), which leads to AC = VR. In our experiments, we will first compare the AC values; the higher the better. If two AC values are equal or close, then the FAR values are compared; the lower the better. The FRR values are presented for reference but not analyzed for comparison.

V. Experimental design

The experiments were conducted on the Alcorn State University MultiSpectral Stereo (ASUMSS) dataset, which is a particular subset within the ASU MultiSpectral (ASUMS) database [7]. The descriptions of face dataset and experimental results are given in the following three subsections.

A. *Face dataset and experimental design*

The ASUMSS dataset currently consists of the stereo face images of two spectral bands (visible and thermal) from 105 subjects (refer to Fig. 4). The stereo images were acquired with two identical cameras that were horizontally placed on a fixture. The two cameras were spaced at 5.5 inches apart, and the middle line of two cameras was

[2] See ISO/IEC JTC SC37 Harmonized Biometric Vocabulary, Section C7 Application and C8 Performance

aimed at the middle line of a subject. The two cameras are referred to as left camera and right camera (according to the view of the photographer), thus the two stereo images taken by left-side camera and right-side camera are called left (face) image and right (face) image, respectively.

The model of two cameras is FLIR SC620, a two-in-one imaging device, wherein the infrared (thermal) camera is of 640×480 pixel original resolution and 7.5~13μm spectral range, and the digital (visible) camera is of 2048×1536 pixel original resolution. The visible dataset is denoted as ASUDC; while the thermal dataset is denoted as ASUIR. Combining with left (L) and right (R) side, the notations of all datasets are ASUDCL, ASUDCR, ASUIRL, and ASUIRR. ASUMSS denotes the entire dataset (refer to Table 1 and Table 2).

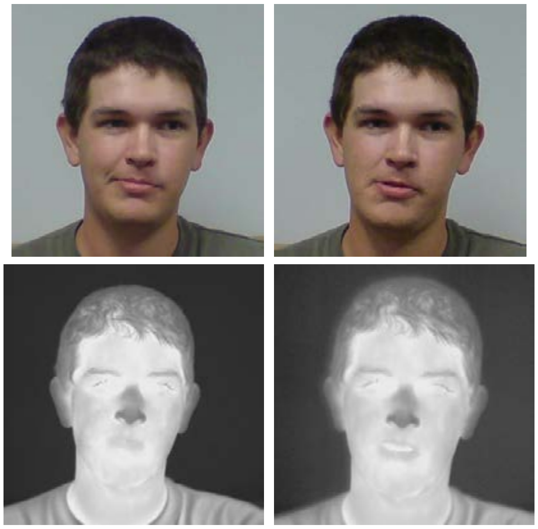

Fig. 4: Sample faces from the ASUMSS face database: Notice that the four images (visible/DC at top, thermal/IR at bottom) were acquired from the same subject. The two images at left (right) column were taken by the left (right) camera. The images are the detected and aligned faces (320×320 pixels).

Four frontal face images were randomly selected from each subject (per side per band), one of which was used as a probe image, and three of which were used as gallery images. For example, a total of 16 images per subject (of two bands and of two sides in stereo imaging) were analyzed in our experiments. Then switch the role between probe and gallery, the reported performance of single algorithms in Table 1 is the average of four rotations. In LDA algorithm, the probe and gallery images serve as test and train images, respectively. All face images were normalized. The face portion was then detected and extracted using face detection algorithms [21], [22]. Notice that the size of an extracted face is usually smaller than the resolution of original image in database. Next, all faces were automatically lined up using image registration algorithms. The visible images (originally in colors as shown in Fig. 4) were converted to grayscale images prior to facial feature extraction.

The performance of single face recognition algorithm is tested on four datasets, ASUDCL, ASUDCR, ASUIRL, and ASUIRR, where the results of mean fusion are also presented in Table 1. The performance of score fusion and decision fusion varying with spectral band (DC, IR), stereo side (L, R), and decision rule (AND, OR) are exhibited in Table 2.

B. *Performance of single face recognition algorithm*

The three selected algorithms (CGF, FPB, LDA) were tested with the same four datasets as shown in Table 1. The FPBs were formed using 8×16 GWT [7], while the FP-CGF were extracted with 4-band CGF. For face matching CGF and FPB utilized Hamming distance, while LDA used Euclidean distance. The experimental results are presented in Table 1, where only the top-1 match (the shortest distance) results were reported.

TABLE 1: The accuracy (AC, %) and FAR (%) of three face recognition algorithms and their mean fusions tested on the 4 subsets of ASUMSS face dataset (of 105 subjects, 4 images per subject)

Dataset\Algorithm	**CGF**	**FPB**	**LDA**	Mean Fusion
ASUDCL	83.57, 16.19	90.24, 9.52	91.67, 8.10	96.19, 3.57
ASUDCR	85.71, 14.05	91.43, 8.33	91.67, 8.10	95.00, 4.76
ASUIRL	76.19, 23.57	89.52, 10.24	43.10, 56.67	93.81, 5.95
ASUIRR	75.00, 24.76	88.33, 11.43	44.05, 55.71	92.62, 7.14
Mean Fusion	96.90, 2.86	98.81, 0.95	93.33, 6.43	**99.29,** **0.48**

From the results shown in Table 1, the FPB is the most credible recognition algorithm across four datasets. In other words, the FPB is the best on the average, Mean(FPB)|AC = 98.81%, and CGF is the second best. LDA is poor and non-credible especially on two thermal datasets (its ACs lower than 50%) although LDA is the single best matcher (SBM), AC = 91.67%.

The results of mean fusion (by averaging all normalized scores) are listed in Table 1, where the matched face is of the shortest distance. At the bottom row, the three results at left are the fusion of four single-matcher scores from four datasets; while the four results at the rightmost column are the fusions of three single-set scores from three matchers. All mean fusions are better than the SBM (91.67%). For example, the mean fusion of LDA scores can achieve AC = 93.33% although its ACs with two IR datasets are very low. The result at low-right corner (AC = 99.29%) is the mean fusion of all 12 scores, which is clearly the best.

FAR values are significantly decreased in all fusion results, especially for the mean fusion of 12 scores (FAR = 0.48%, compared with FAR = 8.10% for SBM). FRR values are not presented in Table 1 since it is not applicable to the face matching by the shortest distance on a closed dataset. Future work could be verification performance measures (such as Equal Error Rate) for each method; however Table 1 shows that all are verified except LDA for ASUIR.

VI. PERFORMANCE WITH SCORE FUSION AND DECISION FUSION

All three score fusion methods require a training process. In addition to score normalization, two more processes are expected. First, the balance of genuine and imposter scores shall be considered in the training process since the majority of scores are impostor. To avoid possible bias in model training, six impostor scores per matcher per band per subject were randomly selected in each dataset for training. All genuine scores (i.e., three genuine scores per matcher per band per subject) as well as the reduced impostor scores are called the "trimmed dataset". The KNN, BLR, and HMM fusions were tested with the *trimmed dataset*. Secondly, a 10-fold cross validation is applied to the three fusions. More specifically, the same randomly split 10 folds were used in all tests shown in Table 2. In contrast, the mean fusion used all normalized scores without any cross validation.

In Table 2, the spectral variations (Datasets) are listed in the top row, wherein the numbers of fusion scores are also given. In each of three fusion methods (KNN, BLR and HMM), there are four different combinations, L (score fusion from left images), R (score fusion from right images), L&R (decision fusion by AND rule), and L|R (decision fusion by OR rule). Three performance values reported in each cell are AC, FAR, and FRR, respectively. To test and calculate FAR, the gallery images corresponding to that probe image must be excluded from the dataset, which simulates an open dataset that does not include the user (intruder).

The notation of KNN(ASUMSS,L&R)|AC = 98.55% denotes the AC value of the AND decision of two KNN fusions from two subsets (Left and Right) of ASUMSS. HMM(ASUMSS,L&R)|AC = 98.04% shows that KNN fusion is better than HMM fusion, both of which are better than BLR fusion, BLR(ASUMSS,L&R)|AC = 97.58%. The fusion results with other two subsets (ASUDC, ASUIR) exhibited the similar patterns. The fusion performance between left (L) and right (R) subsets are quite close (i.e., relatively equivalent performance). It seems that, the more fusion scores, the higher the recognition accuracy. For example, the fusion performance on ASUMSS (6 scores) are better than those on its subsets (ASUDC, ASUIR). Certainly, all fusion methods exhibited in Table 2 definitely improve the recognition accuracy in contrast with the SBM (AC = 91.67%) given in Table 1.

To investigate the impact of decision fusion, we need to compare both AC and FAR values. Look at the rightmost column (ASUMSS), and compare L&R with L|R. Comparing KNN(ASUMSS,L&R)|AC,FAR = 98.55%,0.14% with KNN (ASUMSS,L|R)|AC,FAR = 98.57%,1.65% shows decreases in both AC and FAR. A similar situation occurs in BLR fusions, where the decision fusion of (L&R) has AC = 97.58% and FAR = 0.33%. Compared to BLR(ASUMSS,L|R)|AC,FAR = 98.45%,1.57%, its decision fusion causes a relatively significant decrease of FAR but only a slight decrease of AC. The FRR values of KNN fusions are lower than those of BLR and HMM fusions. In terms of three performance measures (AC, FAR, FRR), the AND-decision upon KNN fusion is superior to that of BLR and HMM fusion.

TABLE 2: The values (%) of AC, FAR, and FRR of three fusion methods tested on variant combinations of subsets in the ASUMSS dataset (Note: ASUMSS = ASUDC + ASUIR; L = Left; R = Right; & = AND; | = OR)

Dataset (# Scores)		**ASUDC** (3)	**ASUIR** (3)	**ASUMSS** (6)
KNN Fusion	L	97.21, 1.76, 5.79	96.53, 3.76, 2.62	98.83, 0.73, 2.46
	R	96.97, 1.81, 6 59	95.50, 5.12, 2.70	98.28, 1.06, 3.65
	L & R	97.05, 0.30, 10.71	98.28, 0.73, 4.60	**98.55, 0.14, 5.32**
	L | R	97.13, 3.28, 1.67	93.74, 8.15, 0.71	98.57, 1.65, 0.79
BLR Fusion	L	97.13, 1.30, 7.46	94.49, 1.71, 16.67	98.14, 0.79, 5.00
	R	96.91, 1.49, 7.78	93.64, 2.25, 18.41	97.88, 1.11, 5.08
	L & R	96.39, 0.51, 12.70	92.43, 0.60, 28.02	97.58, 0.33, 8.57
	L | R	97.66, 2.28, 2 54	95.70, 3.36, 7.06	98.45, 1.57, 1.51
HMM Fusion	L	97.07, 2.25, 4 92	93.50, 3.95, 13.97	98.28, 1.14, 3.41
	R	96.75, 2.74, 4.76	93.05, 5.39, 11.51	97.54, 2.30, 2.94
	L & R	97.11, 0.95, 8 57	93.28, 1.30, 22.62	98.04, 0.70, 5.63
	L | R	96.71, 4.04, 1 11	93.28, 8.04, 2.86	97.78, 2.74, 0.71

The difference of two decision fusions, AND versus OR, are summarized below. The AND decision results a lower FAR but a higher FRR; whereas the OR decision results higher FAR but lower FRR. The ACs of two decisions are close. If a low FAR is crucial (e.g., for security related applications), AND-decision fusion is preferred.

The experimental results with the current ASUMSS dataset shows the KNN fusion performs the best in terms of high AC and low FAR. And the performance of HMM fusion is very close to KNN fusion. In fact, our earlier experiments (tested on different datasets) [6] proved that

HMM fusion is very credible for multimodal biometric score fusion.

Table 1 shows that the performance of mean fusion with 12 scores reaches AC = 99.29% and FAR = 0.48%, which is actually the best in our experiments. The possible reason might be that the genuine scores and the impostor scores are well separated, which makes a linear separation (like mean fusion) ideal. Surprisingly, in another independent research [25], 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. However, the credibility of mean fusion needs further investigation on different databases.

In the future we will expand the ASUMSS dataset and sufficiently test the proposed score fusion and decision fusion on a larger database. We will also investigate and verify the current findings by using other biometric modalities (like fingerprint, iris, more spectral images) and assess computational complexity.

VII. Conclusions

A commercial biometric system typically expects high accuracy and low false accept rate (FAR). We propose a solution (FSADF) for the performance improvement of a face recognition system, which consists of score fusion of multispectral face images and decision fusion of stereo face images. Experimental results show that the score fusion can significantly improve the accuracy, whereas the decision fusion can reduce the FAR. After applying the AND-decision fusion to the score-fusion results from left and right images, the FAR is reduced as well as a slight decrease of accuracy.

A new face pattern extraction using a set of circular Gaussian Filters (FP-CGF) is proposed and tested. Its performance is better than LDA and lower than FPB. The FP-CGF is quite reliable across four face subsets. Finally, we showed that KNN is better than BLR and HMM fusion.

Acknowledgment

This research is funded by the Department of Defense Research and Education Program administered by the Army Research Office and the Office of Naval Research under grant No. W911NF-12-1-0065.

References

[1] H. Chang, A. Koschan, M. Abidi, S. G. Kong, C.-H. Won, “Multispectral visible and infrared imaging for face recognition,” *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition Workshops*, 2008.

[2] A. Bendada, M. A. Akhloufi, “Multispectral Face Recognition in Texture Space,” *Canadian Conf. on Computer and Robot Vision*, pp.101-106, 2010.

[3] G. Bebis, A. Gyaourova, S. Singh and I. Pavlidis, “Face recognition by fusing thermal infrared and visible imagery,” *Image and Vision Computing*, Vol. 24, pp.727-242, 2006.

[4] O. Arandjelovic, R. I. Hammoud, R. Cipolla, “On Person Authentication by Fusing Visual and Thermal Face Biometrics,” *Proc. of the IEEE Int’l Conf. on Video and Signal Based Surveillance*, 2006.

[5] 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.

[6] Y. Zheng, A. Elmaghraby, “A brief survey on multispectral face recognition and multimodal score fusion,” *IEEE Int’l Symp. on Signal Processing and Info. Technology* (ISSPIT) 2011: 543-550.

[7] Y. Zheng, “Orientation-based face recognition using multispectral imagery and score fusion,” *Opt. Eng*. 50, 117202, 2011.

[8] W. Zhao, R. Chellappa, A. Krishnaswamy, “Discriminant Analysis of Principal Components for Face Recognition,” Proc. *IEEE Int’l Conf. on Face and Gesture Recognition*, FG'98, 14-16, pp. 336-341, 1998.

[9] J. Lu, K.N. Plataniotis, A.N. Venetsanopoulos, “Face Recognition Using LDA-Based Algorithms,” *IEEE Trans. on Neural Networks*, Vol. 14, No. 1, pp. 195-200, 2003.

[10] 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.

[11] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg, “Face Recognition by Elastic Bunch Graph Matching,” in L. C. Jain, *et.al* (eds.), *Intelligent Biometric Techniques in Fingerprint and Face Recognition*, pp. 355-396, CRC Press, USA, 1999.

[12] W. Zhao, R. Chellappa, A. Rosenfeld, P.J. Phillips, “Face Recognition: A Literature Survey,” *ACM Computing Surveys*, 35(4), pp. 399-458, 2003.

[13] L.I. Kuncheva, “A Theoretical Study on Six Classifier Fusion Strategies,” *IEEE Trans. Pattern Anal. Mach. Intell*. 24 (2), 281-286, 2002.

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

[15] 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.

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

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

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

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

[20] L.E. Baum, T. Petrie, “Statistical inference for probability functions of finite state markov chains,” *Ann. Math. Stat*. 37, 1554–1563, 1966.

[21] Y. Zheng, “Face detection and eyeglasses detection for thermal face recognition,” *Proc. SPIE* 8300, 2012.

[22] K. Reese, Y. Zheng, A. Elmaghraby, “A Comparison of Face Detection Algorithms in Visible and Thermal Spectrums,” *Int’l Conf. on Advances in Computer Science and Application*, 2012.

[23] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. New York: Chapman & Hall, 1990.

[24] N. Poh, T. Bourlai, et al., Benchmarking Quality-dependent and Cost-sensitive Multimodal Biometric Fusion Algorithms, *IEEE Trans. on Information Forensics and Security*, 4(4), pp. 849–866, 2009.

[25] 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.

[26] B. Kahler and E. Blasch, “Decision-Level Fusion Performance Improvement from Enhanced HRR Radar Clutter Suppression,” *J. of. Advances in Information Fusion,* Vol. 6, No. 2, Dec. 2011.