Pre-print of article that appeared at the IEEE Computer Society Workshop on Biometrics 2008.

The published article can be accessed from:
http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=4563124

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 post-recognition analysis. To generate the surfaces used for comparison, we define a set of new features derived from distance measures or similarity scores. 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

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. This approach does not tell us very much about the conditions for which an algorithm fails, nor does it allow us to automatically compensate for failure conditions. Post-recognition analysis techniques address algorithmic failure prediction, and open the door to more advanced score and feature level fusion techniques.

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, presented 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 Cumulative Match Curve (CMC) format, and showing that the CMC for higher quality data was above that for lower quality data. This does demonstrate the quality/reliability measure has some prediction value.

To date, only a handful of papers have been published directly related to predicting failure in a post-match sense. The notion of biometric failure prediction as an analysis of algorithmic failure, as opposed to an image quality analysis, was first introduced in [2]. 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. [3] [4] successfully apply the failure prediction methodology of [2] to the problem of imagery selection from multiple cameras for face recognition.

The work of [6] 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 [7], 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 [5], 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 char-

acteristic 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 [6].

This paper extends the work of [2] and [6] 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 [6]. 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 describe how similarity surface analysis is the driving force behind our failure prediction system, along with the systemic details of failure prediction analysis. In section 3, we describe four different features that are used for the experimental analysis. 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.

Fig. 1. Three 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.

2. PREDICTING BIOMETRIC SYSTEM FAILURE

2.1. Similarity Surfaces

While the general theory suggests that shape analysis should predict failure, the details of the shapes and their potential for prediction are also a function 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 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, in a post-recognition system, the system "learns" the appropriate similarity shape information for a particular system installation.

The similarity surfaces are composed of n-dimensional data, determined by the number of data points for each feature. As previous work [2] [6] 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. 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 [2]), and sorted scores for single perturbation spaces (as was done in [6]). Eye perturbations exist as fixed-length offsets from the center eye coordinates produced by an eye detector or ground-truth. Figure 3 notes

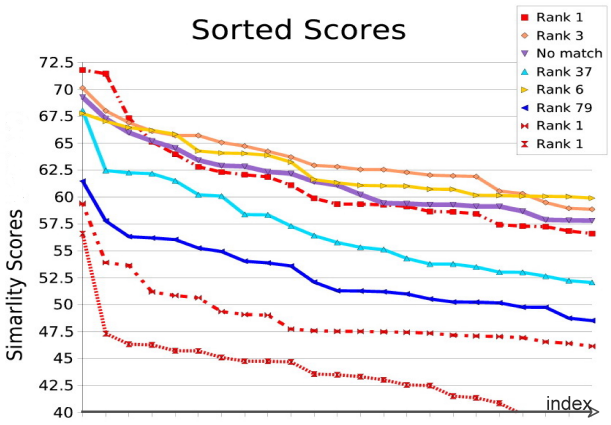

Fig. 2. Sorted similarity scores expressing performance: $\{s(x_i, y_1), s(x_i, y_2), ..., s(x_i, y_n)\}$. Notice that image quality is not always indicative of match performance.

the locations of the eye perturbations used for the experiments presented in this paper.

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. Figures 4 and 5 highlight this, with surfaces constructed from three feature vectors for a single feature space for one individual matching against an entry in the gallery. In these figures, two algorithms, Robust Revocable PCA [11] and a leading commercial algorithm, are shown with a matching surface on the top, a similar non-matching surface in the middle, and a dissimilar non-matching surface on the bottom. As noted above, machine learning is able to discern subtle differ-

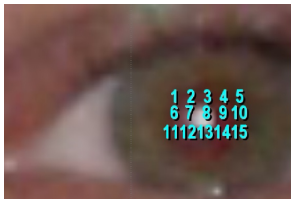

Fig. 3. Locations of perturbed center eye coordinate for 15 different positions. In our experiments, the distance between perturbations is 3 pixels.

ences between surfaces, and thus builds an accurate predictor of system failure.

2.2. 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. [2] suggests the following 4 cases that can be used as the basis of such a curve:

1. "True Accept", wherein the underlying recognition system and the prediction indicates that the match will be successful.
2. "False Accept", when the prediction is that the recognition system will succeed but the ground truth shows it will not.
3. "False Reject", when the prediction is that the recognition system will fail but the ground truth shows that it will be successful.
4. "True Reject", when the prediction system predicts correctly that the system will fail.

The two cases of most interest are Case 3 (quality predicts they will not be recognized, but they are) and Case 2 (quality 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:

$$FPFAR = \frac{|Case3|}{|Case3| + |Case1|} \quad (1)$$

$$FPMDR = \frac{|Case2|}{|Case2| + |Case4|} \quad (2)$$

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

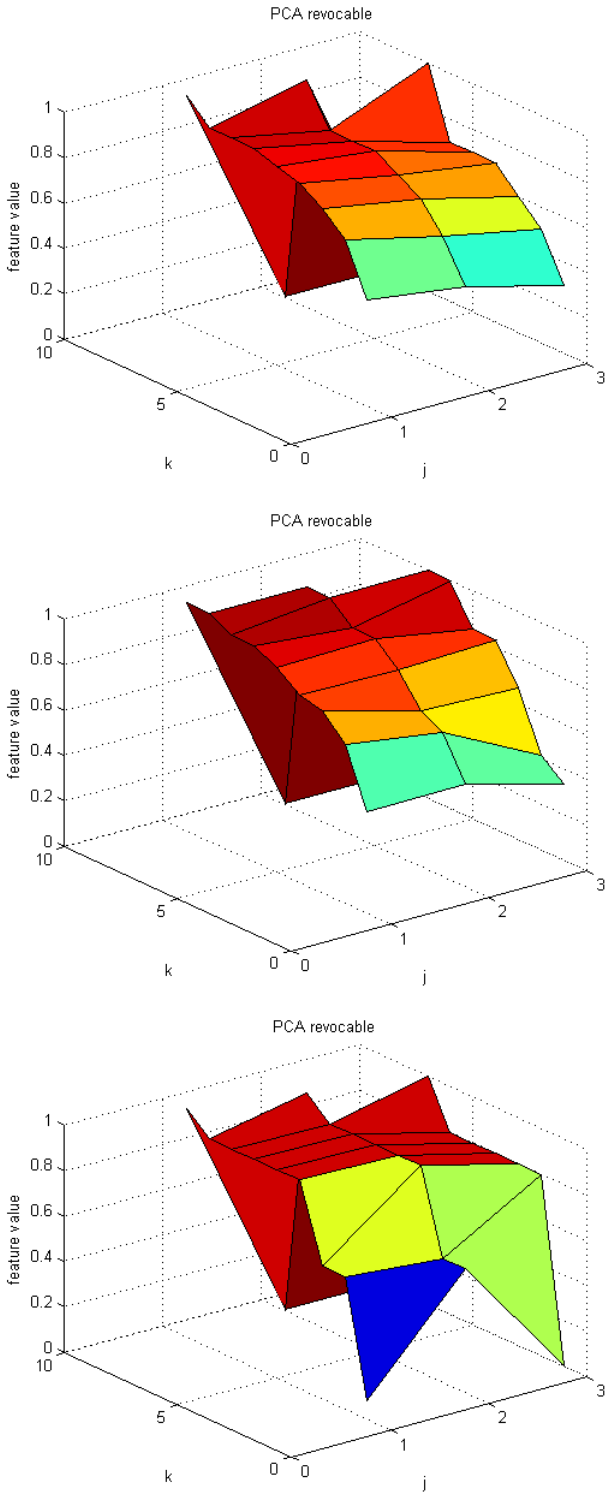

Fig. 4. From top to bottom, a matching surface, a similar non-matching surface, and a dissimilar non-matching surface for Robust Revocable PCA's perturbation space features (feature 3).

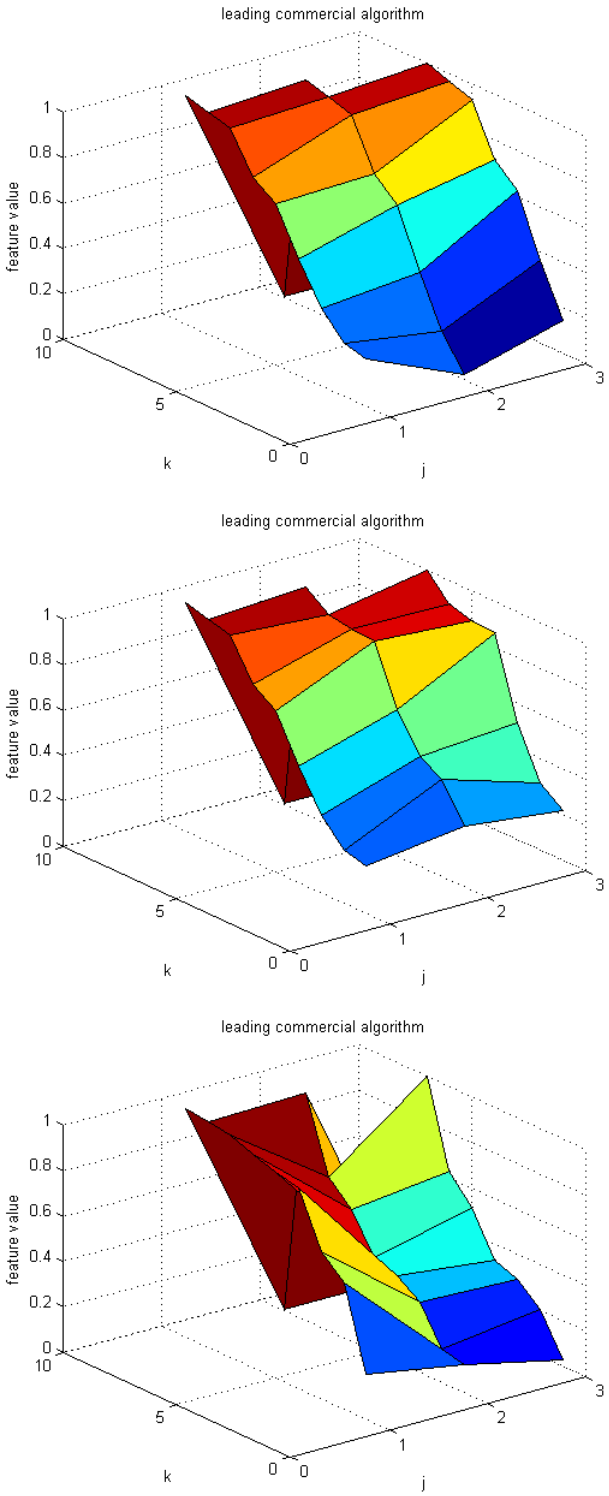

Fig. 5. From top to bottom, a matching surface, a similar non-matching surface, and a dissimilar non-matching surface for a leading commercial algorithm's perturbation space features (feature 3).

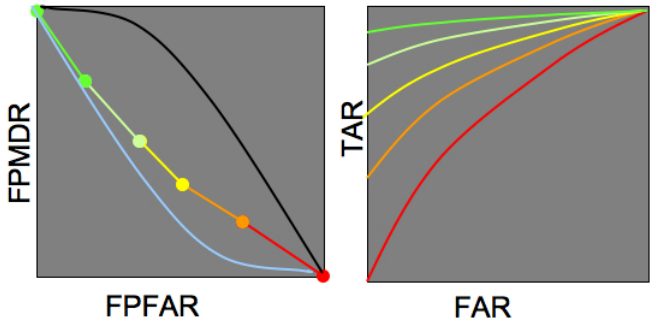

Fig. 6. An example FPROC curve appears on the left, while a traditional ROC curve expressing individual image qualities appears on the right. As can be seen from the ROC curve, segmenting the gallery on quality inflates the difference. Considering full data sets in the FPROC allows us to vary the "quality" threshold.

figure 6. 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).

The advantage of using the FPROC curve as opposed to the traditional ROC evaluation of individual images (figure 6) is that it allows for a more direct comparison of different gallery measures, or a quality measure on different sensors/groups. The ROC evaluation of image quality tends to inflate the 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 and decision making process. We note that it may understate the impact of removing poor quality images from the process.

The impact of switching approaches from a standard ROC evaluation of image quality to the FPROC representation is noted in figure 7, 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 to similarity surfaces, where two approaches are shown to be statistically better over the same data set, compared to the image quality techniques.

3. FEATURES

We have defined a set of features partially in accordance with [2] and [6], and partially new. Each feature is derived from the distance measurements or similarity scores produced by

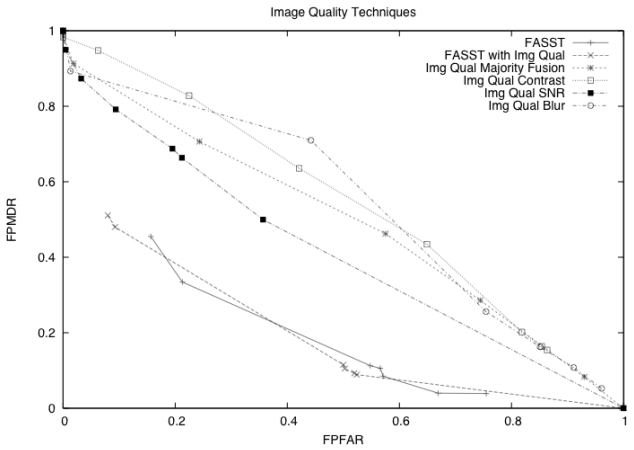

Fig. 7. FPROC for 4 different image quality techniques on 12,000 images.

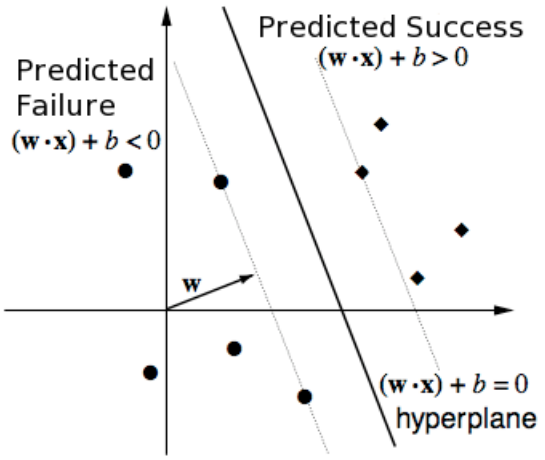

Fig. 8. Hyperplane with a maximal margin generated by a linear SVM. For failure prediction, matching similarity surfaces would be correctly classified on the positive side of the hyperplane, while non-matching similarity surfaces would be correctly classified on the negative side. (image adapted from: http://www.springerlink.com/content/k21rm08555372246/)

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 scores.

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 [6], where the Daubechies wavelet transform was shown to efficiently represent the information contained in a score series.

4. SUPPORT VECTOR MACHINES

Support Vector Machines [8] are a set of supervised learning methods for linear classification and regression of data. Figure 8 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 3, we

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.

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. 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 figure3. 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 9 - 16 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 9 - 16. 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.

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. Feature 2 taken as $\Delta_{2,3...10}$ and $\Delta_{3,4...10}$ also performs poorly,

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

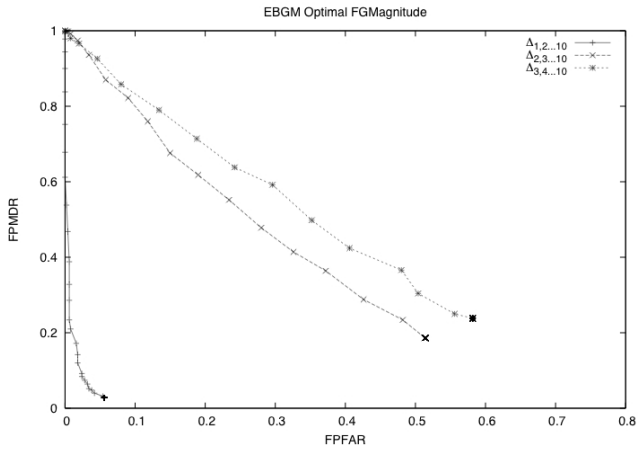

Fig. 9. Three instances of feature 2 for the EBGM Optimal FGMagnitude algorithm. Algorithm rank 1 recognition rate is 0.841.

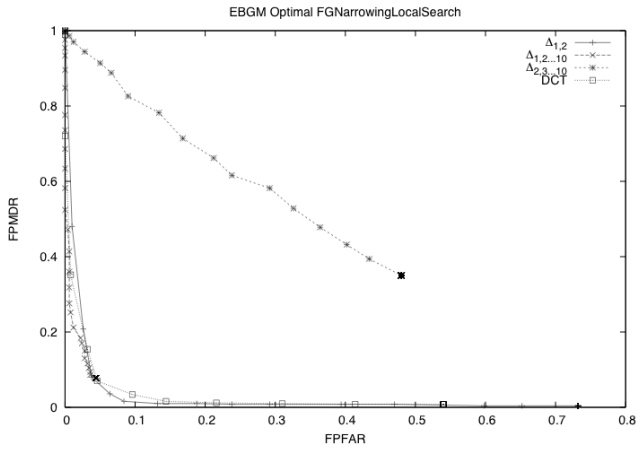

Fig. 10. Feature 1, two instances of feature 2, and feature 4 for the EBGM Optimal FGNarrowingLocalSearch algorithm. Algorithm rank 1 recognition rate is 0.853.

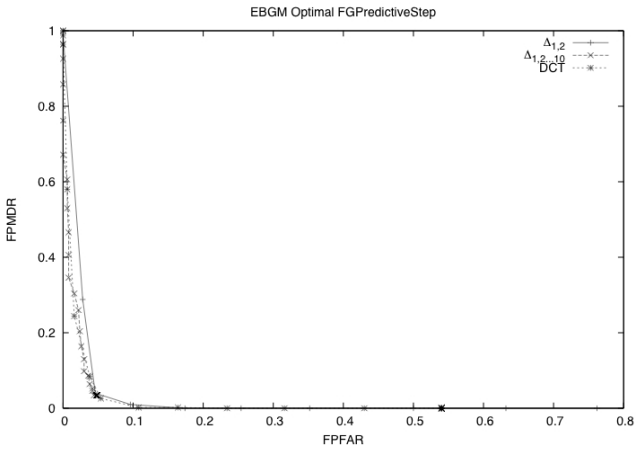

Fig. 11. Feature 1, one instance of feature 2, and feature 4 for the EBGM Optimal FGPredictiveStep algorithm. Algorithm rank 1 recognition rate is 0.817.

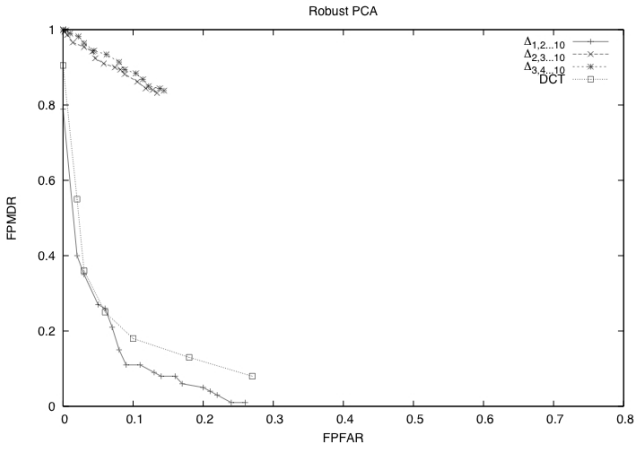

Fig. 12. Three instances of feature 2, and feature 4 for the Robust PCA algorithm. Algorithm rank 1 recognition rate is 0.972.

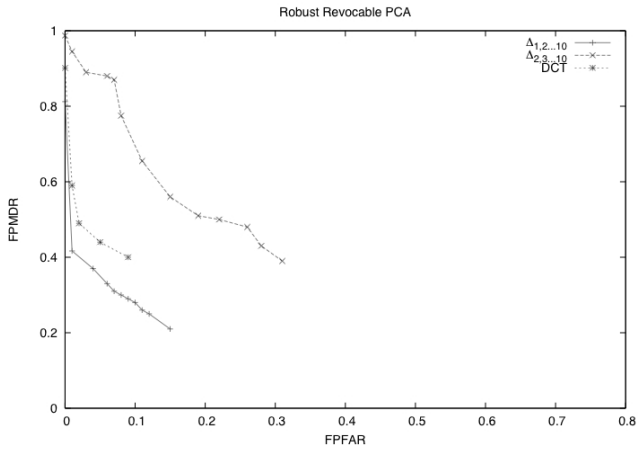

Fig. 13. Two instances of feature 2, and feature 4 for Robust Revocable PCA. Algorithm rank 1 recognition rate is 0.874.

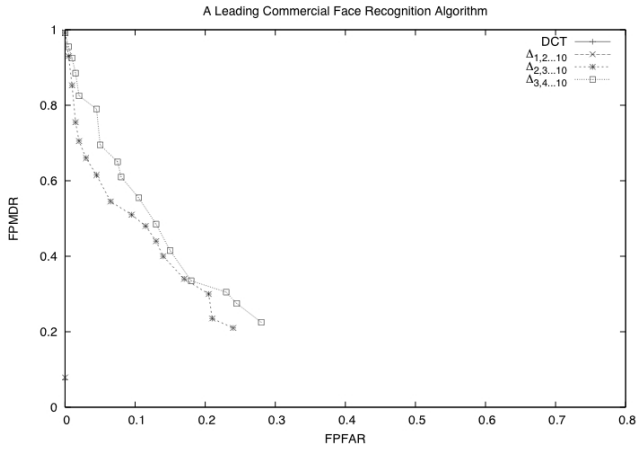

Fig. 14. Three instances of feature 2, and feature 4 for a leading commercial algorithm. Algorithm rank 1 recognition rate is 0.6. Note DCT and $\Delta_{1,2...10}$ maintain a FPFAR rate of 0, suggesting more data is needed for future testing.

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 15 and 16. 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, if not better, to the results reported in [2] [6] [5]. [2], 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 F PFAR for its own features and data sets. [6], 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. [5] reports an overall error rate of between 15% and 25% on FERET FB, FB, and DUP1 for its "perfect recognition" curve scheme.

6. CONCLUSION

In this paper, we have extended the work of [2] and [6] in several ways, further reinforcing the viability and importance of post-recognition biometric failure prediction as a superior alternative to image quality based prediction. Through sur-

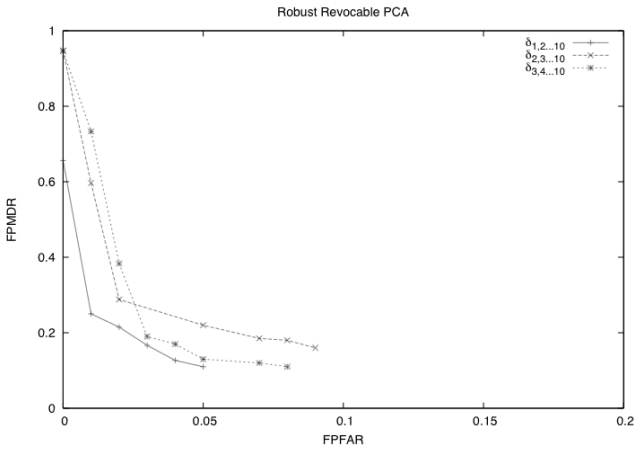

Fig. 15. Three instances of feature 3 for Robust Revocable PCA.

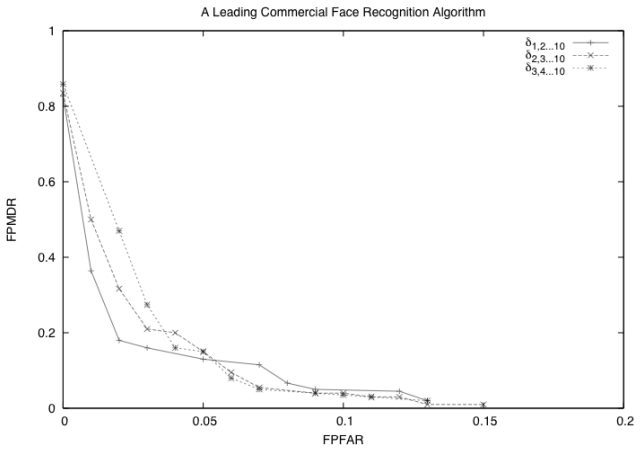

Fig. 16. Three instances of feature 3 for a leading commercial recognition algorithm.

face 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] 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.

[3] B. Xie, T. Boult, V. Ramesh, and Y. Zhu, "Multi-camera face recognition by reliability-based selection," in *In Proc. of the IEEE Conference on Computational Intelligence for Homeland Security and Personal Safety*, 2006.

[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] 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.

[6] 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.

[7] T.P. Riopka and T.E. Boult, "The eyes have it," in *Proc. of the ACM SIGMM Workshop on Biometric Methods and Applications*, 2003.

[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] Kazunori Okada, Johannes Steffens, Thomas Maurer, Hai Hong, Egor Elagin, Hartmut Neven, and Christoph 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. of the 7th IEEE International Conference on 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.