IR and Visible Light Face Recognition

Xin Chen [*] Patrick J. Flynn Kevin W. Bowyer

Department of Computer Science and Engineering

University of Notre Dame

Notre Dame, IN 46556 USA

{xchen2, flynn, kwb}@nd.edu

Abstract

This paper presents the results of several large-scale studies of face recognition employing visible light and infra-red (IR) imagery in the context of principal component analysis. We find that in a scenario involving time lapse between gallery and probe, and relatively controlled lighting, (1) PCA-based recognition using visible light images outperforms PCA-based recognition using infra-red images, (2) the combination of PCA-based recognition using visible light and infra-red imagery substantially outperforms either one individually. In a same session scenario (i.e. near-simultaneous acquisition of gallery and probe images) neither modality is significantly better than the other. These experimental results reinforce prior research that employed a smaller data set, presenting a convincing argument that, even across a broad experimental spectrum, the behaviors enumerated above are valid and consistent.

Keywords

multi-modal, biometrics, face recognition, infrared imaging, principal components analysis

[*] Corresponding author. Mail address: 326 Cushing Hall, University of Notre Dame, Notre Dame, IN 46556 USA. Telephone: (574) 210-9895. Fax: (574) 631-9260.

1 Introduction

IR cameras provide a measure of thermal emissivity from the facial surface and their images are relatively stable under illumination variation [1]. The anatomical information which is imaged by infrared technology involves subsurface features believed to be unique to each person [2], though the twins images are not necessarily substantially different as shown in Figure 1. Those features may be imaged at a distance, using passive infrared sensor technology, with or without the cooperation of the subject. IR therefore provides a capability for identification under all lighting conditions including total darkness [3]. Limitations of IR cameras, such as their resolutions being below that of visible light spectrum cameras, should also be acknowledged [1].

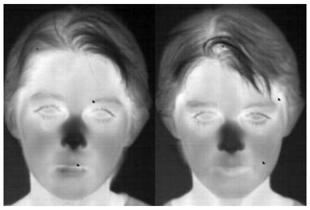


Figure 1: A pair of twins IR images

Face recognition in the thermal IR domain has received relatively little attention in the literature in comparison with recognition in visible light imagery. Wilder *et al.* [4] demonstrated that both visible light and IR imageries perform similarly across algorithms. Early studies by Socolinsky *et al.* in [5], [6] and [7] suggest that long-wave infrared imagery of human faces is not only a valid biometric, but superior to using comparable visible light imagery. However, the testing set size in these studies is relatively small, the training and gallery are composed of disjoint sets of images of the same subjects, and there is no substantial time lapse between gallery and probe image acquisition. Our early studies [8] [9] [10] show that recognition performance is substantially poorer when unknown images are acquired on a different day from the enrolled images. The FRVT 2002 coordinators [11] report that face recognition performance decreases approximately linearly with elapsed time. Hence, the time-lapse issue is one reason why [9] and [10] seem to be at odds with

[5], [6], [7] and [12], since the former shows that PCA-based face recognition using visible light imagery may outperform that using infrared images. The following factors could also contribute to the discrepancies:

1. Chen *et al.* [9] [10] manually locate eye-locations in infrared images. Socolinsky *et al.* [5] [6] [7] [12] use a sensor capable of imaging both modalities simultaneously through a common aperture which enables them to register the face with reliable visible light images instead of infrared images. [10] shows that relatively unreliable face registration degrades performance in IR.

2. Chen *et al.* [9] and [10] used much higher resolution for visible light source images than the infrared images (240 × 320). The resolutions of visible light and infrared images used in [5], [6], [7] and [12] are both 240 × 320.

3. There might be more variations of the facial appearance in [5], [6], [7] and [12] since the images were recorded when the subject pronounced vowels looking directly toward the camera, while the subjects in [9] and [10] are required to demonstrate only neutral and smiling expressions. Infrared face images could be more insensitive to facial expression change. More recent work by Selinger and Socolinsky [13] looks at issues of eye location accuracy in visible-light and infra-red images, and at recognition accuracy in the case of outdoor imagery that may exhibit much greater lighting variation than the indoor imagery in [4,5,6,7,8,9]. They find that although eyes cannot be detected as reliably in thermal images as in visible ones, some face recognition algorithms can still achieve adequate performance [13]. They also find that, while recognition with visible-light imagery outperforms that with thermal imagery when both gallery and probe images are acquired indoors, if the probe image or the gallery and probe images are acquired outdoors, then it appears that the performance possible with IR can exceed that with visible light.

This paper extends our early research analyzing the PCA algorithm performance in infrared imagery, including the impact of illumination change, facial expression change and the short term (minutes) and medium term (days or weeks) change in face appearance. It also presents a comparative study employing visible light imagery.

The rest of this paper is organized as follows. Section 2 describes the data collection process including the experimental equipment, environmental conditions and the subject specifications. Our experimental scheme based on the PCA algorithm is presented in section 3. Section 4 discusses the data preprocessing and its specifications. Section 5 summarizes the PCA technique from a pattern recognition perspective and discusses distance measures and their effect on recognition performance. In Sections 6 and 7, we give the results of two main recognition experiments: same-session and time-lapse recognition. Section 8 compares recognition performance in same-session and time-lapse. We investigate the effect of time dependency on recognition in Section 9. Performance sensitivity to eye center location is studied in Section 10. Since training the subspace in the PCA algorithm is vital to recognition performance, we discuss in Section 11 three important factors: training bias, training set size and eigenvector tuning. We explore the impact of illumination change and facial expression change on recognition performance in Section 12. Combination of visible light and infrared imageries is explored in Section 13. Section 14 compares the performance of PCA and a commercial face-recognition algorithm. We conclude in Section 15.

2 Data Collection

Our database consists of 10916 images per modality (visible light and IR) from 488 distinct subjects. Most of the data was acquired at the University of Notre Dame during the years 2002 and 2003, while 81 images per modality from 81 distinct subjects were acquired by Equinox Corporation. Selinger and Socolinsky [6] describe in detail the acquisition process of the data collected by Equinox Corporation.

Acquisitions were held weekly and most subjects participated multiple times across a number of different weeks. Infrared images were acquired with a Merlin uncooled long-wavelength infrared high-performance camera [1], which provided a real-time, 60Hz, 12 bit digital data stream. It is sensitive in the 7.0-14.0 micron range and consists of an uncooled focal plane array incorporating a 320 × 240 matrix of microbolometer detectors. Three Smith-Victor A120 lights with

[1] http://www.indigosystems.com/product/merlin.html

Sylvania Photo-ECA bulbs provided studio lighting. The lights were located approximately eight feet in front of the subject; one was approximately four feet to the left, one was centrally located and one was located four feet to the right. All three lights were trained on the subject face. The side lights and central light are about 6 feet and 7 feet high, respectively. One lighting configuration had the central light turned off and the others on. This will be referred to as “FERET style lighting” or “LF” [14]. The other configuration has all three lights on; this will be called “mugshot lighting” or “LM”. For each subject and illumination condition, two images were taken: one is with neutral expression, which will be called “FA”, and the other image is with a smiling expression, which will be called “FB”. Due to IR’s opaqueness to glass, we required all subjects to remove eyeglasses during acquisition. Figure 2 (a) shows four views of a single subject in both visible light and infrared imagery acquired at University of Notre Dame. Two images of a single subject in visible light and infrared imagery acquired by Equinox Corporation are illustrated in Figure 2 (b). The infrared images shown in this figure have contrast enhanced for display.

3 Experimental Designs

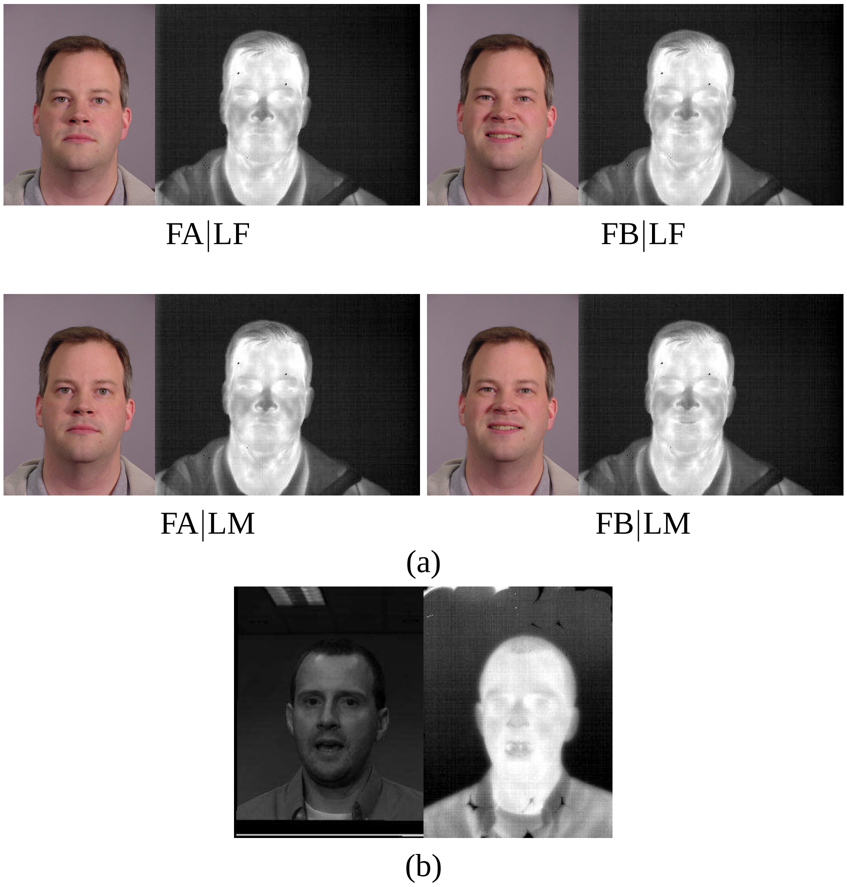
Each face recognition experiment is characterized by three image sets.

a. The *training set* is used to form a face space in which the recognition is performed.

b. The *gallery set* contains the set of “enrolled” images of the subjects to be recognized, and each image is uniquely associated with the identification of a distinct subject in the set.

c. The *probe set* is a set of images to be identified via matching against the gallery. We employ a closed universe assumption; i.e. each probe image will have a corresponding match in the gallery.

In our experiments, the training set is disjoint with the gallery and probe sets (actually, in most of our experiments, the training set would not contain any persons in common with those in the gallery and probe set), which makes the performance worse than otherwise. This is to eliminate any bias that might be introduced in the eigenspace due to subject factors and make the evaluation of the face recognition system more objective.



FA|LF

FB|LF

FA|LM

FB|LM

(a)

(b)

Figure 2: (a) Four views with different lighting and expressions in visible light and infrared imagery, acquired at University of Notre Dame; (b) Two images of a single subject in visible light and infrared imagery, acquired at Equinox Corporation.

According to the lighting and expression situation when the images were acquired, there are four categories: (a) FA expression under LM lighting (FA|LM), (b) FB expression under LM lighting (FB|LM), (c) FA expression under LF lighting (FA|LF) and (d) FB expression under LF lighting (FB|LF). All the subsequent experiments use the valid combinations of two subsets of the image database and each set belongs to one of these four categories.

The three primary face recognition tasks are listed below [11]:

1. Verification: "Am I who I say I am?" A person presents their biometric and an identity claim to a face recognition system. The system then compares the presented biometric with a stored biometric of the claimed identity. Based on the results of comparing the new and stored biometric, the system either accepts or rejects the claim.

2. Identification: "Who am I?" An image of an unknown person is provided to a system. We assume that through some other method we know the person is in the database. The system then compares the unknown image to the database of known people.

3. Watch list: "Are you looking for me?" A face recognition system must first detect if an individual is, or is not, on the watch list. If the individual is on the watch list, the system must then correctly identify the individual.

Our work focuses on the identification task. The main performance measure for identification systems is the ability to identify a biometric signature's owner. More specifically, the performance measure equals the percentage of queries in which the correct answer can be found in the top few matches [15]. Identification of a probe image yields a ranked set of matches, with rank 1 being the best match. Results are presented as cumulative match characteristic (CMC) curves, where the x-axis denotes a rank threshold and the y-axis is the fraction of probes that yields a correct match at ranks equal to or lower than the threshold.

By selecting meaningful data sets as the pairs of galleries and probes, we conducted several experiments to investigate face recognition performance in visible light and infrared imagery. We require that each image involved in the experiment used in one modality should have a counterpart (acquired at the same time, under the same condition and of the same subject) in the other modality. The PCA software suite used in our experiments was developed at Colorado State University (CSU) [2] and slightly modified by us to correctly read 12-bit infrared images.

4 Preprocessing

As very few existing software applications can automatically locate a face in the image and humans generally outperform a computer in this task, we located faces manually by clicking with a mouse on the center of each eye. Figure 2 shows that the features on a human face appear more vague in an infrared image than those in a visible light image and thus the registration in the following normalization step might not be as reliable in IR as in visible light images.

From Figure 2, we notice that the background, some possible transformations of the face (scaling, rotation and translation) and sensor-dependent variations (for example, automatic gain control calibration and bad sensor points) could undermine the recognition performance. This impact can be minimized by normalization, which is implemented in the CSU software.

The CSU software supports several metrics for normalization:

a. *Integer to float conversion.* After the image is read from a file, it is converted to double precision (64 bit) floating point for subsequent image calculations.

b. *Geometric normalization.* This aligns images such that the faces are the same size, in the same position and at the same orientation. Specifically, the image is scaled and rotated to make the eye coordinates coincident with prespecified locations in the output.

c. *Masking.* Masking is used to eliminate parts of the image that are not the face. This is to ensure that the face recognition system does not respond to features corresponding to background, hair, clothing etc. The CSU system uses an elliptical mask that is centered just below eye level and obscures the ears and sides of the face. This is the same mask as used in the FERET experiments [14].

d. *Histogram equalization.* Histogram equalization attempts to normalize the image histogram

[2] http://www.cs.colostate.edu/evalfacerec/

to reduce image variation due to lighting and sensor differences.

e. *Pixel normalization.* This is to compensate for brightness and contrast variations. The CSU code does this by changing the dynamic range of the images such that the mean pixel value is 0.0 and the standard deviation is 1.0.

We found that the recognition system performs best when turning on all the normalizations above with default options and a, b, c, d and e applied in order. Other settings bring no significant performance gain or yield even worse performance. For example, we tried turning off histogram equalization, considering that the original gray value response at a pixel is directly related to thermal emission flux and our algorithm might benefit most from arrays of corresponding thermal emission values rather than arrays of gray values. The result turned out to be no better than turning the histogram equalization on.

5 PCA Algorithm

The PCA method was first described for face image representation by Sirovich and Kirby [16] and adapted to face recognition by Turk and Pentland [17]. The face recognition system in our experiments should be able to do the following:

a. Derive a classification rule from the face images in the training set; i.e. it should be able to develop a discrimination technique to separate images of different subjects.

b. Apply the rules to new face images; i.e. given a set of new enrolled images as the gallery and a set of new unidentified images as the probe, it should be able to use the discrimination technique to map each probe image to one gallery image.

5.1 Definition

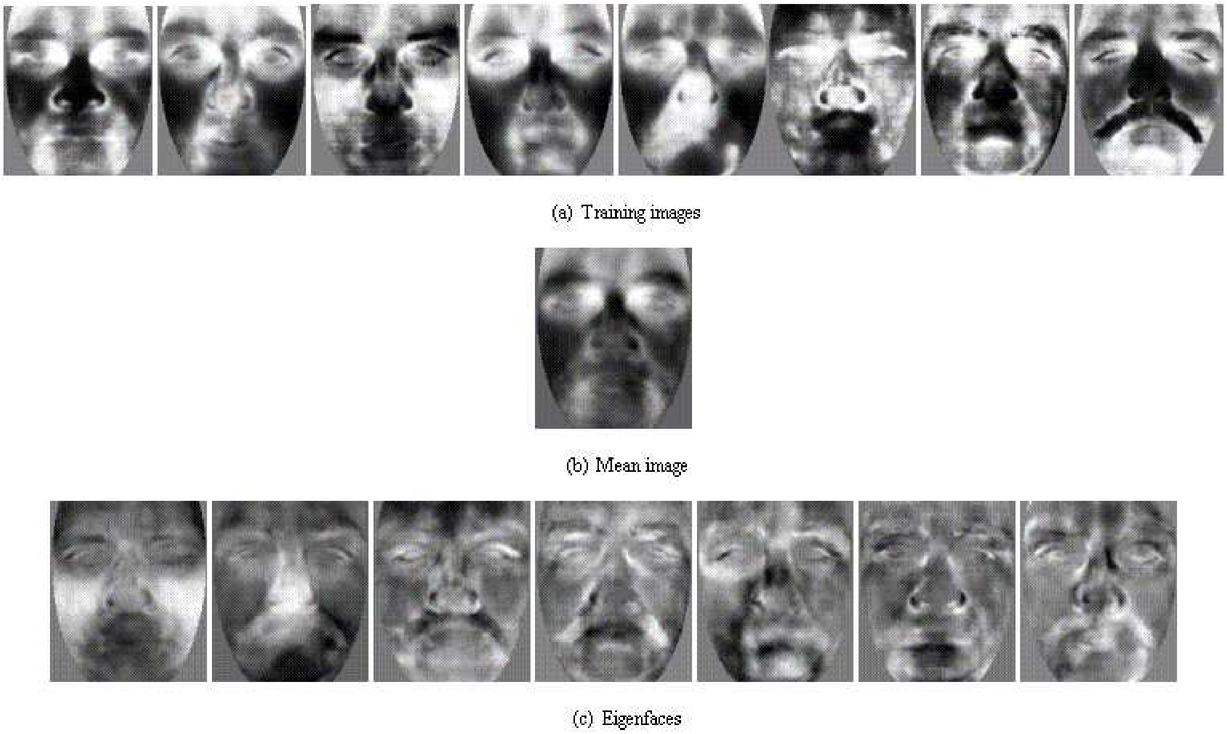
Given a training set of N images $\{x_1, x_2, ..., x_N\}$, all in $\mathbb{R}^n$, taking values in an $n-$dimensional image, PCA finds a linear transformation W^T mapping the original $n-$dimensional image space into an $m-$dimensional feature space, where $m < n$. The new feature vectors have coordinates

$$y_k = W^T x_k, k = 1, 2, ..., N$$

where $W \in \mathbb{R}^{n \times m}$ is a matrix with orthonormal columns. We define the total scatter matrix S_T as:

$$S_T = \sum_{k=1}^{N} (x_k - \mu)(x_k - \mu)^T \tag{1}$$

where N is the number of training images, and $\mu \in \mathbb{R}^n$ is the sample mean of all images. Examples of an infrared image training set and its sample mean image are shown in Figure 3 (a) and (b), respectively.



(a) Training images

(b) Mean image

(c) Eigenfaces

Figure 3: (a) Training images: frontal IR images of eight different subjects. (b) Mean image: average of the eight images in (a). (c) Eigenfaces: principal components calculated from (a) in decreasing eigenvalue order.

After applying the linear transformation W^T, the scatter of the transformed feature vectors

$y_1, y_2, ..., y_N$ is $W^T S_T W$. In PCA the projection W_{opt} is chosen to maximize the determinant of the total scatter matrix of the projected samples, i.e.

$$W_{opt} = argmax_W |W^T S_T W| = [w_1 w_2 ... w_m]$$

where $w_i = 1, 2, ..., m$ is the set of $n-$dimensional eigenvectors of S_T corresponding to the m largest eigenvalues [6]. Since these eigenvectors are face-like in appearance when rearranged so that they follow the original image pixel arrangement, they are commonly referred to as "eigenfaces". They are also referred to as principal components. The Eigenface method, which uses principal components analysis for dimensionality reduction, yields projection directions that successively maximize the total residual scatter across all classes, i.e. all images of all faces [18]. Figure 3 c shows the top seven eigenfaces derived from the input images of Figure 3 a in decreasing eigenvalue order.

5.2 Distance Measures

Any eigenface matches must employ a measure of proximity in the face space. The "MahCosine" (named by the CSU software) and Mahalanobis distance are simply the angle metric [19]

$$d(x, y) = -\frac{x \cdot y}{||x||||y||} = -\frac{\sum_{i=1}^{k} x_i y_i}{\sqrt{\sum_{i=1}^{k} (x_i)^2 \sum_{i=1}^{k} (y_i)^2}}$$

and Euclidean distance measure [19]

$$d(x, y) = ||x - y||^2 = \sum_{i=1}^{k} |x_i - y_i|^2$$

applied in the weighted space, respectively.

The CSU software implements the MahCosine measure, the classical Euclidean distance measure and the city block distance measure [19]

$$d(x, y) = |x - y| = \sum_{i=1}^{k} |x_i - y_i|.$$

We also implemented the Mahalanobis distance measure for comparison with MahCosine. Based on initial experiments, we found that MahCosine offered the best performance, and so this metric is used for all results reported in this paper.

6 Same-session Recognition

The two experiments described in this section use ”same session” images. That is, the gallery and probe images were taken within a minute of each other at the same acquisition session.

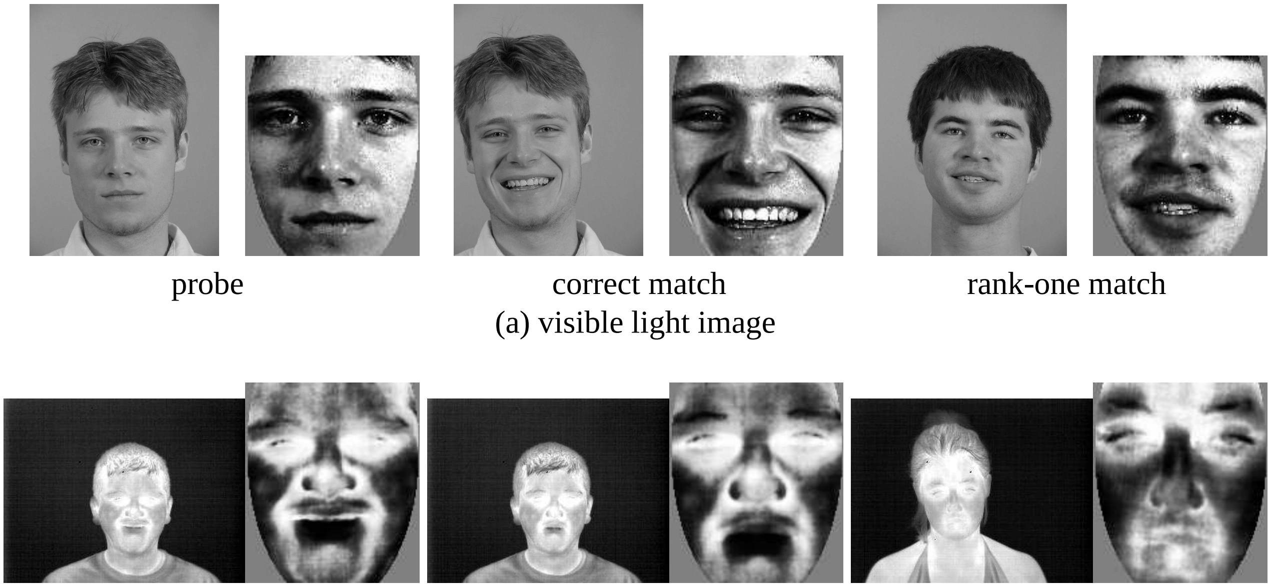
In the first experiment, We used 82 distinct subjects and four images for each subject acquired within one minute with different illumination and facial expressions. These images were acquired during spring 2002. For each valid pair of gallery and probe sets, we computed the rank 1 correct match percentage and the rank at which all the probes were correctly matched. They are reported in Table 1. Each entry in the leftmost column corresponds to a gallery set, and each entry in the top row corresponds to a probe set. The subspace for Table 1 was derived by using 240 images of 240 distinct subjects. These individuals are not in the gallery or probe set.

Probe Gallery	FA|LF	FA|LM	FB|LF	FB|LM
FA|LF		0.98 (2) 0.98 (10)	0.99 (3) 0.98 (10)	0.99 (2) 0.94 (4)
FA|LM	0.99 (2) 0.95 (6)		0.94 (28) 1.00 (1)	0.95 (19) 1.00 (1)
FB|LF	0.96 (4) 0.95 (6)	0.95 (39) 1.00 (1)		1.00 (1) 1.00 (1)
FB|LM	0.98 (2) 0.89 (17)	0.96 (19) 0.98 (3)	1.00 (1) 0.98 (3)	

Table 1: The percentage of correctly matched probes at rank 1 and in parentheses, the smallest rank at which all probes are correctly matched for same session recognition in visible light (bottom) and IR (top)

Table 1 shows that there is no consistent difference between the performance of visible light and IR. IR is better in six instances, visible light is better in four instances, and they are the same in two instances. The overall performance for same session recognition is high for both IR and visible light, and so it is possible that some “ceiling effect” could make it difficult to observe any true difference that might exist.

Figure 4 shows the worst mis-matches for visible light and IR, i.e. the probe image, the correct match and the rank-one match.



probe correct match rank-one match
(a) visible light image

probe correct match rank-one match
(b)IR image

Figure 4: Worst match examples

A similar same-session experiment using 319 distinct subjects and four images for each subject acquired within one minute during spring 2003 with different illumination and facial expressions is conducted and reported in Table 2. The face space for this experiment was derived by using one image for each of 488 distinct subjects and all eigenvectors were retained. Of the 488 training images, 319 (FA|LF) come from spring 2003, which means that the gallery and probe sets of some subexperiments overlap with the training set. The performance of the subexperiments in which the probe set is FA|LF should be ignored, because probe set and training set must be disjoint for a fair comparison[12].

Probe Gallery	FA|LF	FA|LM	FB|LF	FB|LM
FA|LF		0.73 (312) 0.96 (126)	0.76 (312) 0.90 (276)	0.72 (309) 0.89 (223)
FA|LM	N/A N/A		0.78 (226) 0.91 (254)	0.81 (312) 0.93 (259)
FB|LF	N/A N/A	0.80 (231) 0.94 (220)		0.84 (286) 0.96 (110)
FB|LM	N/A N/A	0.83 (312) 0.93 (212)	0.84 (287) 0.96 (97)	

Table 2: The percentage of correctly matched probes at rank 1 and in parentheses, the smallest rank at which all probes are correctly matched for same session recognition in visible light (bottom) and IR (top), using Spring 2003 data

A striking difference from the previous same-session recognition result is the much lower performance of infrared face recognition. The comparable experiment using visible light images still achieves very good performance given a reasonably large face space. Apparently, the visible light face recognition performance degrades slightly when the expressions of the gallery and probe images are different.

Selinger and Socolinsky have looked at automated eye location in visible-light versus thermal imagery [13]. They find that although the error increase from visible to LWIR is large, LWIR values still stay within 15% of the eye size, quite a reasonable bound [13]. Their recognition experiments are based on evaluating recognition performance using a 40-frame video sequence as input, potentially complicating a direct comparison of recognition results.

7 Time-lapse Recognition

Experiments in which there is substantial time passage between gallery and probe image acquisitions is referred to as time-lapse recognition.

Our first experiment of this type uses the images acquired in ten acquisition sessions of Spring 2002. In the ten acquisition sessions, there were 64, 68, 64, 57, 49, 56, 54, 54, 60, and 44 subjects.

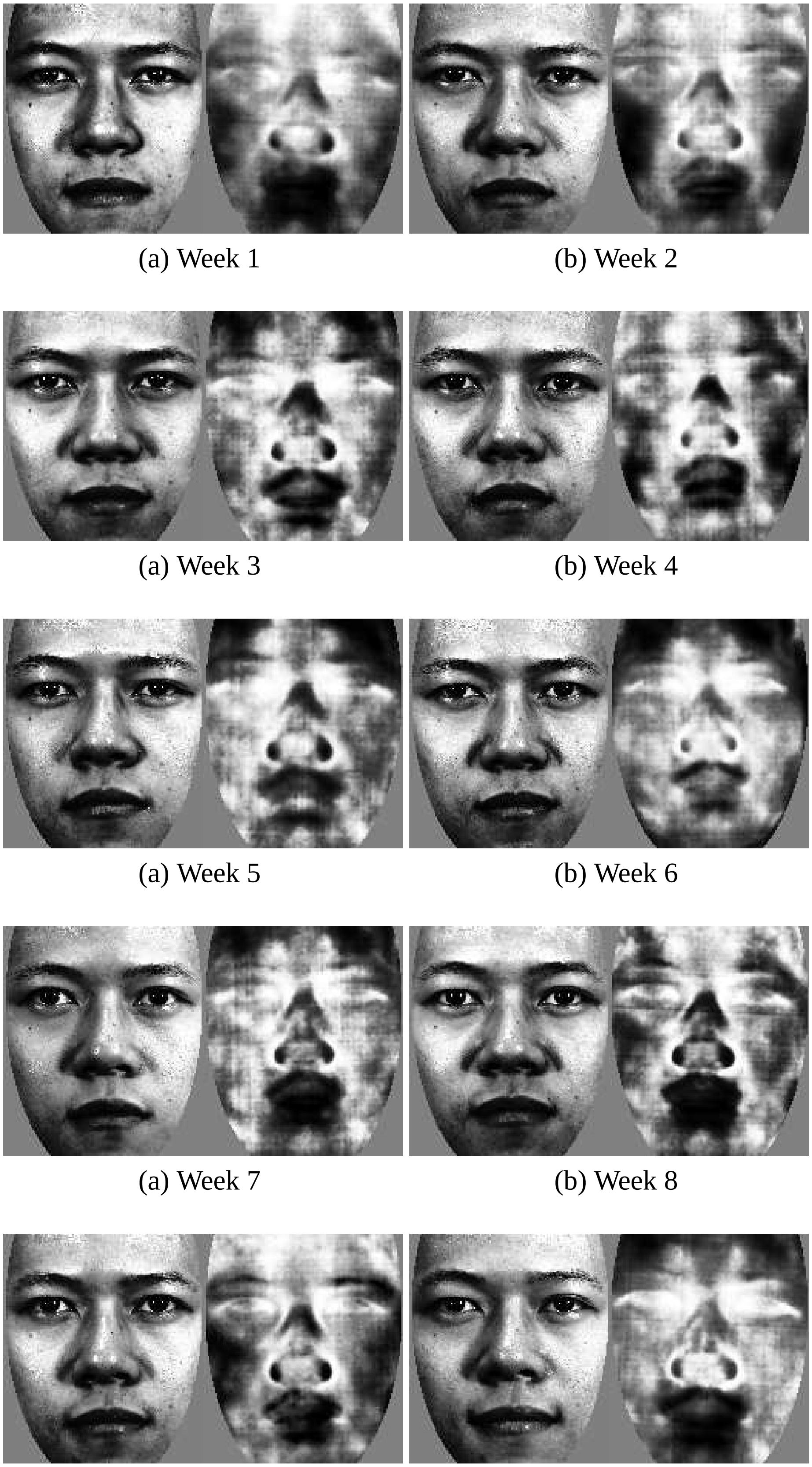
Figure 5 shows the visible light and IR images of one subject across 10 different weeks, which suggests that there may be more apparent variability, on average, in the IR images of a person than in the visible light images. For example, note the variation in IR images in the cheeks and temples between weeks 9 and 10, or between the bridge and sides of the nose in different IR images. Other research [20] has confirmed that there is variability in facial IR images due to startling, gum-chewing, etc. More recently, Socolinsky *et al.* [21][22] have replicated our basic early result [10][9] of lower IR performance in the time-lapse experiments.

The scenario for this recognition is a typical enroll-once identification setup. There are 16 subexperiments based on the exhaustive combinations of gallery and probe sets given the images of the first session under a specific lighting and expression condition as the gallery and the images of all the later sessions under a specific lighting and expression condition as the probe. That is, each gallery set has 64 images from session 1, and each probe set has 431 images from sessions 2-10. The rank-1 correct match percentages are given in Table 3. For each subject in one experiment, there is one enrolled gallery image and up to nine probe images, each acquired in a distinct later session. The same face space is used as in the first ”same-session” experiments.

Probe Gallery	FA|LM	FA|LF	FB|LM	FB|LF
FA|LM	0.83 (41) 0.91 (39)	0.84 (27) 0.93 (54)	0.77 (48) 0.73 (56)	0.75 (43) 0.71(56)
FA|LF	0.81 (38) 0.92 (31)	0.82 (46) 0.92 (28)	0.74 (49) 0.75 (32)	0.73 (43) 0.73 (44)
FB|LM	0.77 (45) 0.77 (33)	0.80 (49) 0.81 (44)	0.79 (39) 0.86 (48)	0.78 (51) 0.85 (47)
FB|LF	0.73 (58) 0.75 (41)	0.76 (58) 0.79 (40)	0.77 (36) 0.90 (27)	0.76 (41) 0.90 (47)

Table 3: Rank 1 correct match percentage for time-lapse recognition in visible light (bottom) and IR (top). Row indicates gallery and column indicates probe.

For IR, Table 3 illustrates a striking difference in performance relative to the same-session recognition results shown in Table 1. Visible light imagery outperforms IR in 12 of the 16 cases, with IR and visible light the same in another two. The rank 1 correct match rate for IR drops by 15% to 20%. The most obvious reason is that the elapsed time caused significant changes in the



(a) Week 1

(b) Week 2

(a) Week 3

(b) Week 4

(a) Week 5

(b) Week 6

(a) Week 7

(b) Week 8

(a) Week 9

(b) Week 10

Figure 5: Normalized FA|LM face images of one subject in visible light and IR across 10 weeks.

thermal patterns of the same subject. Table 3 also shows that the performance degrades for visible light imagery compared with that in same-session recognition.

For one time-lapse recognition with FA$|$LF images in the first session as the gallery set and FA$|$LF images in the second to the tenth sessions as the probe set, we illustrate the match and non-match distance distributions in Figure 6 and Figure 7. The score (distance) ranges from -1.0 to 1.0 since we use the ”MahCosine” distance metric. The match score histogram is the distribution of distances between the probe images and their correct gallery matches. The non-match score histogram is the distribution of distances between the probe images and all their false gallery matches. Essentially, the match score distribution depicts the within-class difference, while the non-match score distribution represents the between-class difference. Hence, for an ideal face recognition, the match scores should be as small as possible and the non-match scores should be much larger than the match scores and they shouldn’t overlap. In this experiment, there is significant overlapping for both IR and visible light, which accounts for the incorrect matches. The match score distribution for visible light is more at the smaller distance area than that for IR, i.e. the within-class difference for visible light images is smaller than that for IR images. The non-match score distributions for these two modalities are about the same, i.e. the between class differences are similar. Thus, visible light imagery performs better than IR in this setup. Note that our experimental setup includes relatively minimal lighting variations. If more drastic lighting variation was considered, the results could well be different. For example, in the extreme case of no ambient light, one would naturally expect IR to perform better.

Another similar time-lapse experiment using the images acquired in 12 acquisition sessions of Spring 2003 is conducted and the results are reported in Table 4. The face space for this experiment is the same as that used in the second same-session experiment.

As is the case with the first time-lapse experiment, the performance with IR images drops substantially in comparison to the same-session performance shown in Table 2: the rank 1 correct match rate drops by 15% to 30%. The most obvious reason is that the elapsed time caused signifi-

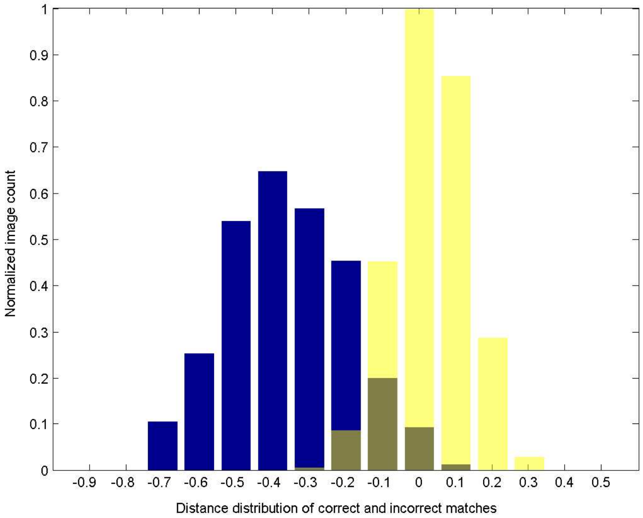


Figure 6: Match and non-match score distributions for one time-lapse recognition in IR, dark color bars represent correct match, light color bars represent incorrect match

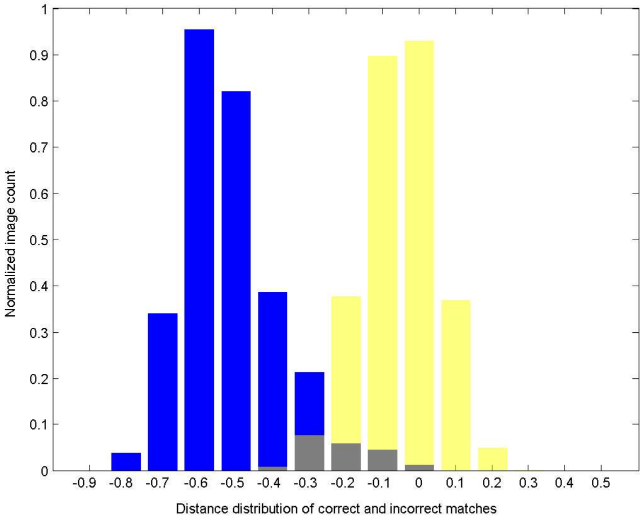


Figure 7: Match and non-match score distributions for one time-lapse recognition in visible light, dark color bars represent correct match, light color bars represent incorrect match

Probe Gallery	FA|LM	FA|LF	FB|LM	FB|LF
FA|LM	0.58 (142) 0.89 (116)	0.56 (143) 0.88 (135)	0.51 (142) 0.68 (136)	0.51 (142) 0.66(106)
FA|LF	0.61 (142) 0.85 (137)	0.62 (142) 0.86 (138)	0.52 (141) 0.64 (136)	0.55 (141) 0.66 (138)
FB|LM	0.55 (143) 0.76 (133)	0.53 (139) 0.76 (138)	0.58 (141) 0.82 (121)	0.57 (142) 0.82 (108)
FB|LF	0.54 (140) 0.74 (134)	0.55 (143) 0.76 (141)	0.58 (143) 0.79 (125)	0.56 (139) 0.79 (134)

Table 4: Rank 1 correct match percentage for time-lapse recognition in visible light (bottom) and IR (top). Row indicates gallery and column indicates probe.

cant changes among thermal patterns of the same subject. In addition, the overall low performance for infrared face recognition is due to the unreliable registration of the eye centers discussed in last section. Table 4 also shows that the performance degrades for visible light imagery compared with that in same-session recognition. Visible light imagery outperforms IR in each subexperiment and performs better when the expressions in gallery and probe are the same.

8 Same-session versus Time-lapse

This experiment tries to make a more direct comparison of performance in the same-session and time-lapse scenarios. This study uses one probe for each gallery image. The gallery sets (FA|LF) are the same in same-session recognition and time-lapse recognition. The probe set for same-session recognition is made up of images (FA|LM) acquired at about the same time (less than one minute difference) as the probe. The probe set for time-lapse recognition is made up of images (FA|LM) acquired in different weeks from when the gallery images were acquired. The face space is the same as the one used in the first ”same-session” experiments.

We conducted 9 experiments of different time delays for time-lapse recognition and for each there is a corresponding same-session recognition experiment for comparison. The number of gallery and probe pairs for each time delay is 55, 58, 48, 45, 47, 48, 48, 50, and 32.

Figure 8 shows the results for visible light and IR. For both modalities, the same session recognition outperforms time-lapse recognition significantly. Note that for same-session recognition there is no clear advantage between IR and visible light. However, in time-lapse recognition visible light generally outperforms IR. The lower performance during week 10 in time-lapse scenario might be due to the small size of the probe set, which makes it not statistically significant.

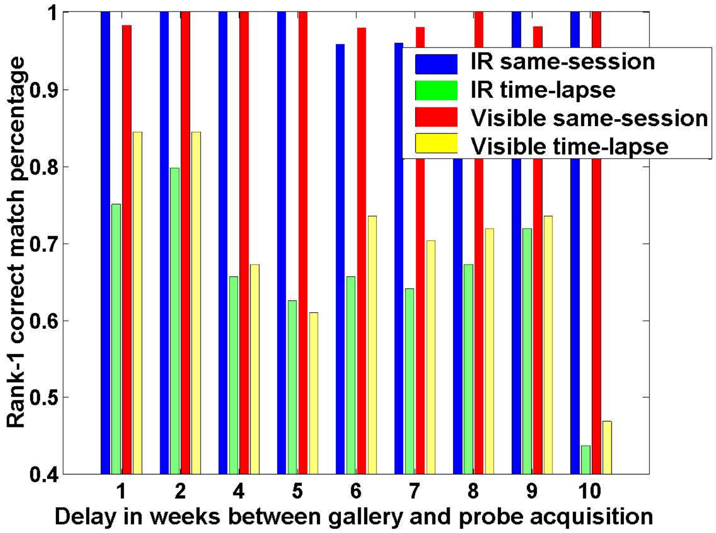


Figure 8: Rank-1 correct match rate for same-session recognition and time-lapse recognition in IR and visible light

9 Assessment of Time Dependency

The face space used in the experiments of this section is the same as the space used in the first ”same-session” experiment. The first experiment is designed to reveal any obvious effect of short-term elapsed time between gallery and probe acquisition on performance. The experiment consists of nine sub-experiments. The gallery set is FA|LF images of session 1. Each of the probes was a set of FA|LF images taken within a single session after session 1 (i.e. sub-experiment 1 used session 2 images in its probes, sub-experiment 2 used session 3 for its probes, and so forth). Figure 9 shows the histogram of the nine rank-1 correct match rates for the nine sub-experiments in IR and visible light imagery. The figure shows differences in performance from week to week, but there is no clearly discernible trend over time in the results. All the rank 1 correct match rates in visible light imagery are higher than in IR. We also conducted a more statistically meaningful experiment by averaging the results of every possible experiment in the context of a specific time delay i.e. to get the result for a one week delay, we used week 1 data as gallery images, week 2 as the probe, week 2 as gallery and week 3 as probe, etc. There is no obviously identifiable trend in this case either.

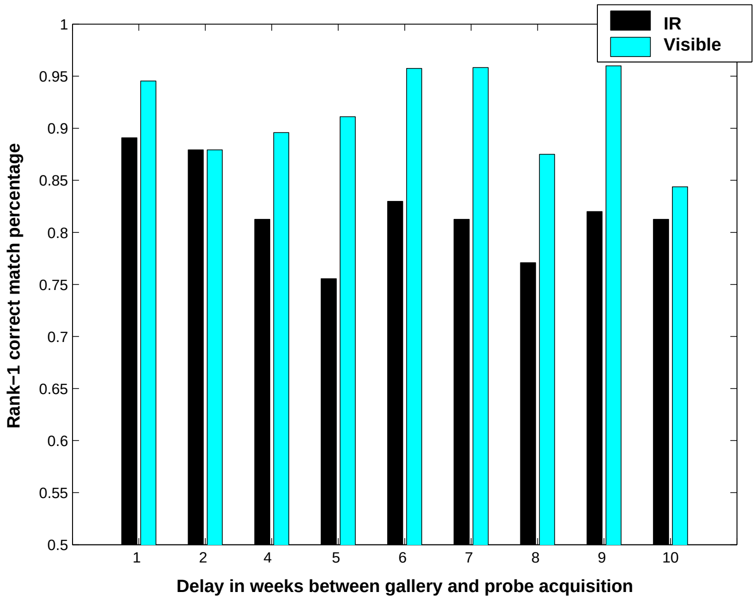


Figure 9: Rank-1 correct match rate for 10 different delays between gallery and probe acquisition in visible light and IR

Another experiment was designed to examine the performance of the face recognition system with a constant delay of one week between gallery and probe acquisitions. It consists of nine sub-experiments: the first used images from session 1 as a gallery and session 2 as probe, the second used session 2 as gallery and session 3 as probe and so on. All images were FA|LF. The rank 1 correct match rates for this batch of experiments appear in Figure 10. We note an overall higher level of performance with one week of time lapse than with larger amounts of time. The visible light imagery outperforms IR in 7 of the 8 sub-experiments.

Together with the time-lapse recognition experiment in Section 7, these experiments show that delay between acquisition of gallery and probe images causes recognition performance to degrade. The one overall surprising result from these experiments is that visible light imagery outperforms

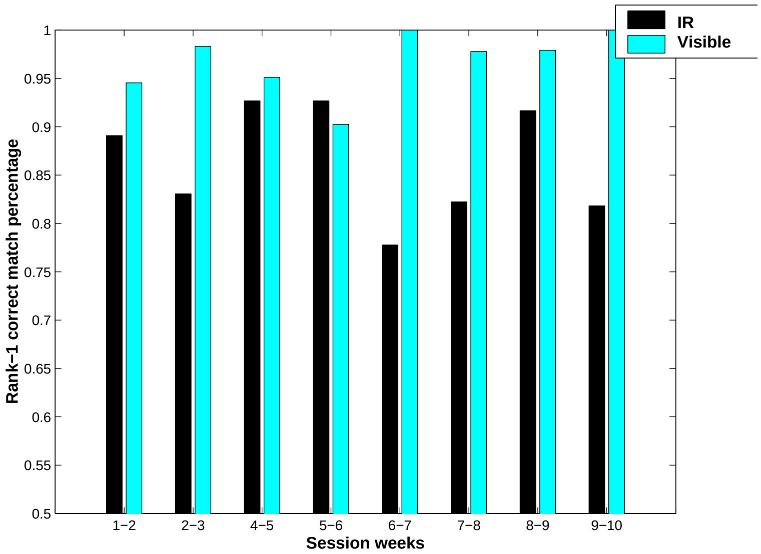


Figure 10: Rank-1 correct match rate for experiments with gallery and probe separated by one week in visible light and IR

IR in the context of time-lapse.

10 Sensitivity to Eye Center Location

We manually located eye centers in visible light and IR images for normalization. It is possible that the errors in eye center location could affect the recognition performance differently in visible light and IR, especially considering that the IR imagery is more vague than visible light imagery and the original resolution for IR is 320 x 240 versus 1600x1200 for the visible light image. This is potentially an important issue when comparing the performance of IR and visible light imagery.

We did a random replacement of the current manually-marked eye centers by another point in a 3x3 (pixel) window, which is centered at the manually-marked position. This is very close to the possible human error when images are truthwritten. The time-lapse recognition results by using images normalized with the randomly perturbed eye centers are shown in Table 5.

When Table 5 and Table 3 are compared, one conclusion is that IR is more sensitive to eye center locations. The correct recognition rates drop significantly compared to the performance where the manually located eye centers are used. For visible light imagery in time-lapse scenario, the performance decrease is at most slight. This suggests that marking eye centers in IR might be harder to do accurately than marking eye centers in visible light, and that this might have affected IR accuracy relative to visible light accuracy in our experiments. Selinger and Socolinsky [13] look at automated eye center location and also report finding greater error for thermal imagery than for visible-light imagery. However, they also find relatively smaller differences in recognition performance than we found, although differences in data set and algorithm complicate a direct comparison.

Probe Gallery	FA|LM	FA|LF	FB|LM	FB|LF
FA|LM	0.67 (52) 0.90 (46)	0.65 (44) 0.91 (54)	0.62 (58) 0.71 (55)	0.57 (59) 0.71 (54)
FA|LF	0.68 (40) 0.91 (50)	0.69 (56) 0.92 (27)	0.60 (55) 0.74 (33)	0.62 (61) 0.72 (44)
FB|LM	0.64 (61) 0.75 (56)	0.67 (60) 0.81 (45)	0.65 (62) 0.86 (49)	0.69 (57) 0.84 (50)
FB|LF	0.63 (57) 0.74 (51)	0.62 (57) 0.78 (40)	0.63 (62) 0.88 (33)	0.65 (55) 0.89 (47)

Table 5: Rank 1 correct match percentage for time-lapse recognition in IR (top) and visible light (bottom). Eye center is randomly replaced by a point in a 3x3 window that is centered at the manually-located eye center

11 Training Set Effects

The discussion of this section assumes that the training set differs from the gallery set. The training set defines the subspace in which recognition will be performed. While the lack of exploitation of class label information means that PCA does not enforce good separability among classes, an implicit assumption of the technique is that between-subject variation is much larger than within-subject variation, i.e. that subjects are well clustered in the projection space and well separated from other subject clusters. PCA's property of variance maximization for a fixed dimension under the basis orthonormality constraint tends to preserve class separation. If training images are representative of the subject classes, separability between subjects in the gallery should also be preserved. Hence, PCA can be argued to provide an efficient representation for images while preserving separability.

Training set size determines the dimensionality of the subspace. Apparently, the more sparsely the gallery images are distributed, the higher separability the recognition system can achieve. Generally, by increasing the dimensionality, the images should be more sparsely distributed.

Hence, there are two schemes of selecting training sets to enhance recognition performance. One is to make the training set more representative of the gallery; the other is to enlarge the training set size in order to increase the dimensionality of the subspace.

There are two kinds of bias associated with training set design. One magnifies within-class variance, which could degrade the performance; that is, the recognition system develops a harmful separability that spreads out the instances of the same subject in the face space. The other bias can enhance the separability and improve the performance if the training set includes gallery images (subjects); the recognition system can specially maximize the variance of the images in the gallery, which of course provides high performance given the same size of training set, but it is not guaranteed to work well when applied to a new gallery. In other words, it is not possible in real applications. Hence, the two biases are both undesirable.

If there are $N > 2$ images in the training set then S_T in Equation 1 has rank

$$r = min(n, N-1) \tag{2}$$

where n is the image space dimension and only r unique eigenvectors can be extracted from S_T. This means that the subspace in which the N images are most separated is r dimensional. By increasing the size of the training set, the maximum feature space dimension grows. This means that the training set becomes more and more sparsely distributed if we retain all the eigenvectors to span the subspace. It is particularly useful to our experimental design, which maps each probe image to one gallery image based on their distances in the face space.

For one time-lapse recognition with FA$|$LF images in the first session as the gallery set and FA$|$LF images in the second to the tenth sessions as the probe set, we examined the eigenvector selection results for IR and visible light images.

For IR, we find that dropping any of the first 10 eigenvectors will degrade the performance. A possible reason is that in IR face images, there is no significant irrelevant variance like the lighting in visible light images and the first eigenvectors can well describe the true variance between images. When retaining 94% of eigenvectors by removing the last eigenvectors, the performance reaches maximum performance of 82.8%, compared with 81.2% when all eigenvectors are retained. This shows that these last eigenvectors encode noise only.

For visible light, dropping the first 2 eigenvectors increases the performance to a peak of 92.6% from 91.4%. It is possible that some significant irrelevant variance, like lighting, is encoded in these eigenvectors. With these two eigenvectors dropped, We find that retaining about 80% of the eigenvectors by removing the last eigenvectors makes the performance increase to 94.4%, which shows that these last eigenvectors are noisy and undermine the performance.

12 Statistical Test on Conditions

In Table 3, probe pairs of differing facial expressions and lighting conditions are analyzed for illumination and facial expression impact. Pairs with the same facial expression but different lighting condition reveal the illumination impact given a gallery of the same facial expressions. Pairs of the same lighting condition but varied facial expression generate facial expression impact in a gallery of constant lighting conditions. Essentially, we make a comparison of the response of matched pairs of subjects, using dichotomous scales, i.e. subjects are grouped into only two categories, correct/incorrect match at rank 1. Hence we choose McNemar’s test [23].

12.1 Illumination Impact

The null hypothesis H_0 is that *there is no difference in performance based on whether the lighting condition for the probe image acquisition is matched to the lighting condition for the gallery image acquisition*. The corresponding $p-$values are reported in Table 6. For IR, what we observed is very likely if the null hypothesis were true and the association between FERET and mugshot lighting conditions for the probe images is NOT significant. There also is no evidence to reject the hypothesis for visible-light imagery. Perhaps the most obvious reason is the relatively small difference between the FERET and the mugshot lighting conditions. Both represent controlled indoor lighting.

Gallery	Probe pair	p-value
FA$\|$LM	FA$\|$LM	0.55
	FA$\|$LF	0.18
FA$\|$LF	FA$\|$LM	0.50
	FA$\|$LF	0.85
FB$\|$LM	FB$\|$LM	0.50
	FB$\|$LF	0.32
FB$\|$LF	FB$\|$LM	0.51
	FB$\|$LF	0.47

Table 6: p-values of McNemar's test for the impact of lighting change in visible light (bottom) and IR (top)

12.2 Facial Expression Impact

The null hypothesis H_0 is that *there is no difference in performance based on whether the facial expression for the probe image acquisition is matched to the facial expression for the gallery image acquisition*. The corresponding $p-$values are reported in Table 7. For visible light imagery, all $p-$values are 0, which means that the null hypothesis is unlikely to be true according to what we observed, i.e. the performance is highly dependent on whether the facial expression for the probe image acquisition is matched to the facial expression for the gallery image acquisition. For IR in the group which used neutral expression as gallery, we have the same conclusion as the visible light imagery. But for IR with a smiling expression as gallery, we failed to reject the hypothesis, which means the expression change impact may not be significant in this scenario.

Gallery	Probe pair	p-value
FA|LM	FA|LM	0.01
	FB|LM	0.00
FA|LF	FA|LF	0.00
	FB|LF	0.00
FB|LM	FB|LM	0.23
	FA|LM	0.00
FB|LF	FB|LF	0.92
	FA|LF	0.00

Table 7: p-values of McNemar's test for the impact of expression change in visible light (bottom) and IR (top)

13 Combination of Visible Light and IR

Table 3 shows that visible light imagery is better than IR in time-lapsed recognition, but the sets of mismatched probes of the two classifiers do not necessarily overlap. This suggests that these two modalities potentially offer complementary information about the probe to be identified, which could improve the performance. It is possible to realize sensor fusion on different levels: sensor data level fusion, feature vector level fusion, and decision level fusion [24]. Since these classifiers yield decision rankings as results, we consider that fusion on the decision level has more potential applications. Our fusion process is divided into the following two stages [24]:

1. Transformation of the score

If the score functions yield values which are not directly comparable, for example, the distance in infrared face space and the distance in visible light face space, a transformation step is required. There exist several score transformation methods, such as linear, logarithmic and exponential. The purpose of these transformations is , first, to map the scores to the same range of values, and, second, to change the distribution of the scores. For example, the logarithmic transformation puts strong emphasis on the top ranks, whereas lower ranked scores, which are transformed to very high values, have a quickly decreasing influence. This is particularly true in our experiments since the top few matches are more reliable than the later ones.

2. Combination and reordering

For every class in the combination set, a combination rule is applied and the classes are reordered in order to get a new ranking. Kittler *et al.* [25] conclude that the combination rule developed under the most restrictive assumptions, the sum rule, outperformed other classifier combination schemes and so we have used the sum rule for combination in our experiments. We implemented two combination strategies, rank based strategies and score based strategies. The former is to compute the sum of the rank for every class in the combination set. The class with the lowest rank sum will be the first choice of the combination classifier. Though the score transformation is primarily thought for the manipulation of the score based strategies, it may be applied to the ranks (interpreted as scores in this case) too. In this way it is possible to change the influence of the ranks significantly. The score based strategy is to compute the sum of the score (distance) for each class and choose the class with the lowest sum score as the first match.

We first used an unweighted rank based strategy for combination. This approach is to compute the sum of the rank for every gallery image. On average, for each probe there are 10-20 rank sum ties (64 gallery images). Since the visible light imagery is more reliable based on our experiments in the context of time-lapse, we use the rank of the visible light imagery to break the tie. The top of each item in Table 8 shows the combination results using this approach. Only in 2 out of 16 instances is the visible light alone slightly better than the combination. The combination classifier outperforms IR and visible light in all the other cases.

For each individual classifier (IR or visible light), the rank at which all probes are correctly identified is far before rank 64 (64 gallery images). Hence, the first several ranks are more useful than the later ranks. We logarithmically transformed the ranks before combination to put strong emphasis on the first ranks and have the later ranks have a quickly decreasing influence. The middle of each item in Table 8 shows the results of this approach. The combiner outperforms visible light and IR in all the sub-experiments and is better than the combiner without rank transformation.

Second, we implemented a score based strategy. We use the distance between the gallery and probe in the face space as the score, which provides the combiner with some additional information that is not available in the rank based method. It is necessary to transform the distances to make them comparable since we used two different face spaces for IR and visible light. We used linear transformation, which maps a score s in a range of $I_s = [smin, smax]$ to a target range of $I_{s'} = [0, 100]$. Then we compute the sum of the transformed distances for each gallery and the one with the smallest sum of distances will be the first match. The bottom entry of each item in Table 8 shows the results. The score based strategy outperforms the rank based strategy and improves the performance significantly compared with either of the individual classifiers (IR and visible light). This shows that it is desirable to have knowledge about the distribution of the distances and the discrimination ability based on the distance for each individual classifier (IR or visible light). This allows us to change the distribution of the scores meaningfully by transforming the distances before combination. This combination strategy is similar to that used by Chang *et al.* [26] in a study of 2D and 3D face recognition.

Probe Gallery	FA|LM	FA|LF	FB|LM	FB|LF
FA|LM	0.91 (25) 0.93 (26) 0.95 (24)	0.95 (23) 0.96 (24) 0.97 (21)	0.83 (45) 0.85 (47) 0.90 (46)	0.81 (44) 0.85 (47) 0.90 (45)
FA|LF	0.91 (18) 0.92 (24) 0.95 (20)	0.93 (19) 0.94 (27) 0.97 (20)	0.85 (41) 0.87 (44) 0.91 (39)	0.83 (23) 0.84 (35) 0.90 (24)
FB|LM	0.87 (20) 0.88 (22) 0.91 (27)	0.92 (34) 0.92 (40) 0.94 (32)	0.85 (23) 0.87 (32) 0.92 (25)	0.86 (32) 0.88 (32) 0.92 (31)
FB|LF	0.85 (43) 0.87 (33) 0.87 (40)	0.87 (40) 0.88 (37) 0.91 (44)	0.88 (12) 0.90 (17) 0.93 (20)	0.90 (36) 0.91 (38) 0.95 (37)

Table 8: Rank 1 correct match percentage for time-lapse recognition of combining IR and visible light. Top: simple rank based strategy; Middle: rank based strategy with rank transformation; Bottom: score based strategy. Row indicates gallery and column indicates probe.

A similar experiment using spring 2003 data as the testing images was conducted applying score-based strategy and the result is reported in Table 9. Again, it improves the performance significantly compared with either of the individual classifiers (IR and visible light).

Gallery \ Probe	FA|LM	FA|LF	FB|LM	FB|LF
FA|LM	0.91 (47)	0.90 (70)	0.80 (134)	0.80 (119)
FA|LF	0.91 (100)	0.91 (110)	0.78 (99)	0.79 (116)
FB|LM	0.85 (101)	0.85 (106)	0.87 (99)	0.87 (73)
FB|LF	0.82 (120)	0.86 (84)	0.87 (119)	0.87 (93)

Table 9: Rank 1 correct match percentage for time-lapse recognition of combining IR and visible light using score based strategy.

Multi-modalities versus multi-samples

From a cost perspective, a multiple sample approach (multiple samples of the same modality, e.g. two visible light face images) will most likely be cheaper than a multiple modal approach (visible light and infrared). Hence, it is particularly important to determine if a multiple modal approach is superior to a multiple sample approach for performance. The following experiment shows that the improvement by combining visible light and infrared modalities is not due purely to using multiple probe images.

For one time-lapse experiment, we use two probe images per modality and combine the decisions using a score based strategy. The results is shown in Table 10.

Modality \ Condition	FA|LM	FB|LF	FA|LF	FA|LM + FB|LF	FA|LM FA|LF
IR	0.92	0.73	0.92	0.90	0.93
Visible	0.81	0.73	0.82	0.85	0.85
IR + Visible	0.95	0.97	0.90	N/A	N/A

Table 10: Top match scores of one time-lapse experiment using one and two probe images; the two probe images either come from two different modalities (IR + Visible) or from the same modality but under two different conditions (FA|LM + FB|LF and FA|LM + FA|LF).

Notice that the performance is worse for combining FA|LM and FB|LF than FA|LM alone. For infrared, the top match score for combining FA|LM and FB|LF probes is 0.85, and 0.85 for combining FA|LM and FA|LF. The scores for FA|LM, FB|LF and FA|LF alone are 0.81, 0.73

and 0.82, respectively. The scores for combining infrared and visible light (also two probes) in FA$|$LM, FA$|$LF and FB$|$LF are 0.95, 0.97 and 0.90, respectively, which are significantly better than combining two probes of the same modality.

14 Comparison of PCA and FaceIt

FaceIt [3] is a commercial face-recognition algorithm that performed well in the 2002 Face Recognition Vendor Test[27]. We use FaceIt results to illustrate the importance of combined IR-plus-visible-light face recognition. We used FaceIt G3 and G5 technologies. The latter is the latest version.

Figure 11 shows the CMC curves for a time-lapse recognition with FA$|$LF images in the first session as the gallery set and FB$|$LM images in the second to the tenth sessions as the probe set by FaceIt and PCA. Note that the fusion method is score-based. We notice that FaceIt G3 and G5 outperform PCA in visible light imagery and IR individually. However, the fusion of IR and visible light can easily outperforms either modality alone by PCA or FaceIt G3. We should take into account the training set PCA used when making this comparison. Given an extremely large unbiased training set which is not often practical or efficient, PCA might eventually outperform FaceIt in visible light imagery.

15 Conclusions

In same session recognition, neither modality is clearly significantly better than the other. In time-lapse recognition, the correct match rate at rank 1 decreased for both visible light and IR. In general, delay between acquisition of gallery and probe images causes recognition system performance to degrade noticeably relative to same-session recognition. More than one week's delay yielded poorer performance than a single week's delay. However, there is no clear trend, based on

[3] http://www.identix.com/products/

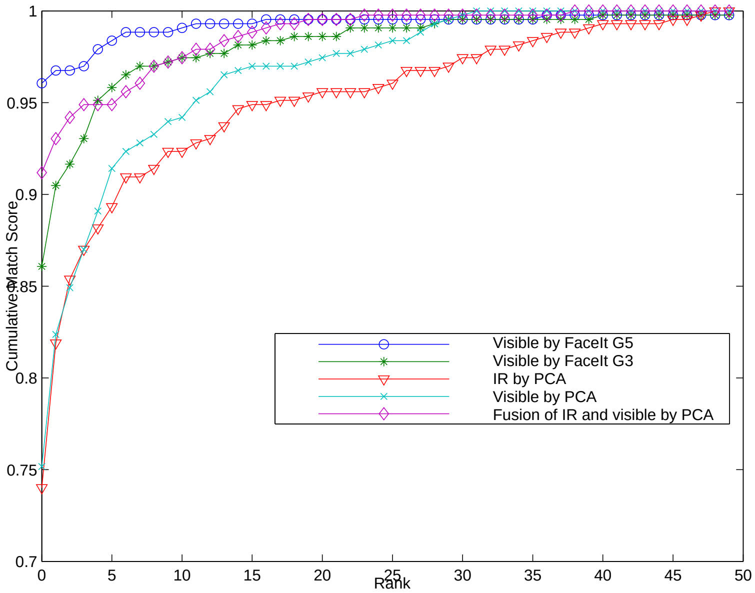


Figure 11: CMC curves of time-lapse recognition using PCA and FaceIt in visible light and IR

the data in this study, that relates the size of the delay to the performance decrease. A longer-term study may reveal a clearer relationship. In this regard, see the results of the Face Recognition Vendor Test 2002 [27].

In time-lapse recognition experiments, we found that: (1) PCA-based recognition using visible light images performed better than PCA-based recognition using IR images, (2) FaceIt-based recognition using visible light images outperformed PCA-based recognition on visible light images, PCA-based recognition on IR images, and the combination of PCA-based recognition on visible light and PCA-based recognition on IR images.

Perhaps the most interesting conclusion suggested by our experimental results is that visible light imagery outperforms IR imagery when the probe image is acquired at a substantial time lapse from the gallery image. This is a distinct difference between our results and those of others [4] [5] [6], in the context of gallery and probe images acquired at nearly the same time. The issue of variability in IR imagery over time certainly deserves additional study. This is especially important because most experimental results reported in the literature are closer to a same-session scenario than a time-lapse scenario, yet a time-lapse scenario may be more relevant to most imagined applications.

Our experimental results also show that the combination of IR plus visible light can outperform either IR or visible light alone. We find that a combination method that considers the distance values performs better than one that only considers ranks.

The likely reason for the success of the technique stems from the fact that face recognition systems depend on accurate localization of facial features, in particular the eyes. The incorporation of multiple images effectively reduces localization errors via averaging. Systems based on eigenface techniques may reap more benefit from such information than other published algorithms such as LFA [28]. It becomes a bottle-neck for infrared face recognition due to the low image resolution and vague infrared imagery.

One could perhaps become confused over the various relative recognition rates reported for visible light and infra-red imaging. The following should be an accurate summary of what is known from various experimental results. Two key elements of experimental design to consider are (a) whether or not there is time lapse between the gallery and probe images, and (b) the degree of lighting variation between gallery and probe images. In studies that use relatively controlled indoor imaging conditions, and for which there is no time lapse between gallery and probe images, the performance from visible and infra-red has been found to be roughly equivalent. In studies that use relatively controlled indoor imaging conditions, and for which there is substantial time lapse between gallery and probe images, the performance from visible light images has been found to exceed that from infra-red images. In studies with greater variation in imaging conditions, such as might occur outdoors with time lapse between gallery and probe, the performance from infra-red images has been found to exceed that from visible light.

The image data sets used in this research will be available to other researchers. See *http://www.nd.edu/˜cvrl* for additional information.

Acknowledgments

The research summarized in this paper was supported by the Defense Advanced Research Projects Agency under Air Force Office of Scientific Research F49620-00-1-0388 and Office of Naval Research grant N00014-02-1-0410, and by the National Science Foundation under grant EIA 01-30839. We thank Dr. Lawrence Wolff and Dr. Diego Socolinsky at Equinox Corporation for providing us with image data and their literature. We thank Dr. Jonathan Phillips at the National Institute of Standards and Technology for his suggestions on this work. We thank Dr. Jaesik Min and Kyong Chang for their hard work on data acquisition and generous help on this work.

References

[1] A. Srivastava and X. Liu, “Statistical hypothesis pruning for identifying faces from infrared images,” in *Image and Vision Computing*, vol. 21, pp. 651–661, July 2003.

[2] F. J. Prokoski, “History, current status, and future of infrared identification,” in *Proceedings IEEE Workshop on Computer Vision Beyond the Visible Spectrum: Methods and Applications*, pp. 5–14, June 2000.

[3] A. Jain, R. Bolle, and S. Pankanti, *Biometrics: Personal Identification in Networked Society*. Kluwer Academic Publishers, 1999.

[4] J. Wilder, P. J. Phillips, C. Jiang, and S. Wiener, “Comparison of visible and infrared imagery for face recognition,” in *2nd International Conference on Automatic Face and Gesture Recognition,Killington,VT*, pp. 182–187, 1996.

[5] D. A. Socolinsky and A. Selinger, “A comparative analysis of face recognition performance with visible and thermal infrared imagery,” in *International Conference on Pattern Recognition*, pp. IV: 217–222, August 2002.

[6] A. Selinger and D. A. Socolinsky, “Appearance-based facial recognition using visible and thermal imagery:a comparative study,” *Technical Report,Equinox corporation*, 2001.

[7] D. Socolinsky, L. Wolff, J. Neuheisel, and C. Eveland, “Illumination invariant face recognition using thermal infrared imagery,” in *Computer Vision and Pattern Recognition*, vol. 1, pp. 527–534, 2001.

[8] P. J. Flynn, K. W. Bowyer, and P. J. Phillips, “Assessment of time dependency in face recognition: An initial study,” in *International Conference on Audio- and Video-based Biometric Person Authentication*, pp. 44–51, June 2003.

[9] X. Chen, P. Flynn, and K. Bowyer, “Visible-light and infrared face recognition,” *Proceedings of the Workshop on Multimodal User Authentication*, pp. 48–55, 2003.

[10] X. Chen, P. Flynn, and K. Bowyer, “Pca-based face recognition in infrared imagery: Baseline and comparative studies,” *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, pp. 127–134, 2003.

[11] P. Phillips, P. Grother, R. Micheals, D. Blackburn, E. Tabassi, and J. Bone, “Frvt 2002: Overview and summary,” March 2003.

[12] D. Socolinsky, A. Selinger, and J. Neuheisel, “Face recognition with visible and thermal infrared imagery,” in *Computer Vision and Image Understanding*, pp. 72–114, 2003.

[13] A. Selinger and D. Socolinsky, “Face recognition in the dark,” in *Conference on Computer Vision and Pattern Recognition Workshop*, vol. 8, pp. 129–134, June 2004.

[14] 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.

[15] P. Phillips, A. Martin, C. Wilson, and M. Przybocki, “An introduction to evaluating biometric systems,” in *Computer*, pp. 56–63, February 2000.

[16] M. Kirby and L. Sirovich, “Application of the Karhunen-Loeve procedure for the characterization of human faces,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 1, pp. 103–108, 1990.

[17] M. Turk and A. Pentland, “Face recognition using eigenfaces,” in *Computer Vision and Pattern Recognition*, 1991.

[18] P. Belhumeur, J. Hespanha, and D. Kriegman, “Eigenfaces vs. fisherfaces: Recognition using class specific linear projection,” in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 711–720, 1997.

[19] W. S. Yambor, B. A. Draper, and J. R. Beveridge, “Analyzing pca-based face recognition algorithms: Eigenvector selection and distance measures,” in *Empirical Evaluation in Computer Vision*, July 2000.

[20] I. Pavlidis, J. Levine, and P. Baukol, “Thermal imaging for anxiety detection,” in *IEEE Workshop on Computer Vision Beyond the Visible Spectrum: Methods and Applications*, pp. 104–109, 2000.

[21] D. Socolinsky and A. Selinger, “Thermal face recognition in an operational scenario,” in *Computer Vision and Pattern Recognition*, 2004, to appear.

[22] D. Socolinsky and A. Selinger, “Thermal face recognition over time,” in *International Conference on Pattern Recognition*, 2004, to appear.

[23] M. Bland, *An Introduction to Medical Statistics*. Oxford University Press, 1995.

[24] B. Achermann and H. Bunke, “Combination of classifiers on the decision level for face recognition,” *Technical Report IAM-96-002, Insitut fur Informatik und angewandte Mathematik, Universitat Bern, Bern*, 1996.

[25] J. Kittler, M. Hatef, R. Duin, and J. Matas, “On combining classifiers,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1992.

[26] K. Chang, K. Bowyer, and P. Flynn, “Multi-modal 2d and 3d biometrics for face recognition,” *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, pp. 187–194, 2003.

[27] *http://www.frvt2002.org*.

[28] P. Grother, “Frvt 2002: Supplemental report,” February 2004.