INVITED PAPER

Face Recognition Using 2-D, 3-D, and Infrared: Is Multimodal Better Than Multisample?

Use of multiple images of a face appears to improve recognition accuracy regardless of the type of images that are taken.

By Kevin W. Bowyer, *Fellow IEEE*, Kyong I. Chang, Patrick J. Flynn, *Senior Member IEEE*, and Xin Chen

ABSTRACT | This work examines face recognition using normal intensity images, infrared images, three-dimensional shape, and combinations of these. We compare the performance improvement obtained by combining three-dimensional or infrared with normal intensity images (a "multimodal" approach) to the performance improvement obtained by using multiple intensity images (a "multisample" approach). Combining results from different types of imagery gives significantly higher recognition rates than are obtained by using a single intensity image. However, significantly higher recognition rates are also obtained by combining results from multiple intensity images. Overall, initial results indicate that, using an "eigen-face" recognition algorithm and weighted score fusion, multisample techniques can result in a performance increase comparable to that of multimodal techniques.

KEYWORDS | Biometrics; face recognition; information fusion; infrared; multimodal; three-dimensional

Manuscript received September 29, 2005; revised June 29, 2006. This work was supported in part by the National Science Foundation under Grant CNS-0130839, in part by the Central Intelligence Agency, in part by the National Geo-Spatial Intelligence Agency, in part by UNISYS Corp., and in part by the U.S. Department of Justice under Grant 2005-DD-BX-1224 and Grant 2005-DD-CX-K078.

K. W. Bowyer and **P. J. Flynn** are with the Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556 USA (e-mail: kwb@cse.nd.edu; flynn@cse.nd.edu).

K. I. Chang is with the Philips Medical Systems—Ultrasound, Bothell, WA 98041 USA (e-mail: jin.chang@philips.com).

X. Chen is with the Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556 USA and also with Navteq, Chicago, IL 60654 USA (e-mail: xchen2@cse.nd.edu; xin.chen@navteq.com).

Digital Object Identifier: 10.1109/JPROC.2006.885134

I. INTRODUCTION

The vast majority of face recognition research assumes that an attempt to recognize a person is made using a single intensity image of the type taken by standard cameras. A recent broad survey of such face recognition research is given by Zhao [1]. However, evaluations such as the 2002 Face Recognition Vendor Test (FRVT) [2] have shown that the accuracy of face recognition is not yet sufficient for the more demanding applications. Complications that arise from variations in pose, lighting, and facial expression are among the various factors that contribute to decreased performance. This has led some researchers to investigate the use of three-dimensional (3-D) shape information for face recognition [3]–[8]. Some motivations for using 3-D shape are that shape is defined independent of lighting, and that acquiring 3-D shape should allow for accurate pose correction. Other researchers have investigated the use of infrared (IR) images for face recognition [9]–[12]. A major motivation for using IR images is that they are relatively unaffected by changes in lighting. Examples of these different types of face image appear in Fig. 1.

In addition to exploring the use of 3-D shape and IR images as alternatives to normal intensity images, researchers have developed approaches to combining the recognition results from either a 3-D shape model or an IR image with the results from an intensity image. These approaches that combine different types of information for face recognition are commonly, although perhaps imprecisely, referred to as *multimodal*. A general meaning of the term multimodal is simply that different properties are sensed. A more specific possible meaning is that different sensors are used in acquiring the data. Some sensors that give integrated 3-D shape and intensity image

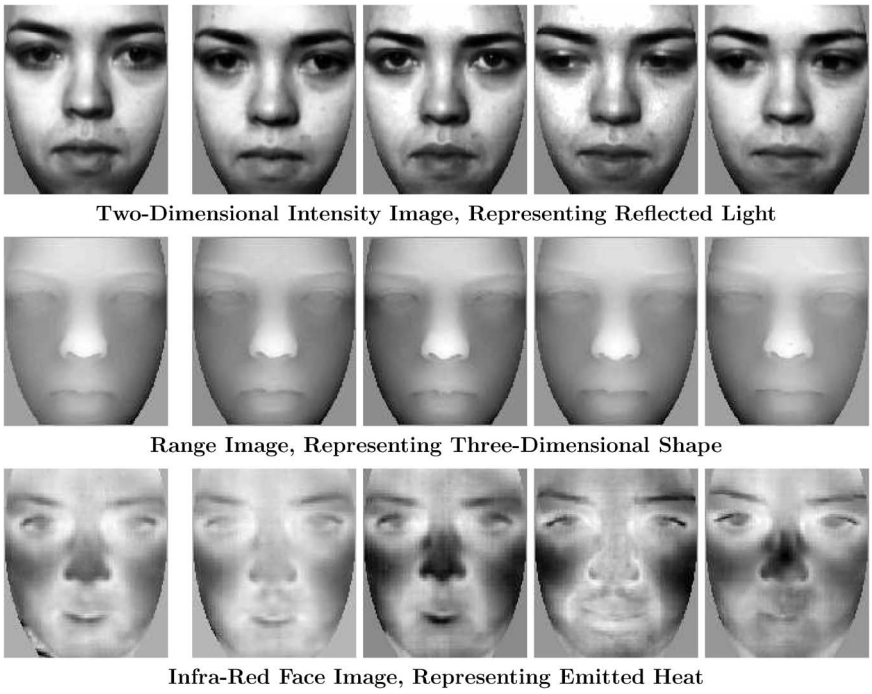

Fig. 1. ***Examples of intensity images, range images representing 3-D shape, and IR images. All images are of the same person. The images along a row are from different acquisition sessions, with one or more weeks of elapsed time between sessions. The images in a column are taken in the same acquisition session. Note the variations between the images in the different modalities and across time.***

data might not be considered multimodal in this more restricted sense.

Publications in the area of multimodal face recognition have uniformly reported that the combination of 3-D shape with intensity images, or the combination of IR images with intensity images, improves performance over the intensity image alone. However, the comparison has generally been made between a system that employs: 1) one normal intensity image plus one additional 3-D shape or IR image and 2) a system that employs one normal intensity image. It has been noted [3] that this type of comparison is biased in favor of seeing a greater recognition accuracy from the multimodal approach, and that a more fair comparison would be between using: 1) one normal intensity image plus one additional 3-D shape or IR image versus 2) two different intensity images of the person. The use of more than one normal intensity image to represent a person will be referred to in the sequel as a *multisample* approach.

One exception to the above general characterization of previous work is that of Socolinsky *et al.* [11] dealing with face recognition from infrared and intensity images. They show results of several experiments in which a person is represented by multiple images in the gallery, and then verified from a single image. For multisample intensity, multisample infrared or multimodal, the equal-error-rate decreases with additional images used to represent a person, up to a six-image maximum. An overall similarity with the work reported here is that multisample intensity with enough samples can achieve performance similar to that of multimodal fusion of intensity and infrared.

This paper considers issues of multimodal face recognition involving the combination of infrared and 3-D with normal intensity images, and the comparison to multisample recognition using normal intensity images. We explore these issues using different instantiations of the same core "eigen-face" recognition algorithm with each type of image. We use an experimental dataset which has intensity, infrared and 3-D images of the same persons. Our results suggest that, at least for the recognition algorithm and score fusion method used here, a multisample approach using traditional intensity images can be competitive with multimodal face recognition using 3-D or infrared in combination with traditional intensity images.

II. BACKGROUND: CATEGORIES OF BIOMETRIC FUSION

Before going into the details of the experiments reported here, it may be useful to place them in the more general context of biometrics and biometric fusion. (Some of the discussion in this section follows that in [13].) The simple approach to biometrics is to sense a single sample (image)

of a biometric source (body part) from a person and then process that to obtain a recognition result. The vast majority of face recognition research has implicitly assumed this framework. The term "multimodal biometrics" is used in the literature with various meanings. Perhaps the least ambiguous example of multimodal biometrics would be two different sources on a person, say face and fingerprint, sensed by different sensors. Two different properties, say infrared and reflected light, of the same biometric source, say the face, would be another unambiguous example of multimodal. An ambiguous example might be two different biometric sources, say face and ear [14], imaged by the same sensor. Another ambiguous example might be two different properties, say 3-D shape and reflected light, of the same source, say face, sensed by the same sensor. An expansive view would consider all of these variations as "multimodal," and consider "multibiometric" as an equivalent term.

A. Multialgorithm Biometrics

One step beyond a simple biometric is what we might call a multialgorithm approach. This approach still employs a single sensor, and acquires a single biometric sample. Two or more different algorithms process the single sample, and the individual results are fused to obtain an overall recognition result.

The multialgorithm approach would seem to be attractive, both from an application point of view and from a research point of view. From an application perspective, it appears to minimize sensor and sensing cost, since there is only one sensor and only one sample sensed in order to obtain a recognition result. Relatively little work has been done in this area. As one example, the Supplemental Report to the 2002 Face Recognition Vendor Test documented increased performance in two-dimensional (2-D) face recognition by combining the results of different commercial recognition systems [15]. More recently, Gokberk *et al.* have looked at combining multiple algorithms for 3-D face recognition [16]. Xu *et al.* [17] have also combined different algorithmic approaches for 3-D face recognition.

A variation of the multialgorithm approach builds an ensemble of multiple instances the same basic type of algorithm, with intentional random variation between instances. For example, Chawla *et al.* used the random subspaces concept to create an ensemble and obtain improved recognition rates from an eigen-face algorithm [18], [19].

B. Multisample Biometrics

Another approach might be called "multisample" or "multi-instance." Multiple samples of the same biometric are sensed, the same algorithm processes each of the samples, and the individual results are fused to obtain an overall recognition result. Multisample approaches were investigated in the 2002 FRVT [2] and more recently in the Face Recognition Grand Challenge [20]. In this paper, we use a multisample approach with 2-D face images for comparison against a multimodal combination of 2-D, 3-D, and infrared imagery.

A multisample approach has advantages and disadvantages in comparison to the multialgorithm approach. The use of multiple samples may overcome poor performance due to one sample that has unfortunate properties. For example, a person might be blinking in one face image, and this might present problems for the recognition algorithm; if multiple samples in time are used, it is unlikely that the person is blinking in all of them. However, the acquisition of multiple samples requires either multiple copies of the sensor, or that the user be available for sensing over a longer period of time. When compared to multialgorithm approaches, multisample techniques would seem to require either greater expense for sensors, greater cooperation from the user, or a combination of both.

C. Multimodal Biometrics

We will discuss multimodal approaches in three categories. We will call these "orthogonal," "independent," and "collaborative." These are not standard terms, but are perhaps useful because they point up differences in the fusion of results from the individual modes.

One common category of multimodal biometrics can be called "orthogonal." By "orthogonal" we mean the use of biometric sources that involve different parts of the body. An example would be face and fingerprint matching used together. The most publicly visible use of multimodal biometrics is perhaps the (prospective) use of face and fingerprint planned in the "US VISIT" program [21]. In a speech about this program in 2003, an official actually mentioned face, fingerprint, and iris—"We'll do so through a minimum of two biometric identifiers—initially, fingerprints and photographs; later, as the technology is perfected, additional forms such as facial recognition or iris scans may be used as well" [22].

In this category, there appears to be little or no opportunity for interaction between the individual biometrics. For instance, it is difficult to see how the intermediate processing of either face or fingerprint could be used to help the other. As a result, the individual biometrics are combined at the "decision level" or the "score level." In decision-level fusion, a recognition decision is made for each individual biometric, and the individual decisions vote to obtain the overall decision. In score-level fusion, a matching score is obtained for each individual biometric, and the scores are combined to obtain the multimodal decision. Researchers in multimodal biometrics have generally found that score-level fusion performs at least as well as decision-level fusion. In general, score-level fusion must involve a method to normalize the scores from the individual biometrics, followed

by a method to combine the scores. In the multimodal results presented in this paper, we use score-level fusion.

Another category of approach to multimodal biometrics might be called "independent." By "independent" we mean to indicate that the individual biometrics are processed independently of each other. It would seem that orthogonal biometrics are processed independently by necessity. But when the biometric source is the same and different properties are sensed, then the processing may be independent, but there is at least the potential for gains in performance through collaborative processing. As with most multimodal face recognition research to date, the results that we report in this paper fall into the category of independent multimodal biometrics.

A less common approach to multimodal biometrics might be called "collaborative." By "collaborative" we refer to interaction between the intermediate results of processing the individual biometrics. There are some examples in the literature of what might be called weakly collaborative approaches. Husken *et al.* [23] describe an approach to multimodal 2-D + 3-D face recognition that locates the feature points (e.g., eyes) on the face in the 2-D image, and then transfers these locations over to the registered 3-D data to process the features there. Socolinsky *et al.* [11] follow a similar approach in their multimodal infrared and visible-light face recognition. Their sensor is able to obtain registered images from the two modes, and they find the eye location in the visible-light image and transfer the locations over to the infrared image. These approaches are collaborative in the sense that intermediate results of processing in one modality are used to assist the processing in the other modality. But the degree of collaboration in these examples is not extensive, and is only in the direction of 2-D to the other modality.

One can imagine that much more extensive collaborative processing might be possible. Consider the example of 2-D + 3-D face recognition. Artifacts occur in both types of images, and it may be possible to exploit the ease of finding a certain type of artifact in one mode to improve the reliability of processing the other mode. For example, if specular highlights are found in the 2-D face image, this might inform the processing of the 3-D shape of the face, since specular highlights in 2-D often result in artifacts in the 3-D image. Also, once something of the general shape of the face is known, it may be possible to use this to consistently interpret regions of the 2-D image as affected by shadows. In this way, the intermediate results of processing each modality might be used to improve the reliability and accuracy in processing the other. It seems that the area of "collaborative" processing among multimodal biometrics, although relatively less explored currently, could hold potential for important gains.

There are some approaches which do not fit neatly into this independent/collaborative categorization. For instance, Papatheodorou and Reuckert [24] approached multimodal 2-D + 3-D face recognition by treating the data as points in a four-dimensional (4-D) space of (x, y, z, intensity). They were then able to use a 4-D iterative closest point (ICP) algorithm for the matching stage. Thus, the two properties of the face are treated in an integrated manner in the matching, so that it is not quite independent, but also certainly not collaborative in the sense that we want to suggest here. In contrast to the approaches above that generally use decision-level or score-level fusion, this approach might be said to use "data level" or "feature level" fusion.

III. EXPERIMENTAL METHODS AND MATERIALS

This section details the image dataset and the recognition algorithm used to generate the experimental results. The image dataset was acquired at the Computer Vision Research Lab at the University of Notre Dame, and is available to the research community. The approach popularly known as "eigen-faces" was used as the core recognition algorithm with each image modality, and the implementation used is one that is available to the research community [25].

A. Image Dataset

At a given image acquisition session, the intensity, IR, and 3-D images of a subject were all acquired within a period of a few minutes. The intensity images and the 3-D images were both acquired using a Minolta Vivid 900 range sensor [26]. This sensor produces a 640 × 480 sampling of range data, taking a few seconds to acquire a scan. It also acquires a 640 × 480 color intensity image just after sensing the range data. Infrared images were acquired with a Merlin uncooled long-wavelength IR camera, which produces a 240 × 320 IR image, with 12 bits of measurement resolution per pixel. During image acquisition, the subject stood approximately 1.5 m from the sensor, against a plain gray background, in a lab equipped with studio lighting. Subjects were asked to have a neutral facial expression, "F_A" in FERET terminology [27], and to look directly at the sensor.

A total of 191 subjects participated in one or more image acquisition sessions held at weekly intervals over a period of several months. For purposes of the experiments presented in this paper, the image dataset can be considered in three parts. Thirty-five subjects had good quality images in all three modalities in only one of the acquisition sessions. The images from these subjects are used only as part of the "training set" to create the "face space" used in the eigen-face method. Another 29 subjects had good images in all three modalities in each of four different acquisition sessions. The images from these subjects are used for a "tuning set" in creating the face spaces, as described later. Another 127 subjects participated in more than one acquisition session, and the images from these

persons are the data for the reported recognition performance results, the "test data." The images from the earliest acquisition session for each person are used as "gallery" images in the recognition experiment, and the images from later sessions are used as the "probe" images. For each modality, there are 297 probe images in the recognition experiment, with as little as one week and as many as thirteen weeks time lapse between the gallery and the probe.

B. Eigen-Face Recognition Algorithm

The eigen-face algorithm [28], [29] is used for the recognition experiments with each of the three image modalities. One reason for using this algorithm is that it is readily adapted for use with infrared images and with the range image representation of the 3-D data. Another reason is that there is a standard implementation available [25] that has been widely used in the face recognition research community as a "baseline" for evaluating other algorithms [2], [20]. This choice also simplifies the issue of score fusion in the multimodal results, since the scores from the different modalities are naturally scaled to the same range.

We made the methodological decision to use the same core recognition algorithm with each modality, and to use a "tuning set" of images to separately tune the algorithm for each modality. The "tuning" performed here involves selecting an appropriate set of dimensions from the eigen-space for use as the "face space." This is one approach to making a fair comparison between different image modalities. A different approach to a comparison could be to use the current best recognition algorithm for each modality, and so let the core recognition algorithm vary between the different types of images. However, there is no general consensus on the current best algorithm for each modality. Also, there would not necessarily be a standard open implementation of the different algorithms. A variation of this approach is to relax the concern to identify the best algorithm for each modality, and to simply use an algorithm for each modality that has been shown to be better than some "baseline" performance. There are numerous possible combinations of algorithms that might be used in this type of comparison.

In the context of evaluating face recognition performance, there is a set of gallery images and a set of probe images. The gallery images represent the persons enrolled into the system in order to be recognized. Each probe image or gallery image corresponds to a point in the face space. Face recognition systems can be used in two different application scenarios: 1) a recognition scenario, also referred to as identification and 2) a verification scenario, also referred to as authentication. In a recognition scenario, to decide the identity for a given probe, the distance is computed from the probe point in face space to each of the gallery points, and the closest gallery point indicates the identity of the probe. In a verification scenario, the probe comes with a claimed identity, and so the distance is computed just between the probe point in face space and the gallery point for the claimed identity. The claimed identity is then "verified" if the distance is small enough, or rejected otherwise. We use the Mahalanobis angle metric as the distance metric between two points in face space. We have found, as have others, that this metric gives better recognition performance than other metrics such as the Euclidean or Mahalanobis distance [29].

C. Tuning a Face Space for Each Modality

In the eigen-face approach to recognition, a "training set" of images is used to create a "face space" and individual face images are then represented as points in that space. It is common for the face space to have a reduced dimensionality relative to the eigen-space. Dimensions of the eigen-space corresponding to some number of the smaller magnitude eigen-values may be discarded on the basis of having minimal value for the recognition process. One or more dimensions of the eigen-space corresponding to the largest eigen-values may also be discarded, on a similar basis. While the larger eigen-values do represent dimensions of large variation between the images, these variations may have nothing to do with identity of the persons in the images. For example, the largest dimension of variation may be due to variations in lighting across the set of images.

We separately create an intensity image face space, a range image face space, and infrared image face space. For each modality, first the corresponding eigen-space is created from a training set of images. The training set for each modality has the same set of persons represented in its images. Then, for each modality, we determine the dimensions of the eigen-space to be kept in the face space by using a tuning set. The tuning data set consists of a gallery set of single images of 29 distinct subjects and a probe set of three images of the same 29 subjects in the gallery set for each modality. Dimensions of the original eigen-space are first dropped from those corresponding to smaller eigen-values until the rank-one recognition rate on the tuning set begins to decrease. Dimensions of the original eigen-space are then dropped from the larger eigen-value end, again until the rank-one recognition rate on the tuning set begins to decrease. The resulting intensity image face space has four large eigen-vector dimensions discarded and zero small eigen-vector dimensions discarded. The range image face space had one large eigen-vector dimension discarded and four small eigen-vector dimensions discarded. The infrared image face space also had one large eigen-vector dimension discarded and four small eigen-vector dimensions discarded. We have no reason to believe that the similar choices in tuning the range image face space and the infrared image face space are anything other than a coincidence.

Fig. 2 shows the first seven dimensions for each face space. It seems clear that the types of variation represented in the different face spaces are quite different.

D. Image Preprocessing

Face images are preprocessed before being used to create the face space. The example images shown in Fig. 1 reflect the results of this preprocessing. The preprocessing is meant to minimize image variations that are unrelated to identity, and so to focus the approach on variation between individuals. The images are masked to suppress the background and leave only the face region. The face region is also scaled to fill a standard size frame, 130 × 150 pixels in the implementation used. Additional mode-specific normalization is also performed.

The preprocessing of the intensity and infrared images is very similar, but the preprocessing of the 3-D shape to obtain the range image is more complicated. Both 2-D and IR images are normalized for pose variation only around the Z axis, the optical axis. The eigen-face software uses two landmark points (the eye centers) for geometric normalization to correct for rotation, scale, and position of the face. However, while histogram equalization is applied to normalize the brightness level in 2-D images, only geometric normalization is applied to IR images. For the results in this paper, the eye center landmark points are manually marked in each image, so that there are no catastrophic failures due to landmark location.

In the case where 3-D shape information is acquired, there is the opportunity to correct for pose variation around the X, Y, and Z axes. A transformation matrix is first computed based on the surface normal angle difference in X (roll) and Y (pitch) between manually selected landmark points (two eye tips and center of lower chin) and predefined reference points of a standard face pose and location. Each point defined in 3-D space for a range image has a depth value along the Z axis. Only the geometric normalization is needed to correct the pose variation. Pose variation around the Z axis (yaw) is corrected by measuring the angle difference between the line across the two eye points and a horizontal line. At the end of the pose normalization, each person's data is translated to have the nose tip at the same point in 3-D relative to the sensor, and rotated to have the same orientation of the triangle formed by the eye tips and chin point to the camera.

Also in the case where 3-D shape information is acquired, there are some particular potential artifacts that do not occur with intensity or infrared images. The two most common artifacts with the 3-D sensor that we used are called "holes" and "spikes." A hole in the data is a point where no 3-D measurement is made. This can happen when the projected light stripe is not imaged for some reason. A spike is a point where the computed 3-D measurement lies far from the actual surface in the scene. This can happen when, for example, there is interreflection of the projected light stripe in the scene. These problems are addressed by a preprocessing step that attempts to remove spikes by median filtering in a local window and attempts to fill holes by linear interpolation from points on the border of the hole.

IV. EXPERIMENTAL RESULTS FOR INDIVIDUAL MODALITIES

In this section, we present the experimental results for eigen-face recognition using each of the individual image modalities. We report experimental results in two formats.

Dimensions of **2D** face space (dimensions 5 through 11 of the original space)

Dimensions of **3D** face space (dimensions 2 through 8 of the original space)

Dimensions of **IR** face space (dimensions 2 through 8 of the original space)

Fig. 2. ***Eigen-face images of first seven dimensions in each face space.***

One is a cumulative match characteristic (CMC) curve. This form of the results is relevant to a recognition scenario, in which a probe image is matched against each of a set of gallery images in order to select the best match in the gallery as representing the identity of the probe. The rank-one recognition rate is the single number often quoted based on the CMC curve. This is the percentage of the probes for which the closest match in the gallery is the correct identity.

The other form of the experimental results is the receiver operating characteristic (ROC) curve. This form of the results is relevant to a verification scenario, in which a probe image is presented with a claimed identity, the image is matched against the gallery image for the claimed identity, and the claimed identity is considered as verified if the match is within some threshold. One common format of the ROC curve plots the false alarm rate (FAR) against the false reject rate (FRR). The equal error rate (EER) is the single number typically quoted from the ROC curve. This is the point at which the FAR is the same as the FRR.

We present both ROC curves and CMC curves in the results in this paper. It should be noted that the ROC curves allow a more sound comparison with results in other papers than do the CMC curves. This is because ROC curves are inherently less dependent on the gallery size. Because the CMC curve is more dependent on the gallery size, it is problematic to compare CMC style results in the literature that come from different size datasets. However, with either ROC or CMC style results, comparisons are still often problematic because different datasets of the same size can vary greatly in difficulty due to background, lighting, pose, and facial expression.

The CMC curve and the ROC curve for the individual modalities are presented in Fig. 3. In the case of both the ROC curve and the CMC curve, we find that recognition with the 2-D images performs slightly better than with with range images, and that both perform better than with the infrared images. The EER in the verification scenario is 2% for the intensity images, 3% for the range images, and 6% for the IR images. McNemar's test at the 0.05 level was used to test for statistical significance of the observed differences in the rank-one recognition rates [30]. The observed difference in rank-one recognition rates between 2-D and 3-D is not statistically significant. However, IR shows statistically significantly lower performance than either 2-D or 3-D.

There are a number of caveats to consider before drawing general conclusions about the inherent relative power of the different image modalities for face recognition. One is that the state of the art in 3-D sensors seems not nearly as well developed as that in normal cameras for capturing intensity images [4]. Artifacts occur in 3-D sensing that seemingly cause greater problems than the artifacts that occur in intensity images. Also, while 3-D shape is defined independent of lighting or

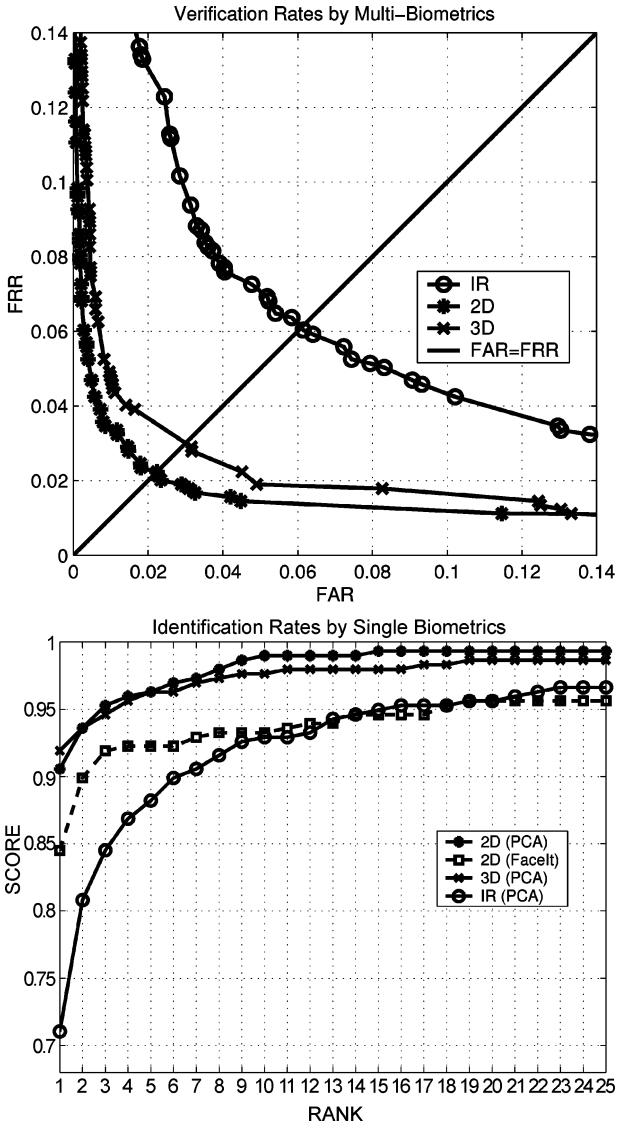

Fig. 3. *ROC and CMC curves for the individual modalities.*

pose, current sensors do not, practically speaking, acquire 3-D shape descriptions independent of lighting conditions or pose [4]. A change in the lighting conditions can induce a change in the sensed shape. Also, surfaces that are nearly tangent to the line of sight of the sensor are not sensed as well as surfaces that are more perpendicular to the line of sight.

The performance of the commercial algorithm FaceIt [31] for the intensity images is shown primarily to give a sense of how easy or difficult the dataset might be considered to be. A direct comparison of the FaceIt performance to that of the other algorithms is complicated by several factors. For the eigen-face algorithms used here, the set of training images includes the gallery images of the persons to be recognized. A commercial off-the-shelf

package such as FaceIt is of course is not able to focus on the particular set of persons in this way.

There are several points to be made with regard to the relative power of infrared face recognition. One is that the controlled indoor lighting conditions used in the data acquisition for this experiment are naturally well suited to normal intensity images. If images are acquired outdoors under highly variable lighting conditions, then infrared images can be expected to give better performance relative to intensity images. On the other hand, while infrared sensing is relatively independent of lighting conditions, the infrared pattern generated by a particular face does naturally vary with physiology, emotion, and other factors. Additional discussion of considerations in comparing infrared and normal visible images for face recognition can be found in [10]–[12].

V. MULTIMODAL EXPERIMENTAL RESULTS

There are many possible ways to combine recognition results from different algorithms. We first discuss the choices in the method of combination and then present experimental results of multimodal recognition.

A. Fusion Method

In our experiments, using the eigen-face approach with the cosine of the Mahalanobis angle as the distance metric, the raw score for a match between two points in face space ranges between -1 and $+1$. Given that fusion of results is done at the score level, there is still a choice of how to combine the scores. Researchers have considered a variety of methods for normalizing the scores prior to combination, including linear, logarithmic, exponential, logistic, etc. [32]. Normalizing the scores to a common range is important, but the particular range used for the normalization seems less essential. The normalized scores can be combined in any of several possible ways, including majority vote, sum, product, median, and min. Depending on the noise properties in the scores, a certain combination rule might be better than another. Many researchers have found that the sum and product rule provide generally good results in biometric applications [32]–[34].

We experimentally compared the use of linear, logarithmic, exponential, and a weighted linear normalization of scores from the individual modalities, in combination with sum, product, and min for combination of the normalized scores. The weighted linear normalization is described below. A summary of the recognition rates resulting from these various choices of score normalization and combination are listed in Table 1. It seems that combining the scores by choosing the minimum of the three scores is a poorer choice than combining by the sum or the product. It also seems that exponential normalization is a poorer choice than the other options considered. Other than these poor choices, the various options perform essentially equally well. The difference between the three entries in Table 1 that have 100% recognition and the three that have 99.7% recognition is just a difference in result for one probe. Any real differences between these combination methods are masked by the "ceiling effect" of the recognition rates approaching 100%.

Based on these results, we chose to linearly normalize the scores from each face space to the same range. The range used was 0 to 100, but the particular range is not essential. We also chose to use the weighted sum combination. The weight for the score from a given face space is based on the distribution of the top three ranks in that space. The motivation of the weighting is that a larger distinction between the first and second ranked matches implies a greater certainty that the first ranked match is correct. For each face space, a weight is computed using the first three rank scores as follows:

$$\text{Weight} = \frac{\text{score}_2 - \text{score}_1}{\text{score}_3 - \text{score}_1}.$$

Score_k is the kth closest distance from a gallery point to the given probe point. A plain sum rule would add the scores for each gallery subject across the three face spaces and select the subject with the smallest sum. The weighted rule sums the weighted scores.

B. Experimental Results

The ROC and CMC curves for the multimodal recognition results are shown in Fig. 4. Results are given for the different pairs of modalities as well as for the combination of all three. For comparison purposes, the individual results are also carried over to this figure.

Based on the ROC curve, the EER for the combination of intensity plus 3-D is 0.5%. The EER for 3-D plus IR is 0.7%. And the EER for intensity plus infrared is 1.3%. Based on the CMC curve, the rank-one recognition rate for intensity plus 3-D is 98.7%. The rank-one recognition rate for IR plus 3-D is 98%. And the rank-one recognition rate for intensity plus IR is 96.6%.

Table 1 Rank-One Recognition Rates 2-D + 3-D + IR for Various Fusion Methods

Normalization rule	Combination rule		
	Sum	Product	Minimum
Linear	99.7%	100.0%	91.6%
Logarithmic	100.0%	100.0%	91.6%
Exponential	91.6%	99.7%	91.6%
Weighted Linear	99.0%	99.7%	96.3%

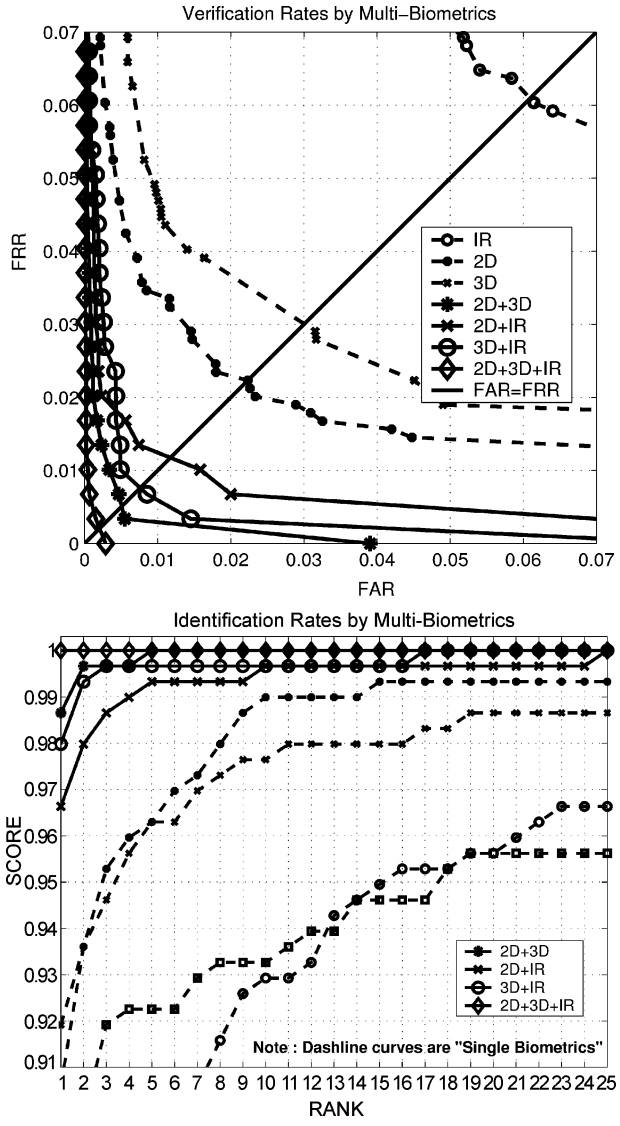

Fig. 4. ***Multimodal biometrics performance results in ROC and CMC curves.***

One "null hypothesis" underlying this experiment might be stated as—there is no significant difference in performance between individual biometrics and multibiometrics. Using McNemar's test for significance of the difference in rank-one recognition rate between using two modalities versus using one modality, we find that the recognition rate from two modalities is statistically significantly greater. Therefore, we would reject the null hypothesis.

The experimental results from combining all three modalities are better than those for any pair of modalities. The ROC curve shows an EER of 0.1% and the CMC curve shows a rank-one recognition rate of 100%. The difference in recognition rate between three modalities versus two modalities is clearly not statistically significant. Given the experimental dataset used here, the recognition rates for two modalities are already so high as to make it difficult to find a statistically significant improvement if one exists. Examples of some images that were incorrectly recognized in one of the individual modalities but correctly recognized in the three-modality results are shown in Fig. 5.

Results for the commercial face recognition system *FaceIt (Version G3)* are included in the ROC curve in Fig. 4. This is simply to give an indication of the relative difficulty of the image dataset used here in comparison to commercial face recognition technology. However, we should also note that the current commercial release of *FaceIt* is now at least version G5.5.

VI. MULTISAMPLE INTENSITY VERSUS MULTIMODAL

As mentioned earlier, in evaluating multimodal recognition results, it is important to compare to the use of multiple samples of a single modality. It is known that combining results from multiple samples of the same modality can improve performance over using a single sample of that modality [3], [33], [35]. In the context of face recognition, proposals for multimodal recognition generally anticipate adding another modality together with 2-D intensity images. Therefore, it is useful to compare the improvement from adding samples of other modalities to the improvement from using multiple intensity images.

In the same image acquisition sessions outlined earlier, each person also had four different intensity images acquired with a Canon Powershot G2 digital camera. These four images varied in lighting condition and facial expression. Two lighting conditions were used, one with three studio lights switched on and positioned one to either side in front of the person, and the other with an additional third studio light positioned straight ahead of the person. The person was requested to make two different facial expressions in each lighting condition, a neutral expression and a smile. The two lighting conditions are referred to as "LM" and "LF," and the two facial expressions are referred to as "FA" and "FB" [36]. So the four image conditions are FALM, FALF, FBLM, and FBLF. Example images that illustrate the expression and lighting variation appear in Fig. 6.

The same eigen-face recognition algorithm, as tuned for the intensity images in the experiments described earlier, was used for experiments with this additional set of intensity images. For this experiment, a person was represented by either two, three, or all four of the images taken in a given acquisition session. When a person is represented by two images for the gallery, and also by two images for the probe, each probe image is matched to each gallery image, for a total of four matches. The sum of the four match scores is then used as the overall score for

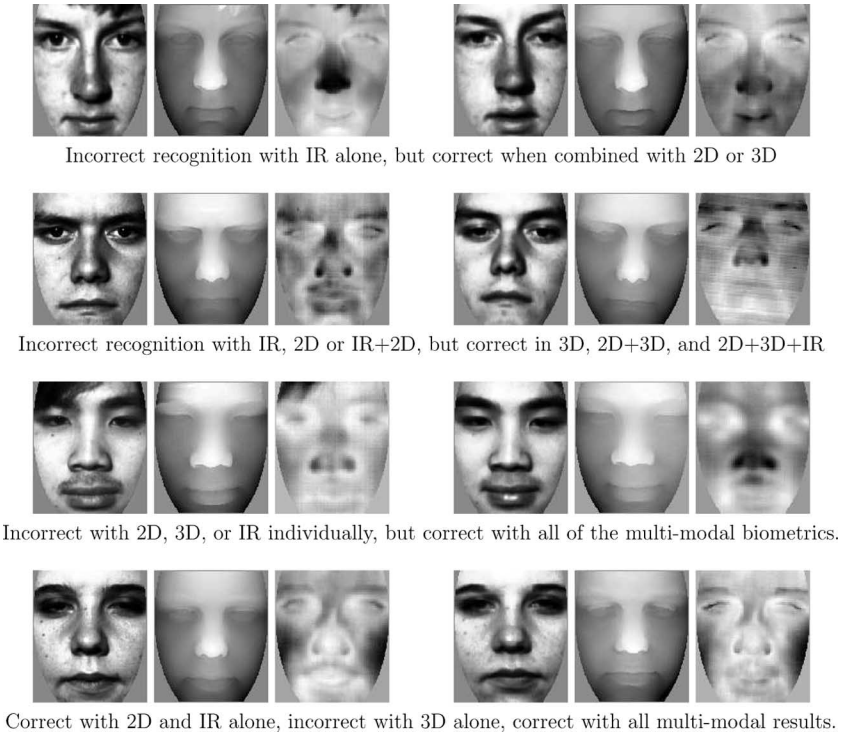

Fig. 5. ***Example matches where multimodal recognition improves over individual. The three images on the left are the gallery of each modality, and the three images on the right are the probe images.***

matching this probe person to this gallery person. Similarly, if each person is represented by three images as a gallery entry or as a probe, then the overall match score is a sum of nine matches. And with a person represented by all four images, the overall match score is a sum of 16 matches. Note that this approach to the overall match score uses more image-to-image matches for one person-to-person match than is possible with the multimodal matching. This is because in the multiple-sample approach, each probe image of a person can be matched to each gallery image of a person.

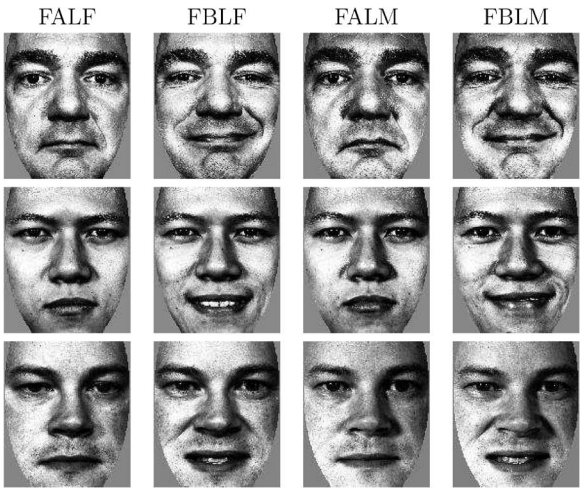

Fig. 6. ***Examples of expression and lighting variation in images for multisample experiment.***

The experimental results of this multiple-sample eigenface recognition are shown in ROC curve form in Fig. 7. Using two images, the FALM and the FALF imaging conditions, to represent a person, the rank-one recognition rate was 96.1%. Using three images, the FALM, FALF, and FBLM conditions, the recognition rate was 98.4%. Using all four image conditions, the recognition rate was 100%. Thus, using four intensity images to represent a person, with the images sampling different lighting and expression conditions as described, achieves the same level of recognition performance as three images that are each drawn from a different modality.

VII. SUMMARY AND DISCUSSION

This is the only work to compare face recognition using normal 2-D images, range images representing 3-D shape, and infrared images, and also the only work to evaluate the multimodal combination of the three types of images. Experimental results are based on an image dataset that has images of the same persons in each of the three

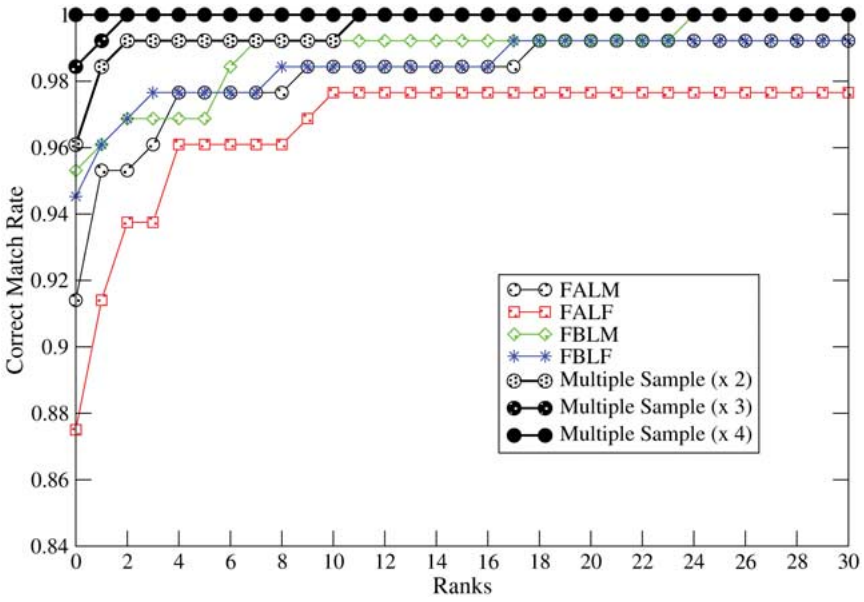

Fig. 7. ***Multisample intensity image recognition results.***

modalities. For a given person at a given acquisition session, the 2-D, 3-D, and IR images are acquired over an time interval of just a few minutes. This should allow the training, gallery, and probe sets for each modality to contain comparable images in the different modalities. We used the same PCA-based recognition engine, with the face space tuned individually for each modality, and all landmark points marked manually for each modality. We found that 3-D resulted in a higher rank-one recognition rate than 2-D, but that the difference was not statistically significant. We also found that the rank-one recognition rate for IR imagery was statistically significantly lower than that for 2-D or 3-D. However, the range of lighting conditions used in the image acquisition was typical of indoor office environments and this may not show off the strength of IR sensing.

We also compared the performance of individual modalities with multiple modalities. We found that each of the multimodal performances improved over all of the individual modalities, and that the multimodal 2-D + 3-D + IR technique performed best of all. The differences between the various multimodal performances were found not to be statistically significant. However, all of the multimodal performances were quite high, making it difficult to reliably detect differences. Additional investigation using a larger and more challenging dataset might reveal performance differences that were not detected here.

The comparison of multimodal performance versus multisample performance raises interesting and difficult issues. In general, it seems that it will be cheaper and more practical to acquire several intensity images than to acquire multiple image modalities. Also, for the experiments described in this paper, which use the same basic eigen-face algorithm and score-level fusion in comparing multisample versus multimodal, the multisample approach with intensity images achieves performance equivalent to the multimodal approach. However, using four identical intensity images will result in the same performance as using one image. The use of multiple intensity images is of value only if there is some variation between the individual images of a person. And very little is known about how to build the "right" degree of variation into a multisample approach. If the range of variation that may appear in the probe images is known, then it should be possible to determine the appropriate number and type of image samples to use to represent a person.

Lastly, to achieve the maximum possible performance, it seems reasonable that the eventual solution could be some combination of multisample and multimodal. Either multisample alone or multimodal alone could be expected to reach a plateau in performance at some number of samples or modalities. Achieving performance greater than this may required a combination of multiple samples of multiple modalities. This effect is also suggested by the work of Socolinsky *et al.* [11] in the context of multimodal fusion of infrared and visible.

It is worth noting once more that the images used in this study were all approximately frontal view and acquired under reasonably good lighting conditions. In conditions of very low light, infrared images can produce results where normal intensity images could not. And in conditions of extreme nonfrontal pose, 3-D face shape may be able to produce useful results where normal intensity images could not. ■

Acknowledgment

Images were acquired under a protocol approved by the University of Notre Dame's Institutional Review Board. Each person signed a consent form for each image acquisition session. The image data set is available to other research groups for noncommercial use. For information about dataset distribution, see the University of Notre Dame Computer Vision Research Lab Web page, http://www.nd.edu/~cvrl. Earlier descriptions of portions of this work appear in the 2004 Face and Gesture Recognition conference [37] and in the 2004 Biometric Technology for Human Identification conference [38].

REFERENCES

[1] W. Zhao, R. Chellappa, A. Rosenfeld, and P. J. Phillips, "Face recognition: A literature survey," *ACM Comput. Surv.*, vol. 35, pp. 399–458, Dec. 2003.

[2] P. J. Phillips, P. Grother, R. Michaels, D. Blackburn, E. Tabassi, and M. Bone. (2003, Mar.). *Facial Recognition Vendor Test 2002: Evaluation report*. [Online]. Available: http://www.frvt2002.org/FRVT2002/documents.htm

[3] K. Chang, K. W. Bowyer, and P. J. Flynn, "An evaluation of multi-modal 2D + 3D face biometrics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 4, pp. 619–624, Apr. 2005.

[4] K. W. Bowyer, K. Chang, and P. J. Flynn, "A survey of approaches and challenges in 3-D and multi-modal 3-D + 2-D face recognition," *Comput. Vis. Image Understand.*, vol. 101, no. 1, pp. 1–15, Jan. 2006.

[5] K. I. Chang, K. W. Bowyer, and P. J. Flynn, "Adaptive rigid multi-region selection for handling expression variation in 3-D face recognition," in *Proc. IEEE Workshop Face Recognition Grand Challenge Experiments*, 2005, vol. 3, p. 157.

[6] X. Lu and A. K. Jain, "Deformation analysis for 3-D face matching," in *Proc. 7th IEEE Workshop Applications of Computer Vision (WACV '05)*, pp. 99–104.

[7] T. Maurer, D. Guigonis, I. Maslov, B. Pesenti, A. Tsaregorodtsev, D. West, and G. Medioni, "Performance of geometrix activeID 3-D face recognition engine on the FRGC data," in *Proc. IEEE Workshop Face Recognition Grand Challenge Experiments*, 2005, vol. 3, p. 154.

[8] K. I. Chang, K. W. Bowyer, and P. J. Flynn, "Multiple nose region matching for recognition under varying facial expression," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 10, pp. 1695–1700, Oct. 2006.

[9] J. Wilder, P. J. Phillips, C. Jiang, and S. Wiener, "Comparison of visible and infrared imagery for face recognition," in *Proc. 2nd Int. Conf. Automatic Face and Gesture Recognition*, 1996, pp. 182–187.

[10] K. W. Bowyer, K. Chang, and P. J. Flynn, "Infra-red and visible light face recognition," *Comput. Vis. Image Understand.*, vol. 99, pp. 332–358, Sep. 2005.

[11] D. Socolinsky, A. Selinger, and J. Neuheisel, "Face recognition with visible and thermal infrared imagery," *Comput. Vis. Image Understand.*, vol. 91, pp. 72–114, 2003.

[12] A. Selinger and D. Socolinsky, "Face recognition in the dark," in *Proc. Workshop on Object Tracking and Classification Beyond the Visible Spectrum*, 2004, pp. 129–134.

[13] K. W. Bowyer, K. I. Chang, P. Yan, P. J. Flynn, E. Hansley, and S. Sarkar, "Multi-modal biometrics: An overview," presented at the 2nd Workshop on Multi-Modal User Authentication (MMUA 2006), Toulouse, France.

[14] K. Chang, K. W. Bowyer, V. Barnabas, and S. Sarkar, "Comparison and combination of ear and face images in appearance-based biometrics," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1160–1165, Sep. 2003.

[15] P. Grother. (2004, Feb.). *Facial Recognition Vendor Test 2002: Supplemental report*. [Online]. Available: http://www.frvt2002.org/FRVT2002/documents.htm

[16] B. Gokberk, A. A. Salah, and L. Akarun, "Rank-based decision fusion for 3-D shape-based face recognition," in *Proc. Int. Conf. Audio- and Video-Based Biometric Person Authentication (AVBPA 2005)*, pp. 1019–1028.

[17] C. Xu, Y. Wang, T. Tan, and L. Quan, "Automatic 3-D face recognition combining global geometric features with local shape variation information," in *Proc. 6th Int. Conf. Automated Face and Gesture Recognition*, 2004, pp. 308–313.

[18] N. V. Chawla and K. W. Bowyer, "Random subspaces and subsampling for 2-D face recognition," in *Proc. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. II:582–II:589.

[19] ——, "Designing multiple classifier systems for face recognition," in *Proc. Multiple Classifier Systems 2005*, 2005, pp. 407–416.

[20] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek, "Overview of the face recognition grand challenge," in *Proc. Conf. Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. I:947–I:954.

[21] Department of Homeland Security, *US VISIT Program*. [Online]. Available: http://www.dhs.gov/us-it

[22] A. Hutchinson, *Remarks by Undersecretary Hutchinson on the Launch of US VISIT*. [Online]. Available: http://www.dhs.gov/dhspublic/interapp/speech/speech_0114.xml

[23] M. Husken, M. Brauckmann, S. Gehlen, and C. von der Malsburg, "Strategies and benefits of fusion of 2-D and 3-D face recognition," in *IEEE Workshop Face Recognition Grand Challenge Experiments*, 2005, vol. 3, p. 174.

[24] T. Papatheodorou and D. Reuckert, "Evaluation of automatic 4-D face recognition using surface and texture registration," in *Proc. 6th Int. Conf. Automated Face and Gesture Recognition*, 2004, pp. 321–326.

[25] R. Beveridge, *Evaluation of Face Recognition algorithms*. [Online]. Available: http://www.cs.colostate.edu/evalfacerec/index.html

[26] Konica Minolta, *Vivid 910*. [Online]. Available: http://www.minoltausa.com

[27] P. J. Phillips, H. Moon, S. Rizvi, and P. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.

[28] M. Turk and A. Pentland, "Eigen-faces for recognition," *J. Cognitive Neurosci.*, vol. 3, no. 1, pp. 71–86, 1991.

[29] W. Yambor, B. Draper, and R. Beveridge, "Analyzing PCA-based face recognition algorithms: Eigen-vector selection and distance measures," presented at the 2nd Workshop on Empirical Evaluation in Computer Vision, Dublin, Ireland, 2000.

[30] G. Givens, R. Beveridge, B. Draper, and D. Bolme, "A statistical assessment of subject factors in the PCA recognition of human faces," presented at the Workshop Statistical Analysis in Computer Vision (CVPR), Madison, WI, 2003.

[31] Identix, Inc., *FaceIt—Face Biometrics*. [Online]. Available: http://www.identix.com

[32] B. Achermann and H. Bunke, "Combination of face classifiers for person identification," in *Proc. Int. Conf. Pattern Recognition*, 1996, pp. 416–420.

[33] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 3, pp. 226–239, Mar. 1998.

[34] G. Shakhnarovich and T. Darrell, "On probabilistic combination of face and gait cues for identification," in *Proc. Int. Conf. Automatic Face and Gesture Recognition*, 2002, pp. 169–174.

[35] N. Poh, S. Bengio, and J. Korczak, "A multi-sample multi-source model for biometric authentication," in *Proc. IEEE Int. Workshop Neural Networks for Signal Processing*, 2002, pp. 375–384.

[36] P. J. Flynn, K. W. Bowyer, and P. J. Phillips, "Assessment of time dependence in face recognition: An initial study," in *Proc. 4th Int. Conf. Audio and Video Based Biometric Person Authetication*, 2003, pp. 44–51.

[37] K. Chang, K. W. Bowyer, and P. J. Flynn, "Multi-biometrics using facial appearance, shape and temperature," in *Proc. Int. Conf. Face and Gesture Recognition*, 2004, pp. 43–48.

[38] K. Chang, K. W. Bowyer, P. J. Flynn, and X. Chen, "Evaluation of multi-modal biometrics using appearance, shape, and temperature," in *Proc. Biometric Technology for Human Identification*, 2004, pp. 1–11.

ABOUT THE AUTHORS

Kevin W. Bowyer (Fellow, IEEE) currently serves as the chair of the Department of Computer Science and Engineering at the University of Notre Dame, Notre Dame, IN. His recent research efforts have concentrated on the areas of data mining and biometrics. The Notre Dame biometrics research group has been active as a part of the support team for the government's Face Recognition Grand Challenge program and Iris Challenge Evaluation program.

Prof. Bowyer's paper "Face Recognition Technology: Security Versus Privacy" published in *IEEE Technology and Society*, was recognized with a 2005 Award of Excellence from the Society for Technical Communication, Philadelphia Chapter. He is a Golden Core Member of the IEEE Computer Society. He has served as Editor-in-Chief of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, and also serves or has served on the editorial boards of *Computer Vision and Image Understanding*, *Image and Vision Computing Journal*, *Machine Vision and Applications*, *International Journal of Pattern Recognition and Artificial Intelligence*, *Pattern Recognition*, *Electronic Letters in Computer Vision and Image Analysis*, and the *Journal of Privacy Technology*. He received an *Outstanding Undergraduate Teaching Award* from the USF College of Engineering in 1991, and *Teaching Incentive Program* awards in 1994 and in 1997.

Kyong I. Chang received the Ph.D. degree in computer science and engineering from the University of Notre Dame, Notre Dame, IN, in 2004.

He is a Principal Engineer in the Ultrasound Division at Philips Medical Systems, Bothell, WA. Currently, he investigates and designs advanced features in ultrasound imaging systems. His present research theme is concentrated on automated breast cancer interpretation in ultrasound images. His Ph.D. research focused on advanced improving biometrics for person identification. His research interests include biometrics, imaging processing, pattern recognition, and medical imaging.

Patrick J. Flynn (Senior Member, IEEE) received the Ph.D. degree in computer science from Michigan State University, East Lansing, in 1990.

He held faculty positions at Washington State University (1991–1998) and Ohio State University (1998–2001) and has been on the Notre Dame faculty since 2001. He is currently Professor of Computer Science and Engineering and Concurrent Professor of Electrical Engineering at the University of Notre Dame, Notre Dame, IN. His research interests include computer vision, biometrics, computer graphics and visualization, and image processing.

Dr. Flynn is a Fellow of IAPR, a past Associate Editor and Associate Editor-in-Chief of IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, and a current Associate Editor of *Pattern Recognition Letters*. He has received outstanding teaching awards from Washington State University and the University of Notre Dame.

Xin Chen received the Ph.D. degree in computer science and engineering from the University of Notre Dame, Notre Dame, IN, in 2006.

He currently works as a software engineer in the Technology and Development Department of NAVTEQ Corporation, Chicago, IL. His recent research efforts have concentrated on computer vision, pattern recognition, and image processing of video and aerial photos. His research at Notre Dame focused on biometrics, including infrared, and two- and three-dimensional face recognition.