To appear in the proceedings of the IEEE International Workshop on Analysis and Modeling of Faces and Gestures (AMFG 2003), October 2003, Nice, France.

PCA-Based Face Recognition in Infrared Imagery: Baseline and Comparative Studies

Xin Chen Patrick J. Flynn Kevin W. Bowyer
Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN 46556
{xchen2, flynn, kwb}@nd.edu

Abstract

Techniques for face recognition generally fall into global and local approaches, with the principal component analysis (PCA) being the most prominent global approach. This paper uses the PCA algorithm to study the comparison and combination of infrared and typical visible-light images for face recognition. This study examines the effects of lighting change, facial expression change and passage of time between the gallery image and probe image. Experimental results indicate that when there is substantial passage of time (greater than one week) between the gallery and probe images, recognition from typical visible-light images may outperform that from infrared images. Experimental results also indicate that the combination of the two generally outperforms either one alone. This is the only study that we know of to focus on the issue of how passage of time affects infrared face recognition.

1. Introduction

Although current face recognition systems have achieved good results for images that are taken in a controlled environment, they perform poorly in less controlled situations [1]. Infrared imagery (IR) may offer better performance than other modalities due to robustness to environmental effects and deliberate attempts to obscure identity. The anatomical information which is utilized by IR involves subsurface features thought to be unique to each person. Also, IR provides a capability for identification under all lighting conditions including total darkness [2].

However, face recognition in the thermal domain has received relatively little attention in the literature in comparison with recognition in visible imagery. This is mainly because of the lack of widely available IR image databases. Previous work in this area shows that well-known face recognition techniques, for example PCA, can be successfully applied to IR images, where they perform as well on IR as on visible imagery [3] or even better on IR than on visible imagery [4] [5]. However, in all of these studies [3] [4] [5], the gallery and probe images of a subject were acquired in the same session, on the same day. In our current study, we also examine performance when there is substantial time between gallery and probe.

The performance evaluation methodology employs the concept of a *training* image set used to develop the identification technique, a *gallery* image set that embodies the set of persons enrolled in the system, and a *probe* image set containing images to be identified. We employ a closed universe assumption, i.e. each probe image will have a corresponding match in the gallery. Identification of a probe image yields a ranked set of matches, with rank 1 being the best match.

Socolinsky and Selinger [4] [5] used 91 subjects and the gallery and probe images were acquired within a very short period of time. These experiments will be called *same session recognition*. Experiments in which the probe and gallery images are acquired on different days or weeks will be called *time-lapse recognition*. Socolinsky and Selinger used a sensor capable of imaging both modalities (visible and IR) simultaneously through a common aperture. This enabled them to register the face with reliable visible images instead of IR images. They emphasized the IR sensor calibration and their training set is the same as the gallery set. In their experiments, several face recognition algorithms were tested and the performance using IR appears to be superior to that using visible imagery.

Wilder *et al.* [3] used 101 subjects and the images were acquired without time lapse. They controlled only for expression change. Several recognition algorithms were tested and they concluded that the performance is not significantly better for one modality than for another.

This study examines more varied conditions and uses a relatively larger database, in both the number of images and the number of subjects, compared with the databases used by Wilder *et al.* and Socolinsky *et al.* [3] [4] [5]. We consider the performance of the PCA algorithm in IR, including the impact of illumination change, facial expres-

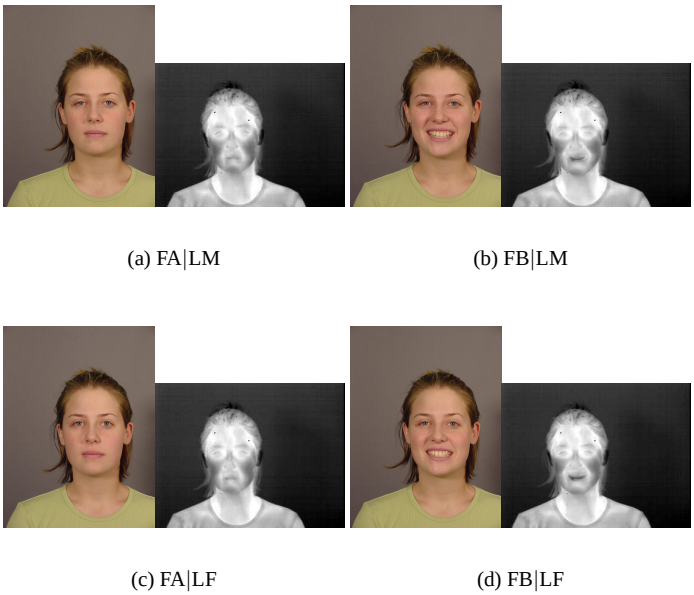
sion change and the short term (minutes) and longer term (weeks) change in face appearance. We also present a comparative study employing visible imagery. Each image involved in the experiment used in one modality has a counterpart (acquired at the same time, under the same conditions and of the same subject) in the other modality. The software suite used in our experiments was developed at Colorado State University [1].

2. Data Collection

Most of the data used to obtain the results in this paper was acquired at University of Notre Dame during 2002, where IR images from 241 distinct subjects were acquired. Each image acquisition session consists of four views with different lighting and facial expressions. Since this data set size is not enough for the training in our experiments, we also used 81 IR and visible-light images of 81 distinct subjects, acquired by Equinox Corporation [6]. Image acquisitions were held weekly for each subject and most subjects participated multiple times. All subjects completed an IRB-approved consent form for each acquisition session. IR images were acquired with a Merlin Uncooled long-wavelength IR camera, which provides a real-time, 60Hz, 12 bit digital data stream, has a resolution of 320×240 pixels and is sensitive in the 7.0-14.0 micron range. Three Smith-Victor A120 lights with 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's 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". 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". For all of these images the subject stood in front of a standard gray background. Due to IR's opaqueness to glass, we asked all subjects to remove eyeglasses during acquisition. According to the lighting and expression, 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). Figure 1 shows one subject in one session under these four conditions.

Selinger and Socolinsky [5] describe in detail the acquisition process of the data collected at Equinox Corporation.

[1] http://www.cs.colostate.edu/evalfacerec/, version 5.0



(a) FA|LM (b) FB|LM

(c) FA|LF (d) FB|LF

Figure 1: Face images in visible and IR under different lighting and facial expression conditions.

The 81 images from Equinox were acquired from 81 distinct subjects, under frontal light while the subject was pronouncing vowels and not wearing glasses. Thus their data set reflects primarily facial expression change.

3 Preprocessing

We located faces manually by clicking on the centers of each eye. The features on a human face are much more vague in IR than those in visible imagery and thus the registration in the following normalization step might not be as reliable in IR as in the visible images. We chose to locate the eye centers since we could not find any other more reliable landmarks on the face. Notice that Socolinsky and Selinger [4] [5] used a sensor capable of capturing simultaneous registered visible and IR, which is of particular significance for their comparison of visible and IR. The fact that they get eye location from visible imagery and use it in IR may make their IR performance better than if they purely used IR.

4 PCA Algorithm

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

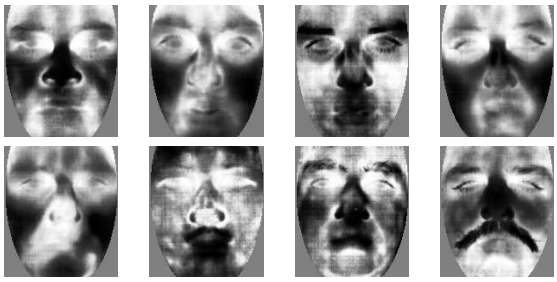
coordinates

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

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

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

where N is the number of sample images, and $\mu \in \mathbb{R}^n$ is the mean image of all samples. An example of a training set and its mean image in IR is shown in Figure 2 (a) and (b), respectively.



(a)

(b)

(c)

Figure 2: (a) Training images: normalized IR images of eight different subjects. (b) Mean image: average of the eight images in (a). (c) Eigenfaces: principle 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 [5]. These face-like eigenvectors are referred to as eigenfaces or principal components. Figure 2 (c) shows the top seven eigenfaces derived from the input images of Figure 2 (a) in decreasing eigenvalue order.

5 Dissimilarity Measures

After deriving a subspace, we project the gallery and probe images onto this space, compute the dissimilarity between the probe image and each of the gallery images, and choose the most similar gallery image as a match of that probe. The relative performance among different dissimilarity measures, such as angle, Euclidean distance, cityblock distance and Mahalanobis distance measures, is consistent throughout the experiments.

Since the contributions of some eigenvectors in describing the variance among images might be much greater than those of some other eigenvectors, it is apparent that the former will dominate the result and the latter could be virtually ignored. In such cases, it is better to use modified eigenspace, i.e. weighting the axes of the space so that all the dimensions are comparable. A bonus of weighting is that more separability is achieved, which can enhance the recognition performance. The "MahCosine" (coined by the CSU software) is the angle measure [7]

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

applied in the weighted space and performs best among all the metrics implemented by the CSU software. Hence, all the subsequent experimental results are based on "MahCosine" metric.

6 Same Session Recognition

We used 83 distinct subjects and four images for each subject acquired within 1 minute with different illumination and facial expressions. 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.

Table 1 shows that there is no consistent difference between the performance of visible and IR. IR is better in six instances, visible is better in four instances and they are the same in two instances. The overall performance for same

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

	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)	

session recognition is high for both IR and visible, and so it is possible that some "ceiling effect" could make it difficult to observe any true difference that might exist.

7 Time-lapse Recognition

Time-lapse recognition experiments use the images acquired in ten acquisition sessions of Spring 2002. Figure 3 depicts the number of the valid subjects in each session. Figure 4 shows the visible 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 images. In particular, the bridge and sides of the nose appear somewhat different in different IR images.

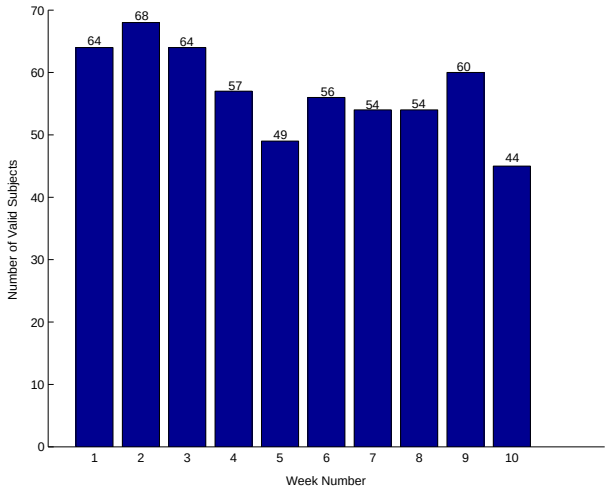
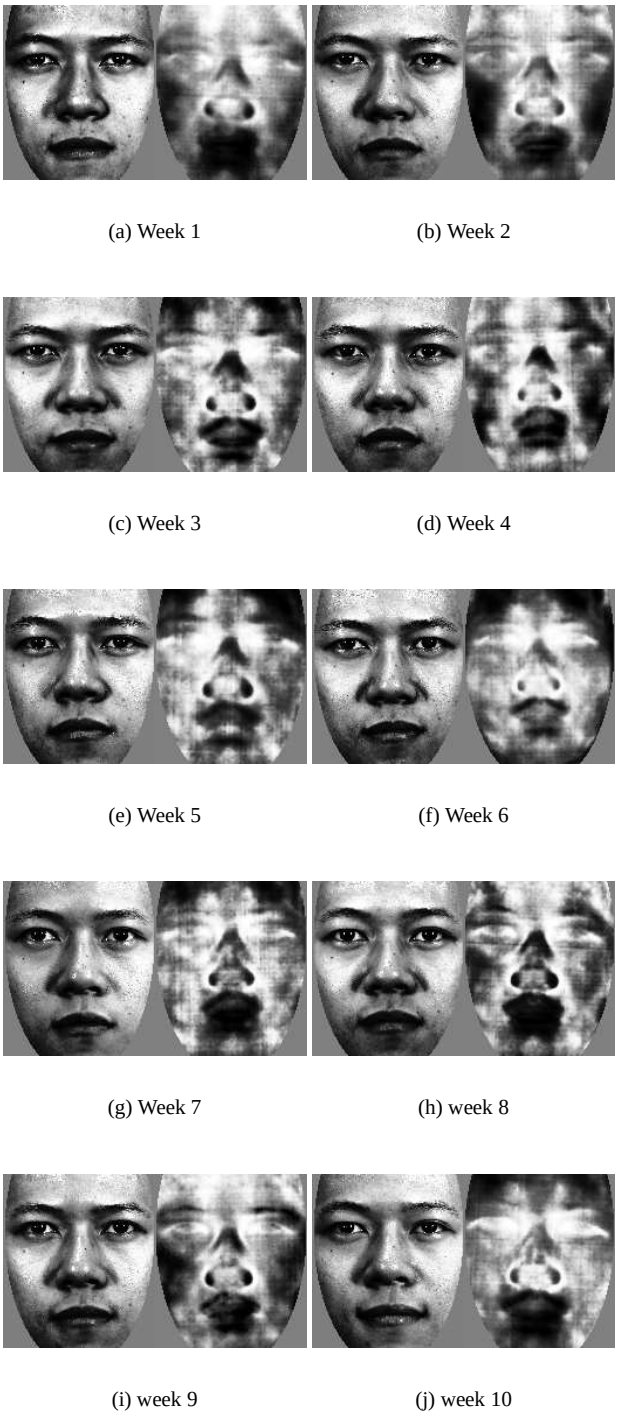


Figure 3: Valid subject counts per session

The scenario for this recognition is a typical enroll-once identification setup. There are 16 experiments based on the exhaustive combinations of gallery and probe sets given the



(a) Week 1 (b) Week 2

(c) Week 3 (d) Week 4

(e) Week 5 (f) Week 6

(g) Week 7 (h) week 8

(i) week 9 (j) week 10

Figure 4: Normalized FA|LM face images of one subject in visible and IR across 10 weeks.

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; each probe set has 431 images from sessions 2-10. The rank-1 correct match percentages are given in Table 2. 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 subspace is derived from a training set of 240 images of 240 distinct subjects.

Table 2: Rank 1 correct match percentage for time-lapse recognition in visible (bottom) and IR (top). Row indicates gallery and column indicates probe.

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

For IR, Table 2 illustrates a striking difference in performance in contrast to same-session recognition results shown in Table 1: the rank 1 correct match rate drops about 20%. The most obvious reason is that the elapsed time caused significant changes among thermal patterns of the same subject; since the probe set size is 431, much larger than 83, the images projected onto the face space became more densely distributed and there is overlap in face space between the clusters of points for different subjects. In addition, it is possible that unreliable registration of the eye centers could have degraded the performance. Table 2 also shows that the performance degrades for visible imagery compared with that in same-session recognition. Visible imagery outperforms IR in 12 of the 16 cases, with IR and visible the same in another two.

8 Combination of Visible and IR

Table 2 shows that visible 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. The performance of either classifier is neither inaccurate (having an individual probability of correct inference less than 0.5) nor highly accurate (having an individual probability of correct inference of greater than 0.95) and so it should be possible to achieve improvement through sensor fusion [8]. Since these classifiers yield decision rankings as results, we first consider fusion on the decision level. Kittler *et al.* [9] 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 first used an unweighted rank based strategy for combination.This approach is to compute the sum of the rank for every gallery image. The gallery image with the lowest rank sum will be the first choice of the combination classifier. However, on average, for each probe there are 10-20 rank sum ties (64 gallery images). Since the visible imagery is more reliable based on our experiments in the context of time-lapse, we use the rank of the visible imagery to break the tie. The top of each item in Table 3 shows the combination results using this approach. Only in 2 out of 16 instances is the visible alone slightly better than the combination. The combination classifier outperforms IR and visible in all the other cases.

For each individual classifier (IR or visible), 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 3 shows the results of this approach. The combiner outperforms visible and IR in all the subexperiments 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. 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 3 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). 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). 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.* [10] in a study of 2D and 3D face recognition.

Table 3: Rank 1 correct match percentage for time-lapse recognition of combining IR and visible. Top: simple rank based strategy; Middle: rank based strategy with rank transformation; Bottom: score based strategy. Row indicates gallery and column indicates probe.

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

9 Assessment of Time Dependency

The first experiment is designed to reveal any obvious effect of 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 5 shows the histogram of the nine rank-1 correct match rates for the nine sub-experiments in IR and visible 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 imagery are higher than in IR.

The second 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 6.

We note an overall higher level of performance with one week of time lapse than with larger amounts of time. The visible imagery outperforms IR in 7 of the 8 subexperiments, shown in in Figure 6.

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. More than one week's delay yielded poorer performance than a single week's delay.

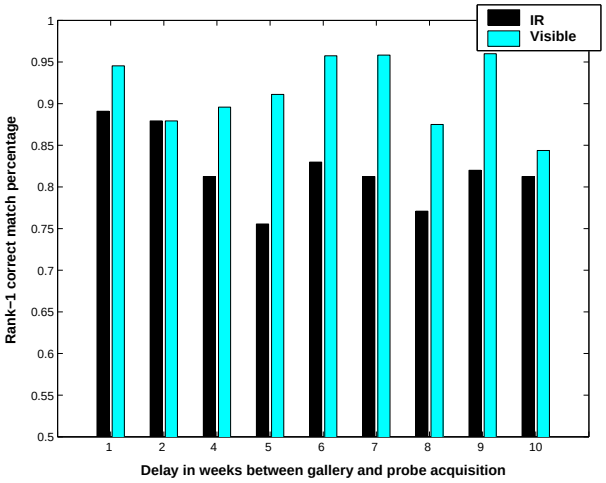


Figure 5: Rank-1 correct match rate for 10 different delays between gallery and probe acquisition in visible and IR

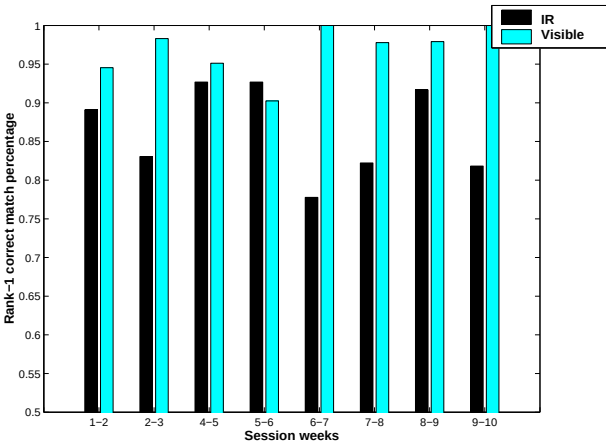


Figure 6: Rank-1 correct match rate for experiments with gallery and probe separated by one week in visible and IR

However, there is no clear trend, using the data in this study, that relates the size of the delay to the performance decrease. This motivates the development of a larger database covering more subjects and a longer period of time. The one overall surprising result from these experiments is that visible imagery outperforms IR in the context of time-lapse.

10 Statistical Test on Conditions

In Table 2, the probe pairs that are of the same facial expression (lighting condition) but different lighting condition (facial expression), given a gallery of the same facial expression (lighting condition), should reveal the illumination (facial expression) impact. Essentially, we make a comparison of the response of matched pairs of subjects, using dichoto-

mous scales, i.e. subjects are grouped into only two categories, correct/incorrect match at rank 1. Hence we choose McNemar's test [11].

10.1 Illumination Impact

Given the null hypothesis being *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 3. 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. However, surprisingly, for visible imagery, there is no evidence to reject the hypothesis either. One reason is that the variance, which is dependent on elapsed-time, dominated over the lighting variance. Another possible reason is that there is not enough difference between FERET and mugshot lighting conditions to produce a noticeable effect. Referring to the images in Figure 1, this explanation seems plausible.

Table 4: p-values of McNemar's test for the impact of lighting change in visible (bottom) and IR (top)

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

10.2 Facial Expression Impact

Given the null hypothesis being *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 4.

For visible 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 imagery. But for IR with a smiling expression as gallery, we failed to reject the hypothesis, which means the expression impact may be significant in this scenario.

Table 5: p-values of McNemar's test for the impact of expression change in visible (bottom) and IR (top)

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

11 Conclusion and Discussion

In same session recognition, neither modality is clearly significantly better than another. In time-lapse recognition, the correct match rate at rank 1 decreased for both visible 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 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 [1].

Perhaps the most interesting conclusion suggested by our experimental results is that visible imagery outperforms IR imagery when the probe image is acquired at a substantial time lapse from the gallery image. This is distinctly different from our experimental results, and those of others [3] [4] [5], 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.

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

Note that the sort of "time lapse" experiment considered here shows very different results than the "same session" experiments. There are at least two important differences. One is that recognition performance is noticeably lower in a time-lapse scenario. Another is that relative performance between modalities observed in a same session scenario does not necessarily hold true in a time-lapse scenario. Both of these are 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.

The image data sets used in this research will eventually

be available to other researchers as part of the Human ID database. See *http://www.nd.edu/~cvrl* for additional information.

Our experimental results motivate the development of a larger IR database covering more subjects and a longer period of time, in order to explore these experimental observations with greater certainty.

Acknowledgments

This work was supported by Office of Naval Research (DARPA) N000140210410 and National Science Foundation EIA 01-30839. We thank Dr. Lawrence Wolff at Equinox Corporation for providing us with image data. We thank Dr. Jonathan Phillips at the National Institute of Standards and Technology for his suggestions on this work. We thank Dr. Jaesik Min for his hard work on data acquisition and his help with McNemar's test.

References

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

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

[3] 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.

[4] 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.

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

[6] *http://www.equinoxsensors.com/products/HID.html*.

[7] 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.

[8] B. Achermann and H. Bunke, "Combination of face classifiers for person identification," *International Conference on Pattern Recognition*, pp. 416–420, 1996.

[9] 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.

[10] K. Chang, K. Bowyer, and P. Flynn, "Multi-modal 2d and 3d biometrics for face recognition," *to appear in IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, 2003.

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