PREDICTING BIOMETRIC FACIAL RECOGNITION FAILURE WITH SIMILARITY SURFACES AND SUPPORT VECTOR MACHINES

W. J. Scheirer[1,2]*, A. Bendale*[1]*, and T. E. Boult*[1,2]

[1]VAST Lab, University of Colorado at Colorado Springs and [2]Securics, Inc.
Colorado Springs, CO.

ABSTRACT

The notion of *quality* in biometric system evaluation has often been restricted to raw image quality, with a prediction of failure leaving no other option but to acquire another sample image of the subject at large. The very nature of this sort of failure prediction is very limiting for both identifying situations where algorithms fail, and for automatically compensating for failure conditions. Moreover, when expressed in a ROC curve, image quality paints an often misleading picture regarding its potential to predict failure.

In this paper, we extend previous work on predicting algorithmic failures via similarity surface analysis. To generate the surfaces used for comparison, we define a set of new features derived from distance measures or similarity scores from a recognition system. For learning, we introduce support vector machines as yet another approach for accurate classification. A large set of scores from facial recognition algorithms are evaluated, including EBGM, Robust PCA, Robust Revocable PCA, and a leading commercial algorithm. Experimental results show that we can reliably predict biometric system failure using the SVM approach.

1. INTRODUCTION

Failure in a biometric system may be defined in a variety of different ways. The most common instance of failure occurs at the recognition portion of a system - a legitimate user may not be able to achieve a good enough similarity score to match their entry in the gallery. If the parameters of the system call for rank 1 recognition, then failure occurs when the top match in the gallery does not correspond to the probe. This notion of failure can be generalized to rank n recognition, where failure occurs if all of the top n matches do not correspond to the probe. All of these failure instances correspond to statistical Type II error - false rejection. Similarly, for rank n recognition, it is possible that an impostor has achieved a match within the gallery - a very dangerous failure condition. This corresponds to statistical Type I error - false accept. Beyond recognition, failure can also occur at a failure prediction level of the system. Some systems incorporate image quality metrics to determine when a recognition system will fail. If image quality is assessed to be poor enough to induce failure, but the recognition system makes a correct match with the submitted image, then failure has occurred within the failure prediction system itself.

The question of "Why predict failure?" in a biometric system is intriguing for a variety of reasons. Failure prediction serves as another metric of "quality". Often, we are interested in feedback to improve a sensor or collection system, and other times, it is the algorithm itself we wish to evaluate and improve. Moreover, failure prediction can aid in the appropriate weighting of results and features for multi-biometric fusion approaches. Traditional evaluation of biometric system quality has relied on image quality to determine system performance.

Probably the best-known existing work on biometric quality and reliability is [1]. In that work, a reliability measure for fingerprint images is introduced, and is shown to have a strong correlation with recognition performance. Various multi-finger fusion techniques have been developed using that quality/reliability measure. The work, while excellent in its overall analysis, presents its results by separating data (probes and galleries) into separate quality bins and then analyzing the performance of each subset separately, e.g., presenting a Receiver Operator Characteristic (ROC) curve format, and showing that the ROC curve for higher quality data was above that for lower quality data.

Further in the quality space, [2] describes methods for the quantitative evaluation of systems that produce quality scores for biometric data. This paper promotes somewhat confusing boxplots and error versus reject curves, targeting specific matching error rates, over detection error trade-off characteristics for quality algorithm assessment. The results of that paper are indeed important for quality assessment, but do not tell us very much about the failure conditions of biometric matching algorithms or the systems that they are components of. While image quality overall predicts success, recognition of failure for an individual image was not addressed. A more comprehensive systems level approach can expose failure conditions that are not caused by what is traditionally classified as poor image quality.

This work was supported by DHS SBIR "Optimizing Remote Capture of Biometrics for Screening Processes," Award Number NBCHC080054

978-1-4244-2340-8/08/$25.00 ©2008 IEEE

To date, only a handful of papers have been published directly related to predicting failure using match scores. The notion of biometric failure prediction as an analysis of algorithmic failure, as opposed to an image quality analysis, was first introduced in [3]. In that work, similarity scores are analyzed to predict system failure, or to verify system correctness after a recognizer has been applied. Adaboost is used to learn which feature spaces indicate failure. Most importantly, the expression of failure prediction as a failure prediction receiver operator characteristic (FPROC) curve is introduced, allowing failure prediction to be analyzed in a familiar manner. [4] successfully applies the failure prediction methodology of [3] to the problem of imagery selection from multiple cameras for face recognition.

The work of [5] takes the idea of failure prediction further by introducing *eye perturbations* as a means to enhance the gallery score distributions in a face identification scenario. An assumption is made, supported by [6], that inaccurate eye coordinates are primarily responsible for incorrect classification. By perturbing the eye coordinates of the gallery images, error in probe eye coordinates may be predicted, and compensated for, by perturbing the probe coordinates. Features are computed using Daubechies wavelets. The perturbation distributions are key to this work; the prediction module, via a neural network classifier, is able to learn the sensitivity of eye algorithms to eye coordinate error.

In [7], failure prediction is presented by first defining perfect recognition similarity scores (PRSS). These scores are obtained by submitting the gallery set as the probe set during matching. Based on these scores, a performance metric f can be computed as a function of system parameters, with a characteristic curve plotted for each value of f. This paper also uses perturbations to enhance system performance to the best value of f, but with worse performance compared to [5].

This paper extends the work of [3] and [5] by introducing support vector machines as a viable learning approach to predicting biometric system failure. Moreover, we articulate the problem of biometric failure prediction to one of similarity surface analysis, and extend this surface analysis to the perturbation space of [5]. We introduce a set of new features for learning and testing, and evaluate their performance over multiple algorithms and a standard data set (FERET).

The rest of this paper is as follows. In section 2, we introduce four different features that are used for the experimental analysis, and then describe how similarity surface analysis is the driving force behind our failure prediction system. The systemic details of failure prediction analysis are presented in Section 3, which introduces the Failure Prediction ROC curve. Section 4 briefly describes SVMs, and how they are applied to our learning problem, before moving on to the actual experimentation presented in section 5, where comprehensive experimental results for all features across 4 different algorithms are presented.

2. PREDICTING BIOMETRIC SYSTEM FAILURE

2.1. Features

We have defined a set of features partially in accordance with [3] and [5], and partially new. Each feature is derived from the distance measurements or similarity scores produced by the matching algorithm. Before each feature is calculated, the scores are first sorted from best to worst. In our system, for features 1, 2 & 4, we take the minimum of minimums over all views and perturbations for each gallery entry as the score for that particular gallery entry. The top k scores are considered for feature vector generation. For Feature 3, the perturbation scores are sorted per view (or across all views, taking the minimum).

1. $\Delta_{1,2}$ defined as (sorted score 1) - (sorted score 2). This is the separation between the top score and the second best score.

2. $\Delta_{i,j\ldots k}$ defined as ((sorted score i) $-$ (sorted score j), (sorted score i) $-$ (sorted sorted score $j+1$), $\ldots$, (sorted score i) $-$ (sorted score k)), where $j = i + 1$. Feature vectors may vary in length, as a function of the index i. For example, $\Delta_{1,2\ldots k}$ is of length $k-1$, $\Delta_{2,3\ldots k}$ is of length $k-2$, and $\Delta_{3,4\ldots k}$ is of length $k-3$.

3. $\delta_{i,j\ldots k}$ defined as (score for person i, perturbation j) - (score for person i, perturbation j), (score for person i, perturbation j) - (score for person i, perturbation $j+1$), $\ldots$, (score for person i, perturbation j) - (score for person i, perturbation k).

4. Take the top n scores and produce DCT coefficients. This is a variation on [5], where the Daubechies wavelet transform was shown to efficiently represent the information contained in a score series.

2.2. Similarity Surfaces

Let S be an n-dimensional similarity surface composed of k-dimensional feature data computed from similarity scores. The surface S can be parameterized by n different characteristics and the features may be from matching data, non-matching data or a mixed set of both.

Similarity Surface Theorem 2.1. *For a recognition system, there exists a similarity surface S, such that surface analysis around a hypothesized "match" can be used to predict failure of that hypothesis with high accuracy.*

While the (empirical) similarity surface theorem 2.1 suggests that shape analysis should predict failure, the details of the shapes and their potential for prediction are unknown functions of the data space. Because of the nature of biometric spaces, the similarity surface often contains features at multiple scales caused by matching with sub-clusters of related data (for example, multiple samples from the same individual over time, from family members, or from people in

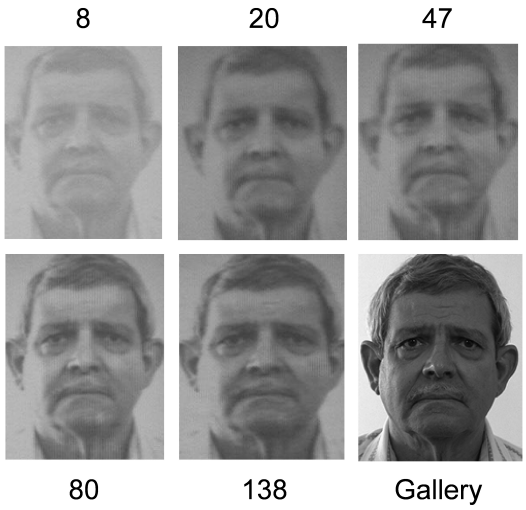


Fig. 1. Five images of varying quality, and associated rank scores, along with the original gallery image for comparison. Note that apparent quality is not always correlated with rank.

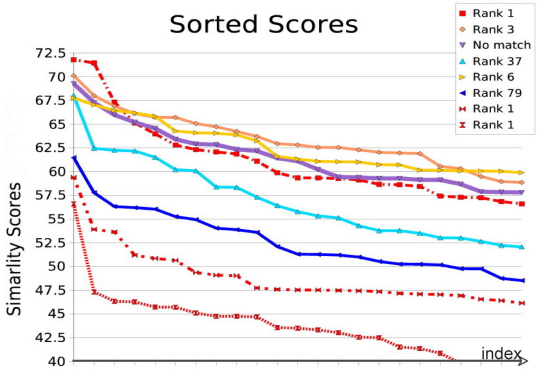
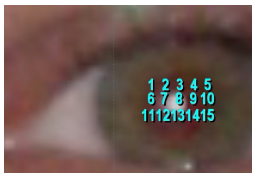


Fig. 2. Sorted similarity scores expressing performance for a single probe image at varying qualities and recognition rates. Notice, following from figure 1, that image quality is not always indicative of match performance.

similar demographic populations). What might be "peaked" in a low-noise system, where the inter-subject variations are small compared to intra-subject variations, might be flat in a system with significant inter-subject variations and a large population. These variations are functions of the underlying population, the biometric algorithms, and the collection system. Thus, with theorem 2.1 as a basis, the system "learns" the appropriate similarity shape information for a particular system installation.

As previous work [3] [5] has shown, similarity scores are not always tied to image quality rank, as is shown in figure 1. Figure 2 also shows this with a plot of sorted similarity scores arranged by image quality rank (the curves in this plot



(a)

(b)

Fig. 3. (a) Locations of perturbed center eye coordinate for 15 different positions. In our experiments, the distance between perturbations is 3 pixels. (b) Example perturbed normalized images. Note that perturbing the eye coordinates produces noticeable visual differences between the resulting normalized images.

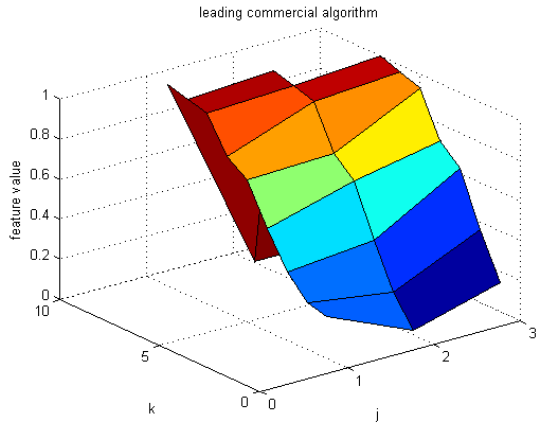
are one-dimensional surfaces for the simple feature of sorted scores). This observation leads us to explore other features over a sorted score space. This paper considers both features computed over sorted scores for an entire gallery (as was done in [3]), and sorted scores for single perturbation spaces (as was done in [5]). Eye perturbations exist as fixed-length offsets from the center eye coordinates produced by an eye detector or ground-truth. Figure 3 notes the locations of the eye perturbations used for the experiments presented in this paper. It is important to note that similarity scores from perturbations do not require a gallery, thus, it is possible to predict failure with just samples from the user at hand. Eye perturbations can also be used to automatically correct for inaccurate detected coordinates after failure has been predicted.

The nature of matching similarity surfaces for a feature class and their difference compared to other non-matching surfaces within same feature class may be explicit, or subtle. Figure 4 highlights this, with surfaces constructed from three feature vectors for a single feature space for one individual matching against an entry in the gallery. As noted above, machine learning discerns subtle differences between surfaces, and thus builds an accurate predictor of system failure.

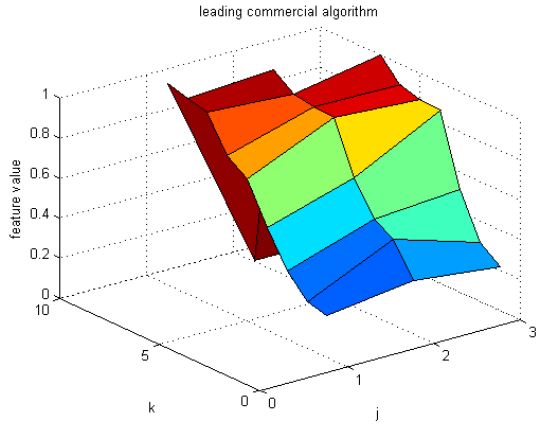
3. FPROC CURVES

As we are measuring system performance, this then suggests that for a comparison of measures what is needed is some form of a Receiver Operator Characteristic (ROC) curve on the prediction/classification performance. [3] suggests the following 4 cases that can be used as the basis of such a curve:

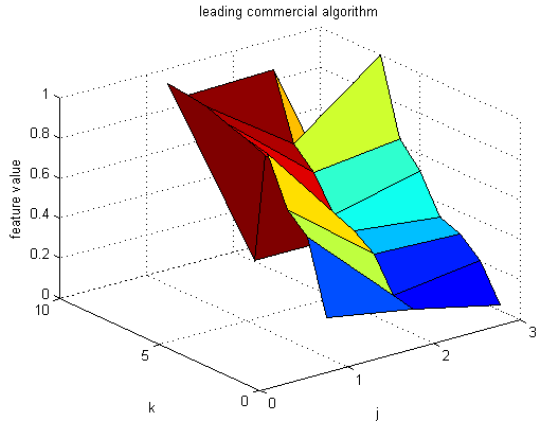
1. "False Accept", when the prediction is that the recognition system will succeed but the ground truth shows it will not.



(a)



(b)



(c)

Fig. 4. (a) a matching surface, (b) a similar non-matching surface, and (c) a dissimilar non-matching surface; all are for a leading commercial algorithm's perturbation space features (feature 3). Each plot is for one individual matching against an entry in the gallery. k represents the index of the feature in the vector, and j represents the index of the vector itself.

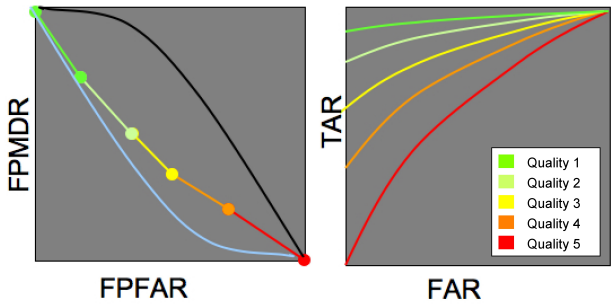


Fig. 5. An example FPROC curve appears on the left, while a traditional ROC curve expressing individual image qualities appears on the right. Segmenting the data on quality inflates the difference. Using full data sets in the FPROC allows us to vary the "quality" threshold.

Type I error of the failure prediction and Type I or Type II error of the recognition system.

2. "False Reject", when the prediction is that the recognition system will fail but the ground truth shows that it will be successful. Type II error of failure prediction.

3. "True Accept", wherein the underlying recognition system and the prediction indicates that the match will be successful.

4. "True Reject", when the prediction system predicts correctly that the system will fail. Type I or Type II error of the recognition system.

The two cases of most interest are Case 2 (system predicts they will not be recognized, but they are) and Case 1 (system predicts that they will be recognized but they are not). From these two cases we can define the Failure Prediction False Accept Rate (FPFAR), and Failure Prediction Miss Detection Rate (FPMDR) (= 1-FPFRR (Failure Prediction False Reject Rate)) as:

$$FPFAR = \frac{|Case2|}{|Case2| + |Case3|} \quad (1)$$

$$FPMDR = \frac{|Case1|}{|Case1| + |Case4|} \quad (2)$$

With these definitions, the performance of the different reliability measures, and their induced classifier, can then be represented in a Failure Prediction Receiver Operating Characteristic (FPROC) curve, of which an example is shown in figure 5. Implicitly, various thresholds are points along the curve and as the quality/performance threshold is varied, predictions of failure change the FPFAR and FPMDR just as changing the threshold in a biometric verification system varies the False Accept Rate and the Miss Detect Rate (or False Reject Rate). High quality data, which usually matches better,

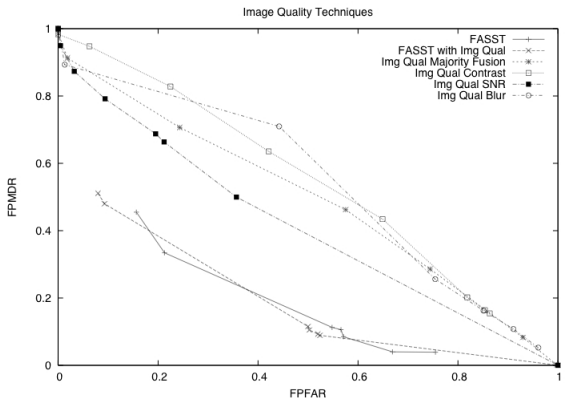


Fig. 6. FPROC for 4 different image quality techniques on 12,000 images, compared with our FASST^{TM} (Failure Analysis from Similarity Surface Theory) technique with and without image quality as a feature dimension.

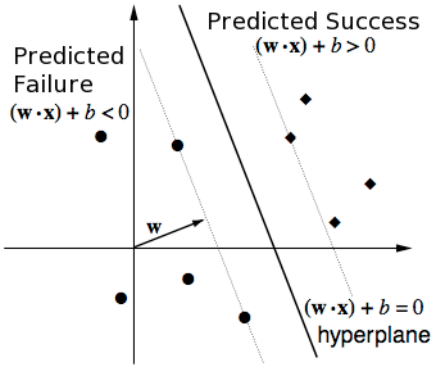


Fig. 7. Hyperplane with a maximal margin generated by a linear SVM. For failure prediction, matching similarity surfaces would be correctly classified on the positive side of the hyperplane, while non-matching similarity surfaces would be correctly classified on the negative side. (image adapted from: http://www.springerlink.com/content/k21rm08555372246/)

will generally be toward the upper right, with low failure prediction false alarms (and lower failures overall), but when good quality data does fail it is harder to predict it so more are missed. Lowest quality data is usually toward the bottom right, with few missed failure predictions, but more false predictions, as poor quality more often results in marginal but correct matches.

The advantage of using the FPROC curve as opposed to the traditional ROC evaluation of individual images (figure 5) is that it allows for a more direct comparison of different measures on the same populatoin, or a quality measure on different sensors/groups. It is impracticle to compare measures or sensors when each one generated 5 ROC curves. The ROC evaluation separated data into image quality tends to inflate the apparant distance by segmenting the gallery into individual curves, while FPROC evaluation allows us to vary the quality threshold over the gallery. The FPROC curve requires an "evaluation" gallery, and depends on the underlying recognition system's tuning, sensors, and decision making process.

The impact of switching approaches from a standard multiple ROC evaluation of image quality to the FPROC representation is noted in figure 6, where three different image quality techniques and a simple image-only fusion scheme are plotted over 12,000 images obtained in varied weather conditions outdoors. As can be seen, none of the techniques are truly suitable for predicting failure, when plotted on the FPROC curve (all four cut through the diagonal of the plot). Further, we make the comparison with failure prediction with similarity surfaces, the two lower curves, where two approaches are shown to be statistically better over the same data set, compared to the image quality techniques.

4. SUPPORT VECTOR MACHINES

Support Vector Machines [8] are a set of supervised learning methods for linear classification and regression of data. Figure 7 shows a simple example of SVM classification, whereby a set of positive examples (in our work, a matching similarity surface) and a set of negative examples (a non-matching similarity surface) are separated by a maximum interclass distance, known as the *margin*, in a hyperplane. The output formula for a linear SVM is:

$$u = w * x + b \tag{3}$$

where w is the normal vector to the hyperplane, and x is the input vector. The goal, as hinted at above, is to maximize the margin. Thus, an optimization problem is formualted:

$$\text{minimize } \frac{1}{2}||w||^2 \text{ subject to } y_i(w * x_i + b) \geq 1, \forall_i \tag{4}$$

where x_i is the i-th training example and $y_i \in \{-1, 1\}$ is, for the i-th training example, the correct output. The notion of "support vectors" comes into play with the training data x_i. For failure prediction, we define the set x as the feature vectors corresponding to successful and non-succesful match occurrences.

SVMs are based on the principle of structural risk minimization. This means SVMs can handle large amounts of input data, without incurring serious penalties for outliers (very common in noisy data). The implication of this is that we have the ability to process thousands of varying inputs for training in a reasonable amount of time, with good results.

5. EXPERIMENTAL RESULTS

In order to assess the performance of the SVM approach to failure prediction, incorporating the features of section 2.1,

Algorithm	Data Sets [1]	Training Samples	Test Samples
EBGM[2]	All	2000	1000
EBGM[3]	All	2000	1000
EBGM[4]	All	2000	1000
Robust PCA	All	2000	1000
Robust Revocable PCA	DUP1, DUP2, FAFC with perturbations	600	200
Commercial Algorithm	DUP1, DUP2, FAFC with perturbations	1000	400

Table 1. Algorithms and corresponding machine learning data information for all experiments.

we performed extensive testing with four different facial recognition algorithms. These algorithms include three variants of the EBGM algorithm [9] from the CSU Face Identification Evaluation Toolkit [10], the Robust PCA and Robust Revocable PCA algorithms introduced in [11], and one of the leading commercial face recognition algorithms. Each algorithm, and information about its learning data is presented in Table 1. In all cases, we assess the success of the prediction of rank 1 recognition. For all experiments, we used the entire set, or subsets, of the NIST FERET data set [12], with training and testing sets created by random sampling.

For Robust Revocable PCA and the commercial algorithm, 225 perturbations were generated per view for each gallery entry in order to assess feature 3. The perturbations for one eye are shown in figure 3. The distance between perturbations is 3 pixels. Considering the full size of the FERET set (3368 images), multiplied by 225, we chose instead to use a subset of FERET consisting of the DUP1, DUP2, and FAFC sets to speed up the rate of experimentation.

The FPROC curves of figures 8 - 15 were generated by considering the output of the SVM learning. By choosing a threshold t, and varying it over a series of increasing marginal distances (starting from the lowest noted distance, and moving to the highest), the margin of the hyperpane is adjusted. With each adjusted margin in the series, cases 1 - 4 can be calculated by determining on which side of the hyperplane each feature vector falls. Formulas 1 & 2 are then used to to calculate the FPFAR and FPMDR values for each margin separation. Only the best performing features are expressed in figures 8 - 15. If the SVM was unable to find good separation between the positive and negative training data, the result often leaves all test data classified on either the positive or negative side of the hyperplane. These cases are not plotted.

[1] Full or subsets of FERET
[2] EBGM Optimal FGMagnitude
[3] EBGM Optimal FGNarrowingLocalSearch
[4] EBGM Optimal FGPredictiveStep

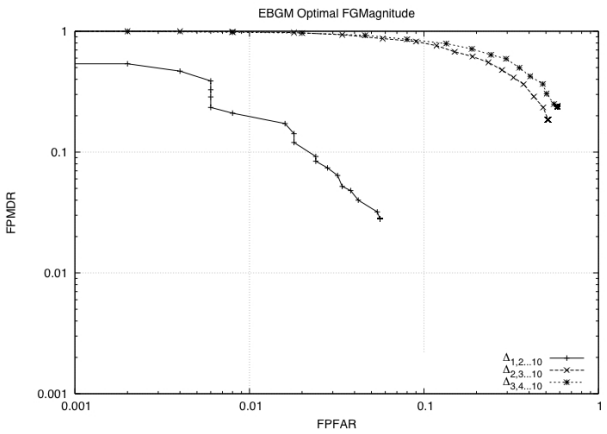


Fig. 8. Varations on feature 2 for the EBGM Optimal FGMagnitude algorithm. Algorithm rank 1 recognition rate is 0.841.

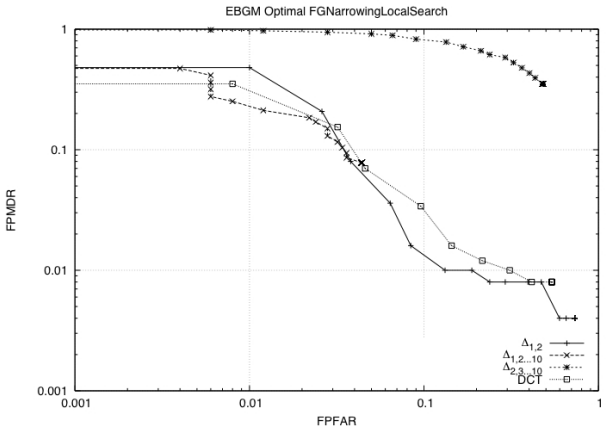


Fig. 9. Feature 1, two instances of feature 2, and feature 4 for the EBGM Optimal FGNarrowingLocalSearch algorithm. Algorithm rank 1 recognition rate is 0.853.

Depending on the security or usability requirements of our application, we can choose a point on the curve that will yield the acceptable failure prediction results. Curves where both FPFAR and FPDMR can be minimized to very low levels are desirable. Overall, feature 2 taken as $\Delta_{1,2...10}$ performs the best across all 4 algorithms, for scores spread across an entire gallery. Feature 4 also performs well in the cases it yielded valid classification results, especially for EBGM Optimal NarrowingLocalSearch and EBGM Optimal Predictive Step. Feature 1 produces valid classification results in only two experiments (EBGM Optimal NarrowingLocalSearch and EBGM Optimal Predictive Step). The lack of performance implies the difference between the first and second scores does not yield enough information to reliably build meaningful surfaces for failure prediction when taken by itself. Fea-

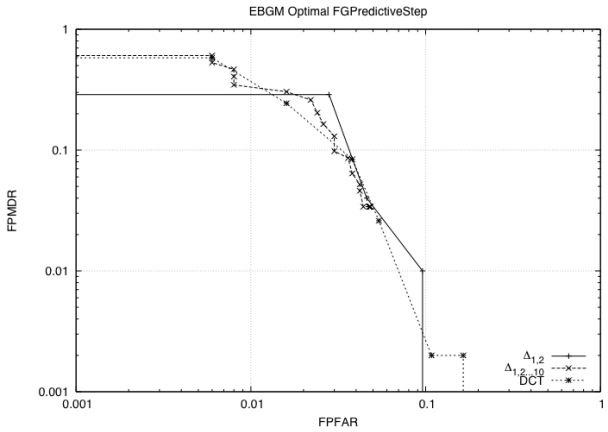


Fig. 10. Feature 1, one instance of feature 2, and feature 4 for the EBGM Optimal FGPredictiveStep algorithm. Algorithm rank 1 recognition rate is 0.817.

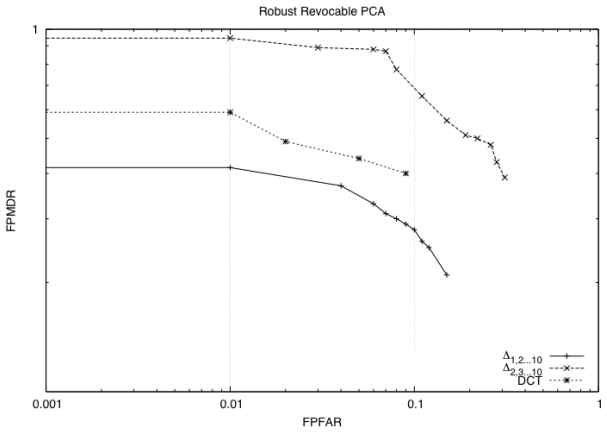


Fig. 12. Two instances of feature 2, and feature 4 for Robust Revocable PCA. Algorithm rank 1 recognition rate is 0.874.

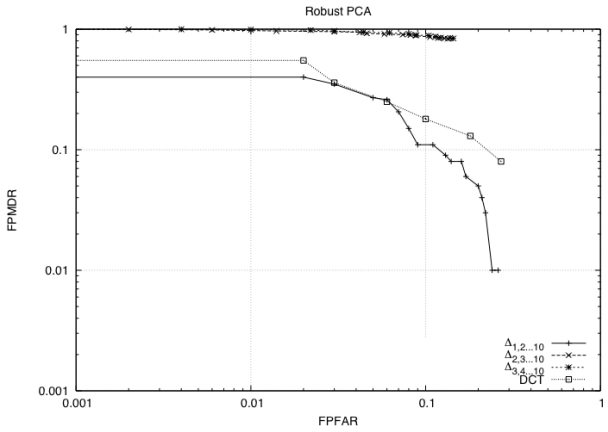


Fig. 11. Variations on feature 2, and feature 4 for the Robust PCA algorithm. Algorithm rank 1 recognition rate is 0.972.

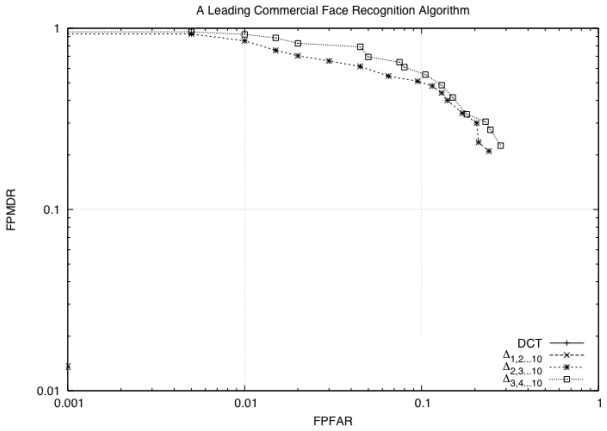


Fig. 13. Variations on feature 2, and feature 4 for a leading commercial algorithm. Algorithm rank 1 recognition rate is 0.6. Note DCT and $\Delta_{1,2...10}$ maintain a FPFAR rate of 0, suggesting more data is needed for future testing.

ture 2 taken as $\Delta_{2,3...10}$ and $\Delta_{3,4...10}$ also performs poorly, which reinforces the notion of $\Delta_{1,2...10}$ as strong performer, taken over the most relevant score as a feature vector of sufficient length. The noted variance in feature performance suggests feature level fusion is a valid approach to further refining failure prediction.

Of even more interest are the results for scores spread across the perturbation space in figures 14 and 15, as this approach does not require a "gallery", the similarity surface is from just data about the match. Both Robust Revocable PCA and the commercial algorithm achieve a FPMFR of around 0.1 around a FPFAR of 0.05. If the tolerance for false failure prediction in an application is higher, the commercial algorithm can reach a FPMFR of nearly 0 by FPFAR of 0.15.

The results presented in this paper are comparable, an in many cases better, to the results reported in [3] [5] [7]. [3], using an Adaboost predictor and minimizing FPMDR, reports "best" predictions of between 0.02 FPMDR and 0.1 FPFAR, and 0.01 FPMDR and 0.5 FPFAR for its own features and data sets. [5], using a neural network predictor, reports a correct classification rate "exceeding 90%" for the entire gallery perturbation space using both EBGM and a commercial algorithm on the FERET data set. [7] reports an overall error rate of between 15% and 25% on FERET FA, FB, and DUP1 for its "perfect recognition" curve scheme.

6. CONCLUSION

In this paper, we have extended the work of [3] and [5] in several ways, further reinforcing the viability and importance of post-recognition biometric failure prediction as a superior alternative to image quality based prediction.

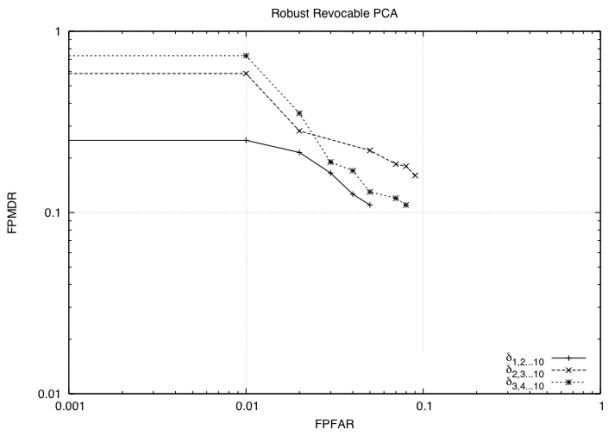


Fig. 14. Variations on feature 3 for Robust Revocable PCA.

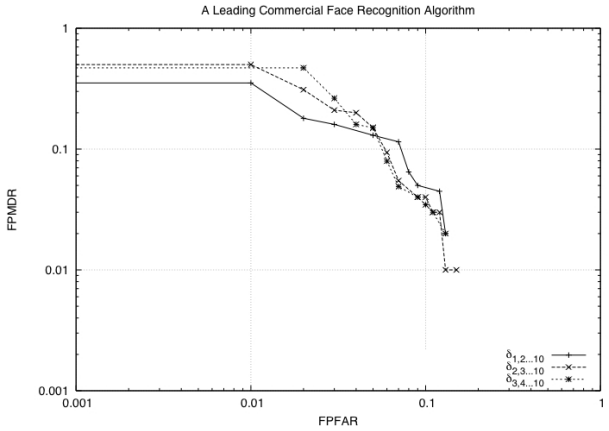


Fig. 15. Variations on feature 3 for a leading commercial recognition algorithm.

Combining the extension herein extension with past papers, we have shown 5 different classes of features sets (hand-defined, wavelets, various delta from sorted similarity scores, delta from perturbation scores, and DCTs), combined with 3 different learning approaches (AdaBoost, NeuralNets and SVM) being used to predict failure over 6 different recognition algorithms (including 2 versions of a leading commercial algorithm) – all with significant success. Based on this we evidence we postulated our empirical similarity surface theory.

Through similarity surface analysis, we have shown an important advantage over image quality, when we plot using an FPROC curve, and explored the potential of the perturbation feature space brought into the FPROC analysis domain. We introduced a new set of four features to be considered for failure prediction, and used them as inputs to an SVM framework - a new learning approach for this sort of failure prediction. The results of our experiments using four face recognition algorithms are extremely promising, and we are are currently investigating multi-feature fusion to enhance failure prediction as an extension of this current work.

7. REFERENCES

[1] E. Tabassi, C.L. Wilson, and C.I. Watson, "Fingerprint Image Quality, NFIQ," in *National Institute of Standards and Technology, NISTIR 7151*, 2004.

[2] P. Grother and E. Tabassi, "Performance of Biometric Quality Evaluations," *IEEE TPAMI*, vol. 29, no. 4, pp. 531–543, 2007.

[3] W. Li, X. Gao, and T.E. Boult, "Predicting Biometric System Failure," in *Proc. of the IEEE Conference on Computational Intelligence for Homeland Security and Personal Safety (CIHSPS 2005)*, 2005.

[4] B. Xie, V. Ramesh, Y. Zhu, and T. Boult, "On Channel Reliability Measure Training for Multi-Camera Face Recognition," in *In Proc. of the IEEE Workshop on the Application of Computer Vision (WACV)*, 2007.

[5] T.P. Riopka and T.E. Boult, "Classification Enhancement via Biometric Pattern Perturbation," in *IAPR Conference on Audio- and Video-based Biometric Person Authentication (Springer Lecture Notes in Computer Science)*, 2005, vol. 3546, pp. 850–859.

[6] T.P. Riopka and T.E. Boult, "The Eyes Have It," in *Proc. of the ACM SIGMM Workshop on Biometric Methods and Applications*, 2003.

[7] P. Wang and Q. Ji, "Performance Modeling and Prediction of Face Recognition Systems," in *In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2006)*, 2006.

[8] C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[9] K. Okada, J. Steffens, T. Maurer, H. Hong, E. Elagin, H. Neven, and C. von der Malsburg, "The Bochum/USC Face Recognition System And How it Fared in the FERET Phase III test," in *Face Recognition: From Theory to Applications*, H. Wechsler, P. J. Phillips, V. Bruce, F. Fogeman Soulié, and T. S. Huang, Eds., pp. 186–205. Springer-Verlag, 1998.

[10] R.J. Beveridge, D. Bolme, M. Teixeira, and B. Draper, "The CSU Face Identification Evaluation System Users Guide: Version 5.0," *Technical report, Colorado State University*, 2003.

[11] T.E. Boult, "Robust Distance Measures for Face Recognition Supporting Revocable Biometric Tokens," in *In Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition, Southampton, UK*, 2006.

[12] P.J. Phillips, H. Moon, S.A. Rizvi, and P.J. Rauss, "The FERET Evaluation Methodology for Face-Recognition Algorithms," *IEEE TPAMI*, vol. 22, no. 10, pp. 1090–1104, 2000.