

Thermal Face Recognition Over Time^{*}

Diego A. Socolinsky[†]

Andrea Selinger[‡]

[†]Equinox Corporation
207 East Redwood Street
Baltimore, MD 21202

[‡]Equinox Corporation
9 West 57th Street
New York, NY 10019

{diego, andrea}@equinoxsensors.com

Abstract

We present a comparative study of face recognition performance with visible and thermal infrared imagery, emphasizing the influence of time-lapse between enrollment and testing images. Most previous research in this area, with few exceptions, focused on results obtained when enrollment and testing images were acquired in the same session. We show that the performance difference between visible and thermal recognition in a time-lapse scenario is smaller than previously believed, and in fact is not statistically significant on existing data sets.

1 Introduction

Face recognition with thermal infrared imagery has recently enjoyed renewed interest. While the volume of literature on the subject is notably smaller than that related to visible face recognition, there is nonetheless a steady stream of research [1, 2, 3, 4, 5, 6]. These papers have established that thermal imagery of human faces constitutes a valid biometric signature, though mostly relying on databases limited both in size and variability, due to the expense and complexity of extensive data collection. Early results were based on gallery and probe sets collected indoors during a single session. In that respect, they resemble the **fa/fb** tests in the FERET program [7].

More recently, a study involving imagery collected indoors in a laboratory setting over multiple weeks was presented in [4, 8]. In that study, the authors note that when using a PCA-based recognition system, visible face recognition of time-lapse images yields better results than its thermal counterpart. They go on to conjecture, based on their visual analysis of the thermal imagery, that large variations of the thermal emission patterns of the face over time were responsible for the degraded performance. The current pa-

per seeks to reproduce and extend some of the results in [4, 8]. In particular, we show that while those results are reproducible, it may be premature to attribute the performance difference to a modality-specific phenomenon. The results below demonstrate that a statistically significant performance difference between modalities can be measured when recognition is performed using PCA. However, when a more sophisticated algorithm is used, no such difference is measurable. This indicates that the authors of [4, 8] may have observed a measurement effect, and that the “inherent” value of visible and thermal imagery for time-lapse face recognition under controlled conditions is equivalent.

2 Data Collection and Normalization

The data used in this study was generously provided by the authors of [4, 8]. A complete description of the data collection procedure can be found in the references, and we include a brief summary here. Visible and longwave IR (LWIR) images of 240 distinct subjects were acquired under controlled conditions, over a period of ten weeks. During each weekly session, each subject was imaged under two different illumination conditions (FERET and mugshot), and with two different expressions (“neutral” and “other”). Visible images were acquired in color and a 1200×1600 resolution. Thermal images were acquired at 320×240 resolution and 12 bit depth.

Eye coordinates for all images, both visible and thermal, were manually located by the authors of [4, 8]. These coordinates were used to affinely register the images to a standard geometry with fixed eye locations and image size of 99×132 pixels. All necessary interpolation was performed bilinearly. The visible and thermal cameras were bore-sighted during data collection, therefore eye coordinates on corresponding images may not match exactly, as they had to be manually located in each modality separately. After alignment, all images were masked to remove all but the inner face, excluding ears and hair. Images used for the PCA

^{*}This research was supported by the DARPA Human Identification at a Distance (HID) program, contract # DARPA/AFOSR F49620-01-C-0008.

experiments were further histogram-equalized, in order to match the processing in [4, 8]. Since the other algorithm does its own internal image processing, no equalization was performed on images before recognition.

3 Thermal Infrared Phenomenology

While the nature of face imagery in the visible domain is well-studied, particularly with respect to illumination dependence [9], its thermal counterpart has received less attention. In [4], the authors show some variability in thermal emission patterns during time-lapse experiments, and properly blame it for decreased recognition performance. Figure 1 shows comparable variability within data collected with our own LWIR sensor. Each column shows images acquired in different sessions. It is clear that thermal emission patterns around the eyes, nose and mouth are rather different in different sessions. Such variations can be induced by changing environmental conditions. For example, exposed to cold or wind, capillary vessels at the surface of the skin contract, reducing the effective blood flow and thereby the surface temperature of the face. Also, when a subject transitions from a cold outdoor environment to a warm indoor one, a reverse process occurs, whereby capillaries dilate, suddenly flushing the skin with warm blood in the body’s effort to regain normal temperature. We have no knowledge of the environmental conditions during the data collection by the authors of [4], although we presume that they were fairly constant throughout all sessions.

Additional fluctuations in thermal appearance are unrelated to ambient conditions, but are rather related to the subject’s metabolism. Vigorous physical activity, consumption of food, alcohol or caffeine may all affect the thermal appearance of a subject’s face. Also, high temporal frequency thermal variation is associated with breathing. The nose or mouth will appear cooler as the subject is inhaling and warmer as he or she exhales, since exhaled air is at core body temperature, which is several degrees warmer than skin temperature.

Much like recognition from visible imagery is affected by illumination, recognition with thermal imagery is affected by a number of exogenous and endogenous factors. And while the appearance of some features may change, their underlying shape remains the same and continues to hold useful information for recognition. Thus, much like in the case of visible imagery, different algorithms are more or less sensitive to image variations. Proper compensation for those variations is a critical step of any successful face (or generally object) recognition algorithm, regardless of modality. Clearly, the better algorithms for thermal face recognition will perform equivalent compensation on the infrared imagery prior to comparing probe and gallery samples.

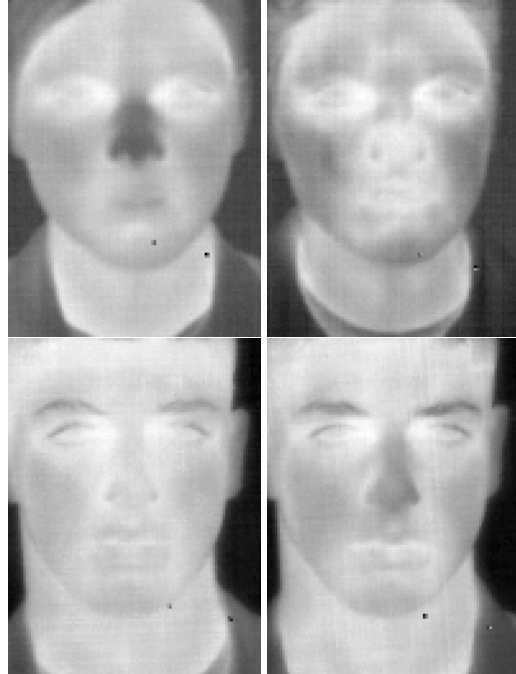


Figure 1: Variation in facial thermal emission from two subjects in different sessions. Left column is the enrollment image and right column is the test image.

4 Algorithms Tested

We performed experiments with two different algorithms in each of the two modalities: PCA with Mahalanobis angle distance and the (blinded for review) algorithm. The first is a standard algorithm with performance evaluations widely available in the literature, including [2], in which the authors present a comprehensive analysis of its performance on visible and thermal infrared imagery in a same-session recognition scenario. The second one is a commercial algorithm made available for testing in binary form.¹

The training set for both algorithms was completely disjoint from gallery and probe images, provided by the authors of [4], in time, space and subjects. That is, the training set was collected at an earlier time, in a different location and used a disjoint set of subjects. This insures that the results reported below are indicative of real-world performance. We should also note that the training set was different from that used in [4], since their complete training set was not available to us. We chose to use a larger set of images collected over the last several years with our own visible and thermal cameras. This further increases the realism of the results, since one cannot usually expect to have

¹This algorithm was made available for testing purposes at [http://\(blinded for review\)](http://(blinded for review)).

training imagery from the same camera as the testing imagery. As a result of these divergences from [4], our PCA results are somewhat different. However, the qualitative nature of the results, as seen below, agrees strongly with those of [4].

5 Experimental Results and Discussion

In order to evaluate recognition performance with time-lapse data, we performed the following experiments. The first-week frontal illumination images of each subject with neutral expression were used as the gallery. Thus the gallery contains a single image of each subject. For all weeks, the probe set contains neutral expression images of each subject, with mugshot lighting. The number of subjects in each week ranges from 44 to 68, while the number of overlapping subjects with respect to the first week ranges from 31 to 56. We computed top-rank recognition rates for each of the weekly probe sets with both modalities and algorithms. The results are shown in Figures 2 and 3. Note that the first data point in each graph corresponds to same-session recognition performance.

Focusing for a moment on the performance curves, we notice that there is no clear trend for either visible or thermal modalities, encompassing weeks two through ten. That is, we do not see a clearly decreasing performance trend for either modality. This appears to indicate that whatever time-lapse effects are responsible for performance degradation versus same-session results are roughly constant over the ten week trial period. Other studies have shown that over a period of years face recognition performance degrades linearly with time [10]. Our observation here is simply that the slope of the degradation line is small enough as to be nearly flat over a ten week period (except for the same-session result, of course). Following that observation, we assume that weekly recognition performances for both algorithms and modalities are drawn independently and distributed according to a (locally) constant distribution, which we may assume to be Gaussian. Using this assumption, we estimate the standard deviation of that distribution, and plot error bars at two standard deviations.

Figure 2 shows the week by week recognition rates using PCA-based recognition. We see that, consistently with the results in [4, 8], thermal performance is lower than visible performance. In fact, for at least six out of nine time-lapse weeks that difference is statistically significant. Table 1 shows mean recognition rates over weeks two through nine for each algorithm and modality. As shown in the last column, we see that mean visible performance is higher than the mean thermal performance by 2.17 standard deviations. This clearly indicates that thermal face recognition with PCA under a time-lapse scenario is significantly less

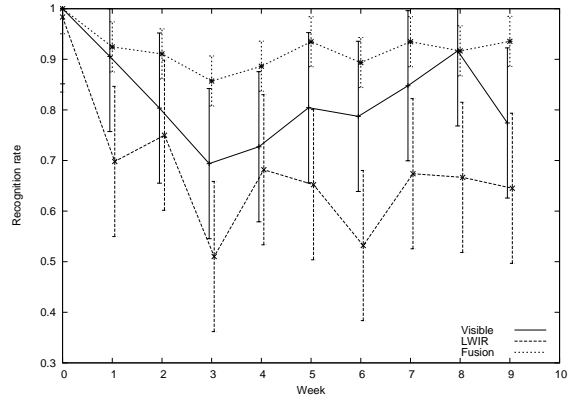


Figure 2: Top-rank recognition results for visible, LWIR and fusion as a function of weeks elapsed between enrollment and testing, using PCA. Note that the x -coordinate of each curve is slightly offset in order to better present the error bars.

reliable than its visible counterpart.

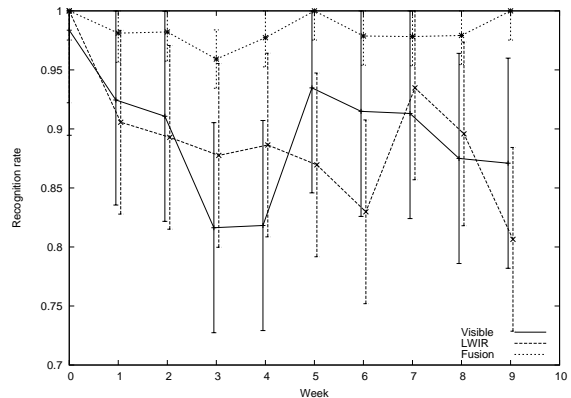


Figure 3: Top-rank recognition results for visible, LWIR and fusion as a function of weeks elapsed between enrollment and testing, using (blinded for review) algorithm. Note that the x -coordinate of each curve is slightly offset in order to better present the error bars.

Turning to Figure 3, we see the results of running the same experiments with the (blinded for review) algorithm. Firstly, we note that overall recognition performance is markedly improved in both modalities. More importantly, we see that weekly performance curves for both modalities cross each other multiple times, while remaining within each other's error bars. This indicates that the performance difference between modalities using this algorithm is not statistically significant. In fact, looking at Table 1, we see that the difference between mean performances for the modalities is only 0.21 standard deviations, hardly a significant result. We should also note that the mean visible time-lapse performance with this algorithm is 88.65%,

	Vis	LWIR	Fusion	Δ/σ Vis vs LWIR
PCA	80.67	64.55	91.04	2.17
(blinded)	88.65	87.77	98.17	0.21

Table 1: Mean top-match recognition performance for time-lapse experiments with both algorithms.

compared to approximately 86.5% for the FaceIt algorithm, as reported in [4]. This shows that the (blinded for review) algorithm is competitive with the commercial state-of-the-art on this data set, and therefore provides a fair means of evaluating thermal recognition performance, as using a poor visible algorithm for comparison would like thermal recognition appear better.

Figures 2 and 3, as well as Table 1 also show the result of fusing both imaging modalities for recognition. Following [2] and [4] we simply add the scores from each modality to create a combined score. Recognition is performed by a nearest neighbor classifier with respect to the combined score. As many previous studies have shown [1, 2, 4], fusion greatly increases performance.

6 Conclusions

The main conclusion of this paper is that one must be cautious when evaluating the value of an imaging modality for a specific recognition task. Ideally, this question should be framed as that of estimating the Bayes optimal error for a classification problem. Inevitably, that estimate is based on an empirical measure of performance which inextricably tied to a particular classifier. While such an estimate can provide us with a valuable upper bound on the Bayes error, it cannot separate classifier effects from data-specific behavior. In this case, we show that while the results in [4] are reproducible, they do not imply that time-lapse face recognition with thermal infrared imagery is inferior to that performed with visible imagery. We have shown by example that, at least on this data set, the Bayes errors for each modality are comparable. Are more detailed analysis will surely require a much larger pool of subjects.

Based on the preceding analysis, and recent results by the authors on time-lapse recognition with a more challenging, larger and diverse data set [11], we firmly believe that the use of thermal imagery of faces for biometric authentication is not only viable, but in certain circumstances even preferable over the use of visible images. Without a doubt, the used of fused visible and thermal imagery provides a level of performance not attainable by either alone.

References

[1] D. A. Socolinsky and A. Selinger, “A comparative analysis of face recognition performance with visible

and thermal infrared imagery,” in *Proceedings ICPR*, Quebec, Canada, August 2002.

- [2] D.A. Socolinsky and A. Selinger, “Face recognition with visible and thermal infrared imagery,” *Computer Vision and Image Understanding*, July - August 2003.
- [3] Joseph Wilder, P. Jonathon Phillips, Cunhong Jiang, and Stephen Wiener, “Comparison of Visible and Infra-Red Imagery for Face Recognition,” in *Proceedings of 2nd International Conference on Automatic Face & Gesture Recognition*, Killington, VT, 1996, pp. 182–187.
- [4] X. Chen, P. Flynn, and K. Bowyer, “Visible-light and infrared face recognition,” in *Proceedings of the Workshop on Multimodal User Authentication*, Santa Barbara, CA, December 2003, to appear.
- [5] B. Abidi, “Performance comparison of visual and thermal signatures for face recognition,” in *The Biometrics Consortium Conference*, Arlington, VA, September 2003.
- [6] 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*, Hilton Head, 2000.
- [7] P. Jonathon Phillips, Hyeonjoon Moon, Syed A. Rizvi, and Patrick J. Rauss, “The FERET Evaluation Methodology for Face-Recognition Algorithms,” Tech. Rep. NISTIR 6264, National Institute of Standards and Technology, 7 Jan. 1999.
- [8] X. Chen, P. Flynn, and K. Bowyer, “PCA-based face recognition in infrared imagery: Baseline and comparative studies,” in *International Workshop on Analysis and Modeling of Faces and Gestures*, Nice, France, October 2003.
- [9] Yael Adini, Yael Moses, and Shimon Ullman, “Face Recognition: The Problem of Compensating for Changes in Illumination Direction,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 721–732, July 1997.
- [10] P. J. Phillips, D. Blackburn, M. Bone, P. Grother, R. Micheals, and E. Tabassi, “Face recognition vendor test 2002 (FRVT 2002),” Available at <http://www.frvt.org/FRVT2002/default.htm>, 2002.
- [11] Authors, “Thermal face recognition in an operational scenario,” in *Submitted to CVPR 2004*.