Hindawi Publishing Corporation
EURASIP Journal on Image and Video Processing
Volume 2007, Article ID 38205, 11 pages
doi:10.1155/2007/38205

Research Article

Fusion of Appearance Image and Passive Stereo Depth Map for Face Recognition Based on the Bilateral 2DLDA

Jian-Gang Wang,[1] Hui Kong,[2] Eric Sung,[2] Wei-Yun Yau,[1] and Eam Khwang Teoh[2]

[1] *Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore 119613*
[2] *School of Electrical and Electronic Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798*

Received 27 April 2006; Revised 22 October 2006; Accepted 18 June 2007

Recommended by Christophe Garcia

This paper presents a novel approach for face recognition based on the fusion of the appearance and depth information at the match score level. We apply passive stereoscopy instead of active range scanning as popularly used by others. We show that present-day passive stereoscopy, though less robust and accurate, does make positive contribution to face recognition. By combining the appearance and disparity in a linear fashion, we verified experimentally that the combined results are noticeably better than those for each individual modality. We also propose an original learning method, the bilateral two-dimensional linear discriminant analysis (B2DLDA), to extract facial features of the appearance and disparity images. We compare B2DLDA with some existing 2DLDA methods on both XM2VTS database and our database. The results show that the B2DLDA can achieve better results than others.

1. INTRODUCTION

A great amount of research effort has been devoted to face recognition based on 2D face images [1]. However, the methods developed are sensitive to the changes in pose, illumination, and face expression. A robust identification system may require the fusion of several modalities because ambiguities in face recognition can be reduced with complementary multiple-modal information fusion. A multimodal identification system usually performs better than any one of its individual components, particularly in noisy environments [2]. One of the multimodal approaches is 2D plus 3D [3–7]. A good survey on 3D, 3D-plus-2D face recognition can be found in [8]. Intuitively, a 3D representation provides an added dimension to the useful information for the description of the face. This is because 3D information is relatively insensitive to change in illumination, skin-color, pose, and makeup; that is, it lacks the intrinsic weakness of 2D approaches. Studies [3–7, 9] have demonstrated the benefits of having this additional information. On the other hand, 2D image complements well 3D information. They are localized in hair, eyebrows, eyes, nose, mouth, facial hairs, and skin color precisely, where 3D capture is difficult and not accurate.

There are three main techniques for 3D facial surface capture. The first is by passive stereo using at least two cameras to capture a facial image and using a computational matching method. The second is based on structured lighting, in which a pattern is projected on a face and the 3D facial surface is calculated. Finally, the third is based on the use of laser range finding systems to capture the 3D facial surface. The third technique has the best reliability and resolution while the first has relatively poor robustness and accuracy. The attraction of passive stereoscopy is in its nonintrusive nature which is important in many real-life applications. Moreover, it is low cost. This serves as our motivation to use passive stereovision as one of the modalities of fusion and to ascertain if it can be sufficiently useful in face recognition. Our experiments, to be described later, will justify its use.

Currently, the 3D facial surface data quality obtained from the above three techniques is not comparable to that of the 2D images from a digital camera. The reason is that the 3D data usually have missing data or voids in the concave area of a surface, eyes, nostrils, and areas with facial hair. These issues are not problematic to an image from a digital camera. The facial surface data available to us from the XM2VTS database is also coarse (∼4000 points) compared to a 2D image (3 to 8 million pixels) from a digital camera and also compared to other 3D studies [3, 4], where they had around 200 000 points on the facial surface area. The cost of a 3D scanner is also much higher compared to a digital camera for taking 2D images.

While a lot of work has been carried out in face modeling and recognition, 3D information is still not widely used for recognition [10–12]. Initial studies concentrated on curvature analysis [13–15]. The existing 3D face recognition techniques proposed [10, 11, 16–22] assume the use of active 3D measurement for 3D face image capture. However, active methods employ structured illumination (structure projection, phase shift, etc.) or laser scanning, which is not desirable in many applications. Thanks to the technical progress in 3D capture/computing, an affordable real-time passive stereo system has become available. In this paper, we set out to find out if present-day passive stereovision in combination with 2D appearance images can match up to other methods relying on active depth data. Our main objective is to propose a method of combining appearance and depth face images to improve the recognition rate. While 3D face recognition research dates back to before 1990, algorithms that combine results from 3D and 2D data did not appear until about 2000 [17]. Pan et al. [23] used the Hausdorff distance for feature alignment and matching for 3D recognition. Recently, Chang et al. [3, 4, 16] applied principal components analysis (PCA) with 3D range data along with 2D image for face recognition. A Minolta Vivid 900 range scanner was used to obtain 2D and 3D images. Chang et al. [16] investigated the comparison and combination of 2D, 3D, and IR data for face recognition based on PCA representations of the face images. We note that their 3D data were captured by active scanning. Tsalakanidou [5] developed a system to verify the improvement of the face recognition rate by fusing depth and color eigenfaces on the XM2VTS database. The 3D models in the XM2VTS database are built using an active stereo system provided by the Turing Institute [24]. It can be seen that the recognition performance has been improved by using 3D information from the mentioned literature.

PCA and Fisher linear discriminant analysis (LDA) are common tools for facial feature extraction and dimension reduction. They have been successfully applied to face feature extraction and recognition [1]. The conventional LDA is a 1D feature extraction technique, and so a 2D image must first be vectorised before the application of LDA. Since the resulting image vectors are high-dimensional, LDA usually encounters the small sample size (SSS) problem in which the within-class scatter matrix becomes singular. Liu et al. [25] substituted $S_t = S_w + S_b$ for S_b to overcome the singularity problem. Yang et al. [26] proposed a 2DPCA for face recognition. Recently, some 2DLDA methods have been published [27–30] to solve SSS problem. In contrast to the $\mathbf{S}_b$ and $\mathbf{S}_w$ of 1DLDA, the corresponding $\mathbf{S}_b$ and $\mathbf{S}_w$ obtained by 2DLDA are not singular. Ye et al. [27] developed a scheme of simultaneous bilateral projections, L and R, and an iteration process to solve the two optimal projection metrics. This simultaneous bilateral projection is essentially a reprojection of a body of discriminant features that will discard some information. The performance of Ye's method depends on the initial choices of the transform matrix, R_0, and may lead to a local optimal solution although they suggested an initial R_0 based on their experiments. The focus of Ye's method is on the reduction of computational complexity of the conventional LDA method. Comparing with the conventional Fisherfaces (PCA plus LDA), Ye et al. found that the improvement in recognition accuracy by their 2DLDA method is not significant [27]. Yang et al. [29] and Visani et al. [30] developed a similar 2DLDA. These methods applied LDA in horizontal direction, and then applied LDA on the final left-projected features. This reprojection, however, may discard some discriminant information.

We proposed a novel 2DLDA framework containing unilateral 2DLDA (U2DLDA) and bilateral 2DLDA (B2DLDA) to overcome the SSS problem [28]. In this paper, we adopt the B2DLDA to extract facial features of the appearance and disparity images. Face is recognized by combining the appearance and disparity in a linear fashion. Differing from the existing 2DLDA [27, 29, 30], the B2DLDA keeps more discriminant information because the two sets of optimal discriminant features, which are obtained from either step of the asynchronous bilateral projection, are combined together for classification. We have compared our method to Ye's method in this paper. It shows better performance than Ye's 2DLDA because of the larger amount of discriminant information. In this paper, we also extended our work in [28] by comparing it with the existing 2DLDA approaches on stereo face recognition.

2. STEREO FACE RECOGNITION

So far, the reported 3D face recognition [3, 10, 16, 17] is based on active sensor (structure light, laser), however, they are not desirable in many applications. In this paper, we used SRI stereo engine [31] that outputs a high enough range resolution (≤0.33 mm) for our applications. Our objective is to combine appearance and depth face images to improve the recognition rate. The performance of such fusion was evaluated on the commonly used database XM2VTS [32] and our own database collected by the real-time passive stereo vision system (SRI stereo engine, Mega-D [31]). The evaluation compares the results from appearance alone, depth alone, and the fusion of them, respectively. The performance using fused appearance and depth is the best among the three tests with a marked improvement of 5–8% accuracy. This justifies our method of fusion and also confirms our hypothesis that both modalities contribute positively. In Sections 2.1 and 2.2, we will discuss the generation of the 3D information of the XM2VTS and a passive stereo vision system. In Section 2.3, we will discuss the normalization of the 2D and 3D.

2.1. XM2VTS database

The XM2VTS is a large multimodal database. The faces are captured onto a high-quality digital video. It contains recordings of 295 subjects taken over a period of four months. Each recording contains a speaking head shot and a rotating head shot. Besides the digital video, the database provides high-quality color images, 32 KHz 16-bit sound files, and a 3D model, which deals with access control by the use of multimodal identification of human faces. The goal of using a multimodal recognition scheme is to improve the recognition efficiency by combining single modalities. We adopted

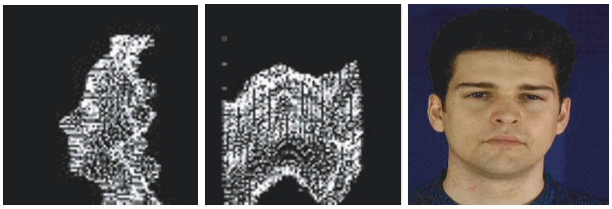


Figure 1: VRML model of a person's face.

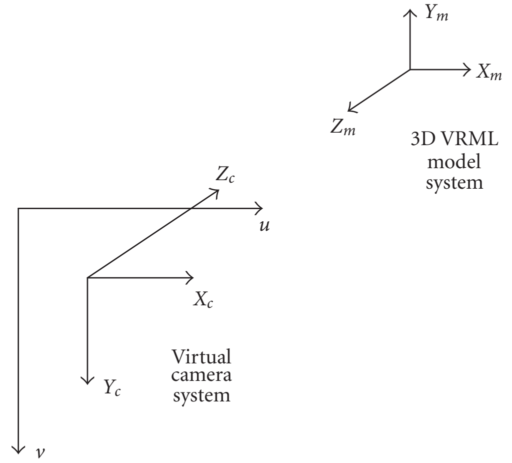


Figure 2: Geometric relationships among the virtual camera, 3D VRML model, and the image plane.

this database because 3D VRML models of subjects are provided and they can be used to generate the depth map for our algorithm. The high-precision 3D model of the subjects' head was built using an active stereo system provided by the Turing Institute [24]. In the following, we will discuss the generation of depth images from VRML model in the XM2VTS database.

A depth image is an image where the intensity of a pixel represents the depth of the correspondent point with respect to the 3D VRML model coordinate system. A 3D VRML model which contains the 3D coordinates and texture of a face in the XM2VTS database is displayed in Figure 1. There are about 4000 points in the 3D face model to represent the face. The face surface is triangulated with these points. In order to generate a depth image, a virtual camera is put in front of the 3D VRML model (Figure 2). The coordinate system of the virtual camera is defined as follows: the image plane is defined as the X-Y plane, the Z-axis is along the optical axis of the camera and pointing toward the frontal object. The camera plane, Y_c-Z_c, is positioned parallel to Y_m-X_m plane of the 3D VRML model. The Z_c coordinate aligns with Z_m coordinate, but in the reverse direction. X_c is antiparallel to X_m and Y_c is antiparallel to Y_m.

The intrinsic parameters of the camera must be properly defined in order to generate a depth image from a 3D VRML model. The parameters include (u_0, v_0), the coordinates of the image-center point (principle point); f_u and f_v, the scale factors of the camera along the u-axis and v-axis, respectively. The origin of the camera system under the 3D VRML model coordinate system is also set at (x_0, y_0, z_0).

The perspective projection pin-hole camera model is assumed. This means that for a point $F(x_m, y_m, z_m)$ in a 3D VRML model of a subject, the 2D coordinates of F in its depth image are computed as follows:

$$\begin{aligned} u &= u_0 + \frac{f_u x_m}{z_0 - z_m}, \\ v &= v_0 - \frac{f_v y_m}{z_0 - z_m}. \end{aligned} \tag{1}$$

In our approach, the z-buffering algorithm [33] is applied to handle the face self-occlusion for generating the depth images.

In the XM2VTS database, there is only one 3D model for each subject. In order to generate more than one view for learning and testing, some new views are obtained by rotating the 3D coordinates of the VRML model away from the frontal (about the Y_m axes) by some degrees. In our experiments, the new views are obtained at ±3°, ±6°, ±9°, ±12°, ±15°, ±18°.

2.2. Database collected by Mega-D

Here, we had used the SRI stereo head [31], in which the stereo process interpolates disparities up to 1/16 pixels. The resolution of the SRI stereo cameras is 640×480. Both intrinsic and extrinsic parameters are calibrated by an automatic calibration procedure. The smallest disparity change, Δd, is $(1/16)\times 7.5\,\mu\text{m} = 0.46875\,\mu\text{m}$. Here a pixel size of 7.5 μm. We used the Mega-D stereo head, where the baseline, b, is 9 cm and the focus length, f, is 16 mm. Hence when the distance from the subject to the stereo head, r, is 1 m, the range resolution, namely the smallest change in range that is discernable by the stereo geometry, is

$$\begin{aligned} \Delta r &= \left(\frac{r^2}{bf}\right)\Delta d \\ &= \left(1\,\text{m}^2/(90\,\text{mm} \times 16\,\text{mm})\right) \times 0.46875\,\mu\text{m} * 10^{-3} \\ &\approx 0.33\,\text{mm}. \end{aligned} \tag{2}$$

The range resolution is high enough for our face recognition applications. The manual of the SRI Small Vision System can be found in [31].

A database, called the Mega-D database, is collected using the SRI stereo head. The Mega-D database includes the images of 106 staff and students of our institute, with 12 pairs of appearance and disparity images for each subject. Two pairs per person are randomly selected for training while the remaining ten pairs are for testing. The recognition rate is calculated as the mean result of the experiments on these groups.

2.3. Normalizations of appearance and disparity images

Normalization is necessary to prevent the failure of similar face images of different sizes of the same person to be recognised. The normalization of an appearance image of the XM2VTS or the Mega-D database is as follows: the appearance image is rotated and scaled to occupy a fixed size array of pixels using the image coordinates of the outer corners of the two eyes. The eye corners are extracted by our morphologically based method [34] and should be horizontal in the normalized images.

The normalization of a depth image in the XM2VTS database is as follows. The z values of the all pixels in the image are subtracted by a value in order that the distances between the nose tip and the camera are the same for all images.

In order to normalize a disparity image in the Mega-D database, we need to detect the outer corners of the two eyes and the nose tip in the disparity image. In the SRI stereo head, the coordinates of a pixel in the disparity image are consistent with the coordinates of the pixel in the left appearance image. Hence we can (more easily) detect the outer eye corners in the left appearance image instead of in the disparity image. The tip of the nose can be detected in the disparity image using template matching [11]. From the coplanar stereo vision model, we have

$$D = \frac{bf}{d}, \tag{3}$$

where D represents the depth, d is the disparity, b is the baseline, and f is the focal length of the calibrated stereo camera. The parameters b and f can be calibrated by the small vision system automatically. Hence we can get the depth image of a disparity image with (3). Thereby the depth image is normalised, similar to that in the XM2VTS database, using the depth of the nose tip. After that, the depth image is further normalized similarly by the outer corners of the two eyes.

In our approach, the normalized color images are changed to the gray-level image by averaging three channels:

$$I = \frac{R+G+B}{3}. \tag{4}$$

The parameters in (1) are set as

$$\begin{aligned} u_0 &= v_0 = 0,\\ f_x &= f_y = 4500,\\ x_0 &= y_0 = 0,\\ z_0 &= 20. \end{aligned} \tag{5}$$

Problems with the 3D data are alleviated to some degree by a preprocessing step to fill in holes (a region where there is missing 3D data during sensing) and spikes. We remove the holes by a median filter followed by linear interpolation of missing values from good values around the edges of the holes.

Some of the normalized face image samples in the XM2VTS database are shown in Figure 3, where color face images are shown in Figure 3(a) and the corresponding depth images are shown in Figure 3(b). The size of the normalized image is 88 × 64. We can see significant changes in illumination, expressions, hair, and eye glasses/no eyeglasses due to longer time lapse (four months) in photograph taking.

Samples of the normalized face images in the Mega-D database are shown in Figures 4 and 5. Both color face images and the corresponding disparity images are shown in Figure 4. The resolution of the images is 88 × 64. The distance between the subjects and the camera is about 1.5 m. We can see some changes in illumination, pose, and expression in Figure 5.

3. FEATURE EXTRACTION

We have proposed a bilateral two-dimensional linear discriminant analysis (B2DLDA) [28] to solve the small sample size problem. In this paper, we apply it to extract features of appearance and depth images. Here, we will extend the work in [28] by comparing it with existing 2DLDA approaches [27, 29, 30].

3.1. B2DLDA algorithm

The pseudocode for the B2DLDA algorithm is given in Algorithm 1.

For face classification, $\mathbf{W}_l$ and $\mathbf{W}_r$ are applied to a probe image to obtain the features B_l and B_r. The B_l and B_r are converted to 1D vector, respectively. PCA is adopted to classify the concatenated vectors of $\{B_l, B_r\}$. It is noted that PCA or LDA can be used in this step. Ye et al. [27] adopted LDA to reduce the dimension of 2DLDA, since a small reduced dimension is desirable for efficient querying. We used PCA because we try to keep as much structure of the features (variance). There are at most $C-1$ discriminant components corresponding to nonzero eigenvalues. Their numbers, m_l and m_r, can be selected using the Wilks Lambda criteria, which is known as the stepwise discriminant analysis [35]. This analysis shows that the number of discriminant components required by left and right transforms for our case is 20. So for our experiments, we set $m_l = m_r = 20$. We used the same number of principal components for classification. This choice was verified experimentally as using more than 20 discriminant components did not improve the results.

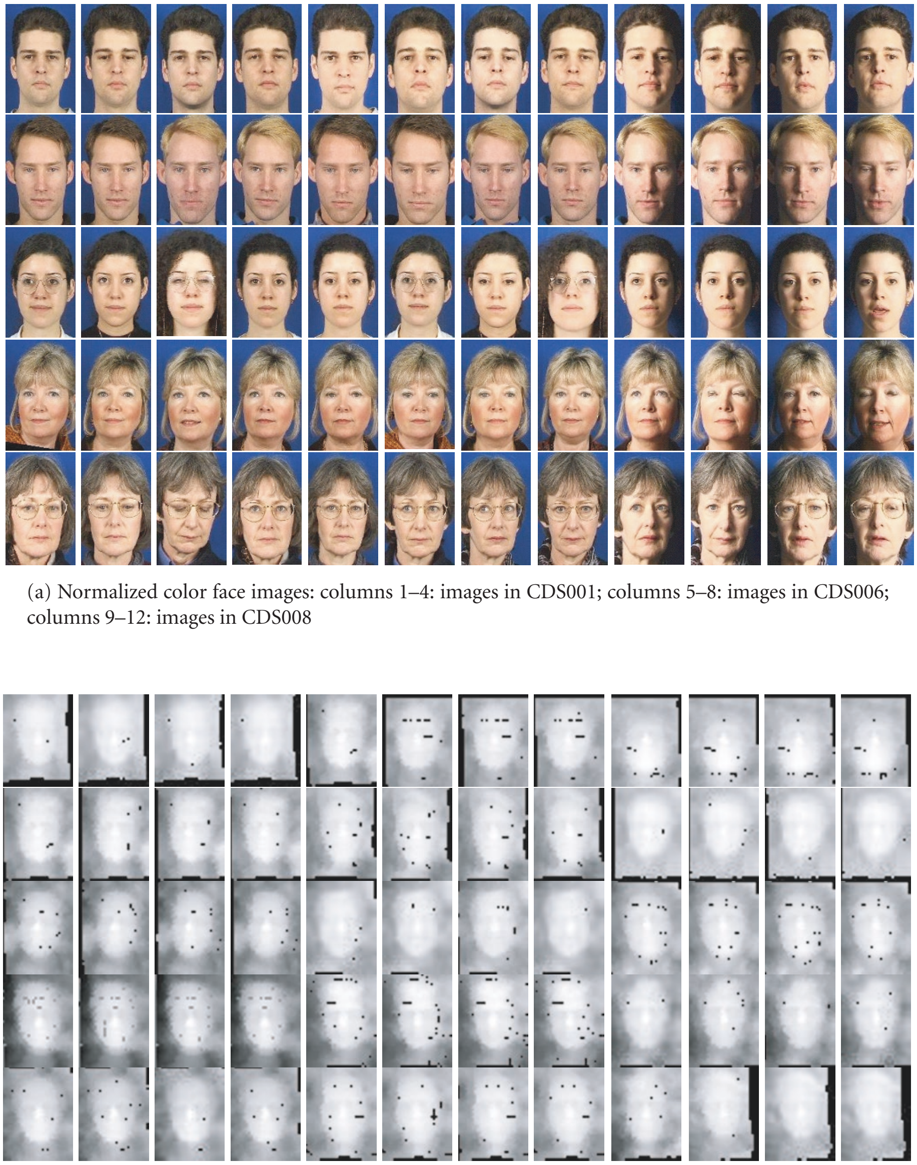
3.2. The complexity analysis

We can see that the most expensive steps in Algorithm 1 are in lines 3, 6, 9. The comparisons of computational complexity of Fisherfaces, Ye's 2DLDA, Yang's 2DLDA, and the proposed 2DLDA are listed in Table 1.

The computational complexity of Fisherfaces increases cubically with the size of the training sample size. The computational complexity of B2LDA is the same as Yang's method, and both of them depend on the image size. However, it is higher than Ye's method.

4. FUSION OF APPEARANCE AND DEPTH/DISPARITY

We aim to improve the recognition rate by combining appearance and depth information. The matter of how to fuse two or more sources of information is crucial to the



(a) Normalized color face images: columns 1–4: images in CDS001; columns 5–8: images in CDS006; columns 9–12: images in CDS008

(b) Normalized depth images corresponding to (a)

Figure 3: Normalized 2D and 3D face images in the XM2VTS database: (a) appearance images, (b) depth images.

performance of the system. The criterion for this kind of combination is to fully make use of the advantages of the two sources of information to optimize the discriminant power of the whole system. The degree to which the results improve performance is dependent on the degree of correlation among individual decisions. Fusion of decisions with low mutual correlation can dramatically improve the performance. There is a rich literature [2, 36] on fusing multiple modals for identity verification, for example, combining voice and fingerprints, voice and face biometrics [37], and visible and thermal imagery [38]. The fusion can be done at the feature level, matching score level, or decision level. In this paper, we are interested in the fusion at the matching score level. There are some ways of combining different matching scores to achieve the best decision, for example, by majority vote, sum rule, multiplication rule, median rule, minimum rule, and average rule. It is known that sum and multiplication rules provide general plausible results. In this paper, we use the weighted sum rule to fuse appearance and depth information. Our rationale is that appearance information and depth information are quite highly uncorrelated. This is clear since depth data yields surface or terrain of the observed scene while the appearance information records the texture of the surface. Though the normals to the surface affects the reflectivity of light and thereby the surface illumination, this has minimal effect on the surface texture. Therefore, a certain linear combination will be sufficient to extract a good set of features for the purpose of recognition. Nevertheless, there will be a small correlation between them in the sense that the general terrain of the face (i.e., depth map) has

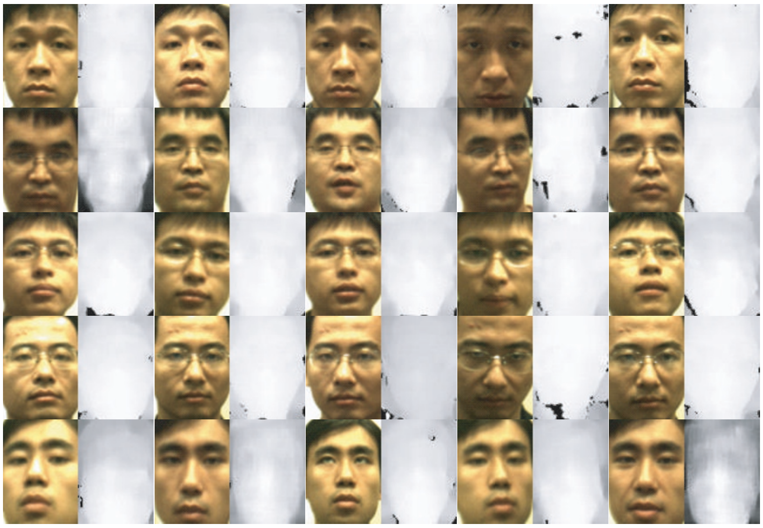


Figure 4: Normalized appearance and disparity images captured by the Mega-D stereo head.

Figure 5: Normalized appearance images captured by a Mega-D stereo head.

a bearing on the shading of the appearance image. We investigate the complete range of linear combinations to reveal the interplay between these two paradigms.

The linear combination of the appearance and depth in our approach can be explained using Figure 6. We optimize the combination of the depth and intensity discriminant Euclidean distances by minimizing the weighted sum of two discriminant Euclidean distances.

Given the gallery of depth images and appearance images, they are trained, respectively, by B2DLDA. The Euclidean distance between the test image and the templates are measured as the inverse of similarity score to decide whose face it is. Assuming the eigenvectors of face image k and i are represented as $\mathbf{v}_k$ and $\mathbf{v}_i$, respectively,

$$S^{-1}(k,i) = \operatorname{dist}(k,i) = \left\|\mathbf{v}_k - \mathbf{v}_i\right\|_2. \tag{6}$$

A probe face, F_T, is identified as a face, F_L, of the gallery if the sum of the weighted similarity scores (appearance and depth) from F_T to F_L is the maximum among such sums from F_T to all the faces in the gallery. This can be expressed as

$$\max_{\text{gallery}} \left\{w_1 S_{2\text{D}} + \left(1 - w_1\right) S_{3\text{D}}\right\}, \tag{7}$$

where $S_{2\text{D}}$ and $S_{3\text{D}}$ are the similarity scores for intensity and depth images, respectively. The weight w_1 is determined to be optimal through experiments. In general, a higher value of $(1 - w_1)$ reflects the fact that the variance of the discriminant Euclidean distance of a depth map is relatively smaller than the one for the corresponding appearance face image.

5. EXPERIMENTAL RESULTS

The face recognition experiments are performed on the XM2VTS database and the Mega-D database, respectively, to verify the improvement of the recognition rate by combining 2D and 3D information. We assess the accuracy and efficiency of B2DLDA and compare it with Ye's 2DLDA [27], Yang's 2DLDA [29], Fisherfaces [34], and Eigenfaces [3–5].

```
Input: $A_1, A_2, \ldots, A_n, m_l, m_r$     % $A_i$ are the $n$ images, and $m_l$ and $m_r$ are the number of the
                                             % discriminant components of left and right B2DLDA transform
Output: $\mathbf{W}_l, \mathbf{W}_r, B_{l1}, B_{l2}, \ldots, B_{ln}, B_{r1}, B_{r2}, \ldots, B_{rn}$     % $\mathbf{W}_l$ and $\mathbf{W}_r$ are the left and right
                                             % transformation matrix respectively by
                                             % B2DLDA; $B_{li}$ and $B_{ri}$ are the reduced
                                             % representations of $A_i$ by $\mathbf{W}_l$ and $\mathbf{W}_r$
                                             % respectively
(1) Compute the mean, $\mathbf{M}_i$, of the $i$th class of each $i$
(2) Compute the global mean, $\mathbf{M}$, of $\{A_i\}$, $i = 1, 2, \ldots, n$
(3) Find $\mathbf{S}_{bl}$ and $\mathbf{S}_{wl}$, $\mathbf{S}_{bl} = \sum_{i=1}^{C} C_i \bullet (\mathbf{M}_i - \mathbf{M})^T (\mathbf{M}_i - \mathbf{M})$, $\mathbf{S}_{wl} = \sum_{i=1}^{C} \sum_{j=1}^{C_i} (\mathbf{X}_i^j - \mathbf{M}_\mathbf{i})^T (\mathbf{X}_i^j - \mathbf{M}_i)$
                                             % C is the number of the classes; $C_i$ is the
                                             % number of the samples in the $i$th class
(4) Compute the first $m_l$ eigenvectors $\{\phi_i^L\}_{i=1}^{m_l}$ of $\mathbf{S}_{\mathbf{w}l}^{-\mathbf{1}} \mathbf{S}_{\mathbf{b}l}$
(5) $\mathbf{W}_l \leftarrow [\phi_1^L, \phi_2^L, \ldots, \phi_{m_l}^L]$
(6) Find $\mathbf{S}_{br}$ and $\mathbf{S}_{wr}$, $\mathbf{S}_{br} = \sum_{i=1}^{C} C_i \bullet (\mathbf{M}_i - \mathbf{M})(\mathbf{M}_i - \mathbf{M})$, $\mathbf{S}_{wr} = \sum_{i=1}^{C} \sum_{j=1}^{C_i} (\mathbf{X}_i^j - \mathbf{M}_i)(\mathbf{X}_i^j - \mathbf{M})$
(7) Compute the first $m$ eigenvectors $\{\phi_i^R\}_{i=1}^{m_r}$ of $\mathbf{S}_{\mathbf{w}r}^{-\mathbf{1}} \mathbf{S}_{\mathbf{b}r}$
(8) $\mathbf{W}_r \leftarrow [\phi_1^R, \phi_2^R, \ldots, \phi_{m_r}^R]$
(9) $B_{li} = A_i W_l$, $i = 1, \ldots, n$
    $B_{ri} = A_i' W_r$, $i = 1, \ldots, n$
(10) Return $\mathbf{W}_l, \mathbf{W}_r, B_{li}, B_{ri}$, $i = 1, \ldots, n$
```

ALGORITHM 1: Algorithm B2DLDA $(\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_n, m_l, m_r)$.

TABLE 1: The comparisons of computational complexity of Fisherfaces [39], Ye's 2DLDA [27], Yang's 2D LDA [29], and the proposed 2DLDA [28]. M is the total number of the train samples; r, c are the numbers of the rows and columns of the original image, $\mathbf{A}$, respectively; $l = \max(r, c)$.

Method	Fisherfaces [39]	Ye [27]	Yang [29]	B2DLDA [28]
Computation complexity	$O(M^3)$	$O(rc)$	$O(l^3)$	$O(l^3)$

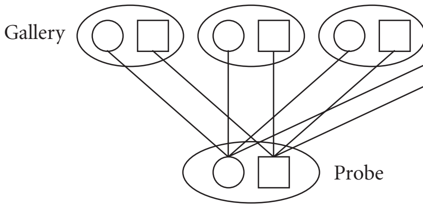


FIGURE 6: Combination of appearance (circle) and depth (square) information.

5.1. Experiment on the XM2VTS database

The XM2VTS consists of the frontal and profile views of 295 subjects. We used the frontal views in the XM2VTS database (CDS001, CDS006, and CDS008 darkened frontal view). CDS001 dataset contains one frontal view for each of the 295 subjects and each of the four sessions. This image was taken at the beginning of the head rotation shot. So there are a total of 1180 color images, each with a resolution of 720 × 576 pixels. CDS006 dataset contains one frontal view for each of the 295 subjects and each of the four sessions. This image was taken from the middle of the head rotation shot when the subject had returned his/her head to the middle. They are different from those contained in CDS001. There are a total of 1180 color images. The images are at a resolution of 720 × 576 pixels. CDS008 contains four frontal views for each of the 295 subjects taken from the final session. In two of the images, the studio light illuminating the left side of the face was turned off. In the other two images, the light illuminating the right side of the face was turned off. There are a total of 1180 color images. The images are at a resolution of 720 × 576 pixels. We used the 3D VRML model (CDS005) of the XM2VTSDB to generate 3D depth images corresponding to the appearance images mentioned above. The models were obtained with a high-precision 3D stereo camera developed by the Turing Institute [24]. The models were then converted from their proprietary format into VRML.

Therefore, a total of 3540 pairs of frontal views (appearance and depth pair) of 295 subjects in X2MVTS database are used. There are 12 pairs of images for each subject. We pick randomly any two of them for the learning gallery while the remainder ten pairs per subject are used as probes. The average recognition rate was obtained over 66 random runs. As only two pairs of face images are used for training, it is clear that LDA will face the SSS problem because the number of the training samples is much less than the dimension of the covariance matrix in LDA. Using two images per person for training could be insufficient for LDA-based or

Table 2: The mean recognition rates (%) on the XM2VTS database versus w_1.

w_1	B2DFDA [28]	Ye's 2D LDA [27]	Yang's 2DLDA [29]	Fisherfaces [39]	Eigenfaces [3–5]
0.0	91.63	90.88	89.88	87.86	84.86
0.1	97.88	96.00	95.00	94.80	93.10
0.2	98.66	97.44	96.44	96.10	94.50
0.3	97.88	96.66	95.66	95.20	92.52
0.4	97.81	96.01	95.01	94.80	91.80
0.5	95.75	94.38	93.92	93.81	90.90
0.6	94.19	93.61	93.01	92.80	90.14
0.7	94.19	93.14	92.14	91.40	89.40
0.8	91.84	91.58	90.58	88.50	87.51
0.9	88.72	88.84	87.84	86.90	85.90
1.0	81.69	80.63	78.63	76.70	75.71

Table 3: The mean recognition rates (%) on the Mega-D database versus w_1.

w_1	B2DFDA [28]	Ye's 2D LDA [27]	Yang's 2DLDA [29]	Fisherfaces [39]	Eigenfaces [3–5]
0.0	90.63	89.87	88.78	89.80	83.82
0.1	97.56	95.44	94.41	94.17	92.51
0.2	96.88	95.00	94.02	93.78	92.13
0.3	96.82	94.62	93.60	93.23	90.51
0.4	95.31	94.01	93.04	92.81	89.78
0.5	93.73	92.81	92.92	90.84	88.92
0.6	92.18	92.01	92.00	90.30	88.17
0.7	92.10	91.14	90.03	88.39	87.41
0.8	89.83	89.60	88.42	86.49	85.53
0.9	86.71	86.79	85.70	85.91	83.91
1.0	79.69	78.58	74.61	78.72	73.73

2DLDA-based face recognition to be optimal. In this paper, we want to show that our proposed method can solve the SSS problem where the number of training sample is less. Therefore, we used the least images per person, that is two, for training. It is fair to compare our algorithm with others because we used the same training set for this comparison. Thus our algorithm is useful in situations where there are only limited numbers of samples for training.

Using the training gallery and probe described above, the evaluations of the recognition algorithms on B2DLDA, Ye's 2DLDA, Yang's 2DLDA, Fisherfaces, and eigenfaces have been done. This includes the recognition evaluation when the weight w_1 in (7) is varied from 0 (which corresponds to depth alone) to 1 (which corresponds to intensity alone) with a step increment of 0.1. Assuming we have N training samples of C subjects (classes), the recognition rates on the XM2VTS database versus the weight w_1 are given in Table 2 or Figure 7. B2DFDA is compared with

(1) Ye's 2D LDA [27],

(2) Yang's 2DLDA [29],

(3) Fisherfaces (PCA plus LDA) [39],

(4) Eigenfaces [3–5].

By fusing the appearance and the depth, the highest recognition rate, 98.66%, happens at $w_1 = 0.2$ for B2DLDA as shown in Table 2. This supports our hypothesis that the combined method outperforms the individual appearance or depth. The results in Table 2 also verified that the proposed B2DLDA outperforms Ye's 2DLDA. Ye reported their method can get the results similar to optimal LDA (PCA + LDA). Here, this can be observed in our results.

5.2. Experiment on stereo vision system

Differing from the existing 3D or 2D + 3D face recognition systems, we used a passive stereovision to get 3D information. A database, called Mega-D, was built with SRI stereo head engine. (We have described the Mega-D database in Section 3.2.) In this section, we evaluate the algorithms on the Mega-D database. We will show that we can get comparable results with the database where 3D information is obtained by an active stereo engine, that is, the XM2VTS database.

A total of 1272 frontal views of 106 subjects in the Mega-D database are used. There are 12 pairs of images for each subject. We use any two randomly selected pairs of them

Table 4: The computation time of Fisherfaces [39], Ye's 2DLDA [27], Yang's 2DLDA [29], and the proposed 2DLDA [28].

Method	Fisherfaces [39]	Ye's 2DLDA [27]	Yang's 2DLDA [29]	B2DLDA [28]
CPU time (s)	75	12.5	24	26

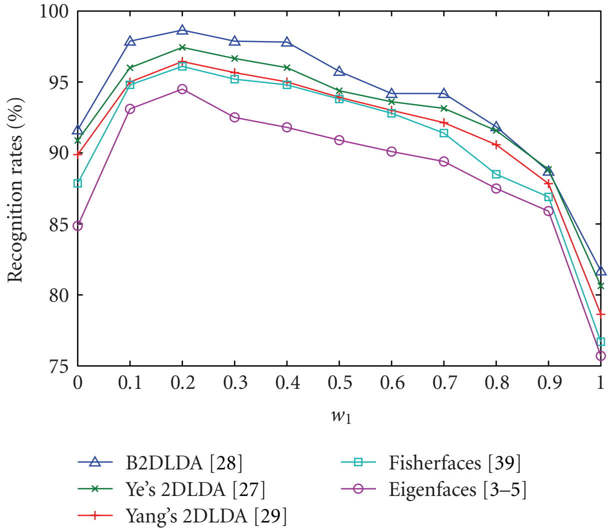


Figure 7: Recognition performance on the XM2VTS database versus w_1. $w_1 = 0$ corresponds to 3D alone, $w_1 = 1$ corresponds to 2D alone.

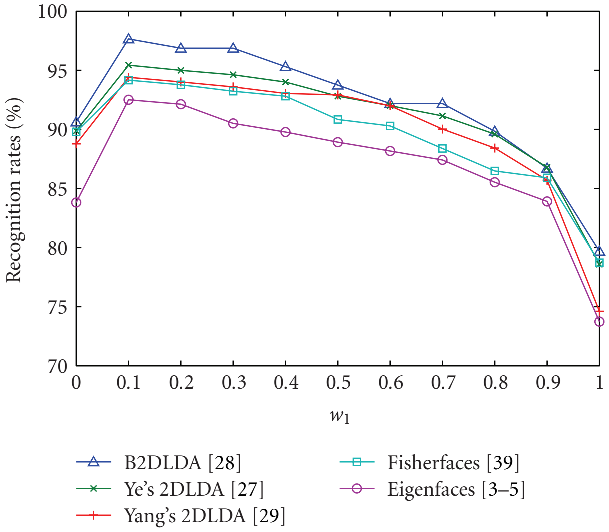


Figure 8: Recognition performance on the Maga-D database versus w_1. $w_1 = 0$ corresponds to 3D alone, $w_1 = 1$ corresponds to 2D alone.

for the learning gallery while the remainder ten are used as probes. Using the gallery and probe described above, the evaluations of the recognition algorithms (2D FDA and 1D FDA) have been done, include the recognition when the weight w_1 in (7) varies from 0 (which corresponds to depth alone) to 1 (which corresponds to intensity alone) with a step increment of 0.1. Similar to the experiments on the XM2VTS database, a total of 66 random trials were performed and the mean of these trails is used in the final recognition result. The recognition rates on the Mega-D database versus the weight w_1 are given in Table 3 or Figure 8.

Similar to the results on the XM2VTS database, the results supported our hypothesis that the combined method outperforms the individual appearance or depth. It also verified that the proposed B2DLDA outperforms Ye's 2DLDA. Ye's method [27] can get the results similar to Fisherfaces. This experiment also illustrated the viability of using passive stereovision for face recognition.

We implemented the algorithms in Visual C++ on a P3 3.4Ghz 1GB PC. The computation time is listed in Table 4.

We can see in Table 4 that our method's processing time costs twice more than that for Ye's method (only one iteration).

6. CONCLUSIONS

In this paper, a novel fusion of appearance image and passive stereo depth is proposed to improve face recognition rates. Different from the existing 3D or 2D + 3D face recognition that used active stereo method to obtain 3D information, comparable results have been obtained in this paper on both the XM2VTS and a large database collected with the passive Mega-D stereo engine. We investigated the complete range of linear combinations to reveal the interplay between these two paradigms. The improvement of the face recognition rate using this combination has been verified. The recognition rate by the combination is better than either appearance alone or depth alone. In order to overcome the small sample size problem in LDA, a bilateral two-dimensional linear discriminant analysis (B2DLDA) is proposed in this paper to extract the image features. The experimental results show that B2DLDA outperforms the existing 2DLDA approaches.

REFERENCES

[1] W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face recognition: a literature survey," *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, 2003.

[2] R. Brunelli and D. Falavigna, "Person identification using multiple cues," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 10, pp. 955–966, 1995.

[3] K. Chang, K. Bowyer, and P. Flynn, "Face recognition using 2D and 3D facial data," in *Proceedings of ACM Workshop on Multimodal User Authentication*, pp. 25–32, Santa Barbara, Calif, USA, December 2003.

[4] K. I. Chang, K. W. Bowyer, and P. J. Flynn, "An evaluation of multimodal 2D+3D face biometrics," *IEEE Transactions on*

Pattern Analysis and Machine Intelligence, vol. 27, no. 4, pp. 619–624, 2005.

[5] F. Tsalakanidou, D. Tzovaras, and M. G. Strintzis, “Use of depth and colour eigenfaces for face recognition,” *Pattern Recognition Letters*, vol. 24, no. 9-10, pp. 1427–1435, 2003.

[6] J.-G. Wang, H. Kong, and R. Venkateswarlu, “Improving face recognition performance by combining colour and depth fisherfaces,” in *Proceedings of 6th Asian Conference on Computer Vision*, pp. 126–131, Jeju, Korea, January 2004.

[7] J.-G. Wang, K.-A. Toh, and R. Venkateswarlu, “Fusion of appearance and depth information for face recognition,” in *Proceedings of the 5th International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA ’05)*, pp. 919–928, Rye Brook, NY, USA, July 2005.

[8] K. W. Bowyer, K. Chang, and P. Flynn, “A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition,” *Computer Vision and Image Understanding*, vol. 101, no. 1, pp. 1–15, 2006.

[9] N. Mavridis, F. Tsalakanidou, D. Pantazis, S. Malassiotis, and M. G. Strintzis, “The HISCORE face recognition application: affordable desktop face recognition based on a novel 3D camera,” in *Proceedings of International Conference on Augmented, Virtual Environments and Three Dimensional Imaging (ICAV3D ’01)*, pp. 157–160, Mykonos, Greece, May-June 2001.

[10] C. Beumier and M. Acheroy, “Automatic face authentication from 3D surface,” in *Proceedings of British Machine Vision Conference (BMVC ’98)*, pp. 449–458, Southampton, UK, September 1998.

[11] G. G. Gordon, “Face recognition based on depth maps and surface curvature,” in *Geometric Methods in Computer Vision*, vol. 1570 of *Proceedings of SPIE*, pp. 234–247, San Diego, Calif, USA, July 1991.

[12] X. Lu and A. K. Jain, “Deformation analysis for 3D face matching,” in *Proceedings of the 7th IEEE Workshop on Applications of Computer Vision / IEEE Workshop on Motion and Video Computing (WACV/MOTION ’05)*, pp. 99–104, Breckenridge, Colo, USA, January 2005.

[13] P. J. Philips, P. Grother, R. J. Micheals, D. M. Blackburn, E. Tabassi, and M. Bone, “Face recognition vendor test 2002,” Tech. Rep. NIST IR 6965, National Institute of Standards and Technology, Gaithersburg, Md, USA, March 2003.

[14] S. A. Rizvi, P. J. Phillips, and H. Moon, “The FERET verification testing protocol for face recognition algorithms,” Tech. Rep. NIST IR 6281, National Institute of Standards and Technology, Gaithersburg, Md, USA, October 1998.

[15] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, “The FERET evaluation methodology for face-recognition algorithms,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090–1104, 2000.

[16] K. I. Chang, K. W. Bowyer, P. J. Flynn, and X. Chen, “Multibiometrics using facial appearance, shape and temperature,” in *Proceedings of the 6th IEEE International Conference on Automatic Face and Gesture Recognition (FGR ’04)*, pp. 43–48, Seoul, Korea, May 2004.

[17] C. Beumier and M. Acheroy, “Face verification from 3D and grey level clues,” *Pattern Recognition Letters*, vol. 22, no. 12, pp. 1321–1329, 2001.

[18] J. C. Lee and E. E. Milios, “Matching range images of human faces,” in *Proceedings of the 3rd International Conference on Computer Vision (ICCV ’90)*, pp. 722–726, Osaka, Japan, December 1990.

[19] Y. Yacoob and L. S. Davis, “Labeling of human face components from range data,” *CVGIP: Image Understanding*, vol. 60, no. 2, pp. 168–178, 1994.

[20] C.-S. Chua, F. Han, and Y. K. Ho, “3D human face recognition using point signature,” in *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG ’00)*, pp. 233–238, Grenoble, France, March 2000.

[21] V. Blanz and T. Vetter, “Face recognition based on fitting a 3D morphable model,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1063–1074, 2003.

[22] V. Blanz and T. Vetter, “A morphable model for the synthesis of 3D faces,” in *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH ’99)*, pp. 187–194, Los Angeles, Calif, USA, August 1999.

[23] G. Pan, Y. Wu, and Z. Wu, “Investigating profile extracted from range data for 3D face recognition,” in *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*, vol. 2, pp. 1396–1399, Washington, DC, USA, October 2003.

[24] C. W. Urquhart, J. P. McDonald, J. P. Siebert, and R. J. Fryer, “Active animate stereo vision,” in *Proceedings of the 4th British Machine Vision Conference*, pp. 75–84, University of Surrey, Guildford, UK, September 1993.

[25] K. Liu, Y.-Q. Cheng, and J.-Y. Yang, “Algebraic feature extraction for image recognition based on an optimal discriminant criterion,” *Pattern Recognition*, vol. 26, no. 6, pp. 903–911, 1993.

[26] J. Yang, D. Zhang, A. F. Frangi, and J.-Y. Yang, “Two-dimensional PCA: a new approach to appearance-based face representation and recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 1, pp. 131–137, 2004.

[27] J. Ye, R. Janardan, and Q. Li, “Two-dimensional linear discriminant analysis,” in *Proceedings of Neural Information Processing Systems (NIPS ’04)*, pp. 1569–1576, Vancouver, British Columbia, Canada, December 2004.

[28] H. Kong, L. Wang, E. K. Teoh, J.-G. Wang, and R. Venkateswarlu, “A framework of 2D fisher discriminant analysis: application to face recognition with small number of training samples,” in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR ’05)*, vol. 2, pp. 1083–1088, San Diego, Calif, USA, June 2005.

[29] J. Yang, D. Zhang, X. Yong, and J.-Y. Yang, “Two-dimensional discriminant transform for face recognition,” *Pattern Recognition*, vol. 38, no. 7, pp. 1125–1129, 2005.

[30] M. Visani, C. Garcia, and J.-M. Jolion, “Two-dimensional-oriented linear discriminant analysis for face recognition,” in *Proceedings of the International Conference on Computer Vision and Graphics (ICCVG ’04)*, pp. 1008–1017, Warsaw, Poland, September 2004.

[31] Videre Design, “MEGA-D Megapixel Digital Stereo Head,” http://users.rcn.com/mclaughl.dnai/sthmdcs.htm.

[32] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, “XM2VTSDB: the extended M2VTS database,” in *Proceedings of International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA ’99)*, pp. 72–77, Washington, DC, USA, March 1999.

[33] E. E. Catmull, *A subdivision algorithm for computer display of curved surfaces*, Ph.D. thesis, Department of Computer Science, University of Utah, Salt Lake City, Utah, USA, 1974.

[34] J.-G. Wang and E. Sung, “Frontal-view face detection and facial feature extraction using color and morphological operations,” *Pattern Recognition Letters*, vol. 20, no. 10, pp. 1053–1068, 1999.

[35] R. I. Jenrich, “Stepwise discriminant analysis,” in *Statistical Methods for Digital Computers*, K. Enslein, A. Ralston, and H.

S. Wilf, Eds., pp. 76–95, John Wiley & Sons, New York, NY, USA, 1977.

[36] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

[37] T. Choudhury, B. Clarkson, T. Jebara, and A. Pentland, "Multimodal person recognition using unconstrained audio and video," in *Proceedings of the 2nd International Conference on Audio- and Video-Based Person Authentication (AVBPA '99)*, pp. 176–181, Washington, DC, USA, March 1999.

[38] D. A. Socolinsky, A. Selinger, and J. D. Neuheisel, "Face recognition with visible and thermal infrared imagery," *Computer Vision and Image Understanding*, vol. 91, no. 1-2, pp. 72–114, 2003.

[39] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.