Pre-print of article that appeared at the IEEE BTAS 2008.

The published article can be accessed from:
http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=4699339

A FUSION-BASED APPROACH TO ENHANCING MULTI-MODAL BIOMETRIC RECOGNITION SYSTEM FAILURE PREDICTION AND OVERALL PERFORMANCE

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

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

ABSTRACT

Competing notions of biometric recognition system failure prediction have emerged recently, which can roughly be categorized as quality and non-quality based approaches. Quality, while well correlated overall with recognition performance, is a weaker indication of how the system will perform in a particular instance - something of primary importance for critical installations, screening areas, and surveillance posts. An alternative approach, incorporating a Failure Prediction Receiver Operator Characteristic (FPROC) analysis has been proposed to overcome the limitations of the quality approach, yielding accurate predictions on a per instance basis.

In this paper, we develop a full multi-modal recognition system integrating an FPROC fusion-based failure prediction engine. Four different fusion techniques to enhance failure prediction are developed and evaluated for this system. We present results for the NIST BSSR1 multi-modal data set, and a larger "chimera" set also composed of data from BSSR1. Our results show a significant improvement in recognition performance with the fusion approach, over the baseline recognition results and previous fusion approaches.

1. INTRODUCTION

For any biometric recognition system, maximizing the performance of recognition is a primary goal. Clearly, we do not want an impostor to be recognized as a legitimate user, nor do we want a misidentification in the case of a watch-list security/surveillance application. Moreover, when a legitimate user attempts to interface with a recognition system for authentication or verification, we expect that they will be identified properly. Any case where an undesirable result occurs in these scenarios is an instance of *failure*.

Image or sample quality has long stood out as the leading predictor of failure in biometric recognition systems. NIST continues to be the most visible organization promoting quality, producing several influential studies [1] [2]. In [1], a reliability measure for fingerprint images is introduced, and is shown to have a strong correlation with recognition performance. In [2], methods for the quantitative evaluation of systems that produce quality scores for biometric data are described. Both works make a strong case for quality as an overall predictor of system of success, and thus, promote the widespread use of quality as a predictor of failure. However, current work emerging from NIST on quality assessment is starting to question the assumption of quality as a universally good predictor.

At the recent Multiple Biometric Grand Challenge workshop, two presentations [3] [4] for NIST commissioned studies on quality assessment made the following claim:

> Quality is not in the eye of the beholder; it is in the *recognition performance figures*!

This assessment was based on the analysis of quality metrics for iris and face recognition. For the iris work [3], three different quality assessment algorithms lacked correlation in resulting recognition performance, indicating a lack of consensus on what image quality actually is. In the face recognition work [4], out of focus imagery was shown to produce better match scores. Both studies produced conclusions that are counterintuitive to traditional notions of quality assessment. Further, the work of [5] also introduces this notion, with a variety of "poor quality" images shown to produce better matching scores than "high quality" images for the same subject. Instead of suggesting a new paradigm for failure prediction, NIST has remained firm in its backing of quality assessment, posing this issue as an open question for researchers.

Reflecting upon this issue of quality a bit deeper, we can begin to understand its limitations. On a per instance basis, [4] showed that what is visually of poor quality produced good recognition results. Thus, "quality" is indeed found in the recognition performance. A compelling alternative approach [6] is to learn when a system fails and when it succeeds, and classify individual recognition instances using the learning as a basis. Based on the decisions made by a machine learning classification system, a Failure Prediction Receiver Operator Characteristic Curve can be plotted, allowing the system operator to vary a quality threshold in a meaningful way. Failure prediction analysis of this sort has been shown to be quite effective for single modalities [6], fusion across sensors for a single modality [7], and across different machine learning techniques [8] [5]. A further goal is to enhance the failure pre-

This work was supported in part by DHS SBIR NBCHC080054, and NSF PFI Award Number 0650251

diction performance through fusion across algorithms, failure prediction features, and modalities, and ultimately, enhance the performance of the underlying recognition system.

A wide variety of methods for fusion have been developed and many have been studied for biometrics. In [9], several score-level fusion techniques are compared with measurements of the effectiveness of fusing data made in various combinations: multi-modal, multi-instance, multi-sample, and multi-matcher. This extensive study, however, was focused on existing ad hoc techniques. It is possible that one or more sensors may fail in recognition, and the goal of fusion is to only use the non-failing result. The ideal approach is one that bases the fusion on the quality of the data and the quality of the biometric matching on the quality of the resulting fusion (see the work of [10], [11], [12], [13], [14], [15] and [16]). The fundamental idea in quality-based fusion is to more heavily weight the results from "higher quality" data. Unfortunately, in today's multimodal biometric systems, there is often no real-time indication as to the quality of data or match results generated from each biometric algorithm/modality in the system. Without such data a "quality" based fusion cannot be applied. Moreover, as we have already discussed, a "poor quality" image may indeed result in a good match score.

The vast majority of fusion work to date has focused on combining consistent data to address limitations of sensors or failure of a modality to correctly identify a subject. We also note that for screening, especially for adversarial threats, failure-prediction based fusion is quite different from any fusion approach that is focused on combining consistent data. If an adversary is actively trying to defeat the screening, then a biometric, say voice, that is providing a more inconsistent answer than another, say multi-view face, does not mean it should be ignored. If one modality predicts success and another predicts failure, we can safely ignore the predicted failures and let the others proceed, consistent or not. The recent case of Ramirez Abadia, in Brazil, who underwent multiple face reconstruction surgeries, but was apprehended by voiceprint recognition, with the help of the DEA, is a timely example[1]. A face-based system would have failed to recognize him, but a voice-based system correctly recognized him. As we shall see, the FPROC-based technique is not just a binary classification, but more of an overall confidence measure (that may be thresholded for classification), so it can be very effectively used to support various approaches to fusion and hybrid classifiers. But unlike the ad-hoc hybrid classifiers, FPROC analysis can provide a more formal way to determine goodness, in the sense of [13].

In this paper, we develop a full multi-modal recognition system integrating a failure prediction fusion-based engine. In section 2, we introduce the fusion-based failure prediction system architecture, with failure prediction features described in 2.1 and specific fusion techniques described in 2.2.

[1] http://www.washingtonpost.com/wp-dyn/content/article/2007/08/10/AR2007081000704.html

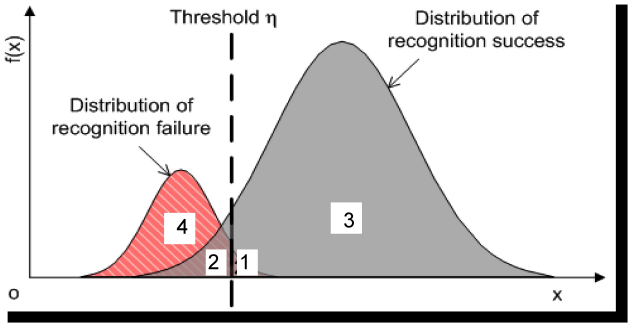


Fig. 1. Thresholding a per datum reliability or "quality" measure to predict recognition system success produces 4 different "cases", depending on the success of the recognition system on a sample and the reliability-based prediction of success or failure associated with that sample.

In section 3, we report extremely promising results for fusion of failure prediction, and show significant improvement in recognition after failure prediction analysis. We conclude on the note that FPROC based analysis is an important alternative to quality analysis, especially in cases where per instance failure prediction is essential.

2. ENHANCING FUSION WITH FAILURE PREDICTION

If a system can predict which input is more likely to fail, that input can be given less weight. Figure 1 shows two distributions of successful recognitions and recognition system failure with the x-axis showing some measure of confidence. The idea of post-recognition failure prediction is based on the construction of a learning system that will predict failure based on prior system performance. The input to such a learning system is a feature vector calculated from the recognition scores. It might seem we could focus on "success-based fusion", since predicting failure is the opposite of predicting success. However, we focus on failure prediction since the failures are, in general, the less frequent outcome, and hence better suited to machine learning approaches. Over a set of recognition scores, each output of such a learning system is in one of the following four cases (depicted in figure 1):

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

4. "True Reject", when the prediction system predicts correctly that the system will fail. Prediction success with 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:

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

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

A variety of features may be used as a basis for an FPROC curve, as long as they have the ability to capture the information contained in the *tails* of the underlying score *distributions* of figure 1. It is these tail regions that represent the scores most likely to cause the system to fail, and the regions our classifier can be tuned to, in order to reduce the Type I or Type II errors of the system. A good feature for failure prediction will be able represent much more than just a raw score, which is often too ambiguous for learning whether or not it is a match or non-match, because it lacks associated distributional information. Failure prediction features are designed to capture distributional information over a series of localized scores. In previous work, Daubechies wavelets [8], DCTs, and difference calculations [5] were applied to localized scores. Further, raw quality itself can be re-introduced in the context of failure prediction as a feature [6], enhancing its somewhat weak prediction capabilities for per instance prediction by building predictors over quality classifications. The previous work has shown that were are not bound to any particular feature, or learning technique, for acceptable performance.

2.1. Failure Prediction Features

Each feature used for experiments in this work is derived from the distance measurements or similarity scores produced by the matching algorithm and is designed to capture information about the nature of the score set. These features were shown to be effective in [5]. Before each feature is calculated, the scores are first sorted from best to worst. In our system, for all features, we take the minimum of minimums over all views for each gallery entry as the score for that particular gallery entry. The top k scores are considered for feature vector generation.

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 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. Take the top n scores and produce DCT coefficients. This is a variation on [8], where the Daubechies wavelet transform was shown to efficiently represent the information contained in a score series.

2.2. Fusion Techniques

Decision level fusion is defined as data processing by independent algorithms, followed by the fusion of decisions (based upon the calculated results) of each algorithm. This idea can be thought of as n different inputs to n different algorithms, producing n decisions that are fused together to produce a final decision that the system will act upon. The power of decision fusion for our system stems from our need to fuse data over independent modalities and corresponding recognition algorithms, as well as independent features over failure prediction. Ultimately, we would like our system to give us a final decision on whether or not the subject was correctly recognized.

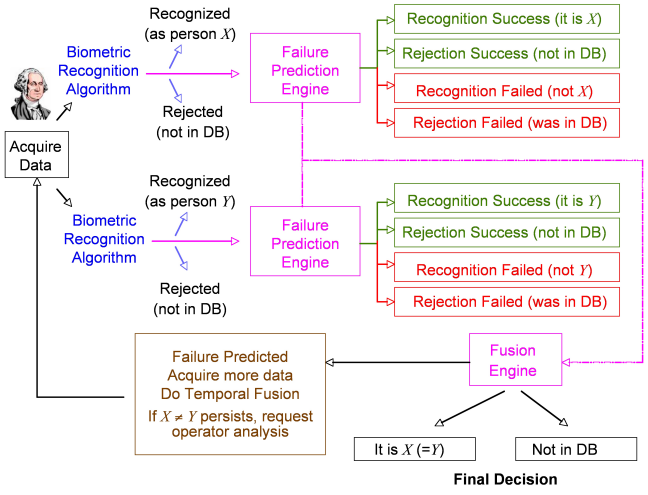


Fig. 2. A multi-modal recognition system incorporating failure prediction based fusion. The failure prediction analysis of our system predicts both individual algorithm/modality failures, drives fusion weighting and predicts overall success or failure of the fusion process.

The multistage nature of our system allows for fusion to take place at various levels throughout the system. Referring to the system diagram of figure 2, we see at the highest level the need to fuse the results of failure prediction across modalities. Thus, if one or more modalities fail, but at least one other modality gives a usable match/distance score, we can

accept the answers. Alternatively, if the highest-level predicts failure across all modalities, we could take corrective action via perturbations [8] or new sample acquisition. Moving to lower levels within the system, we can fuse the recognition algorithm results before failure prediction is performed, in a blending approach similar to [7]. We can also take advantage of our failure prediction features to fuse across all features after failure prediction has taken place. In our implemented system, all features and fusion techniques are processed through a Support Vector Machine learning module. Each fusion technique is described below.

- Threshold over all decisions across features:

$$T\begin{pmatrix} d(\text{feature}_1) \\ d(\text{feature}_2) \\ \dots \\ d(\text{feature}_n) \end{pmatrix}$$

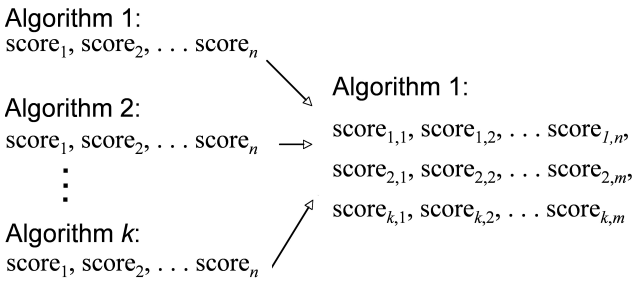
  With this technique, we learn a single threshold over failure prediction decisions across features for a single modality, or for failure prediction decisions across modalities.

- Individual thresholds across all decisions across features:

$$\begin{matrix} T(d(\text{feature}_1)) \\ T(d(\text{feature}_2)) \\ \vdots \\ T(d(\text{feature}_n)) \end{matrix}$$

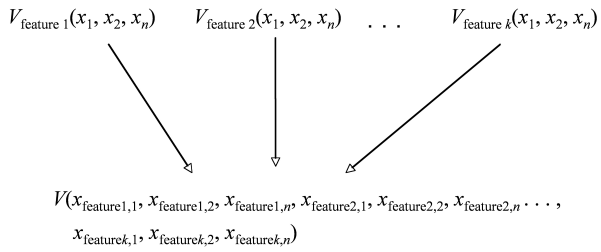
  With this technique, we learn individual thresholds for each failure prediction decision across features for a single modality, or for failure prediction decisions across modalities.

- Combine data from one or more algorithms in another algorithm:



  This technique was used effectively in [7] for a single modality, with some information from one or more algorithms enhancing the performance of another algorithm when added to the data used for its feature computation. Fusion here takes place before feature generation for failure prediction.

- Consider a superset of features as part of one feature vector and fuse the feature vectors that have been calculated for individual features before failure prediction. This blending, including all information for each feature, is an attempt to drive up performance in the machine learning enhancing classification with longer, and ideally more distinct, feature vectors:



The computational efficiency of this system (excluding the underlying recognition system) may be considered in two pieces: training and classification. To sort a series of scores using the quicksort algorithm, $O(n \log n)$ operations are typically required. Computation for our best performing feature, $\Delta_{i,j\ldots k}$ is a simple series of linear operations (subtraction over a set of scores), and is thus $O(M)$ over M score series. The offline training of a SVM is computationally expensive, with a time complexity of $O(M^3)$ over M training examples (feature vectors derived from the M score series). The complete time needed for training the system is $O(nlogn + M + M^3)$ per classifier. SVM classification is a linear operation, of $O(M)$. The complete time needed for classification is $O(n \log n + 2M)$ for fusion before SVM classification, and $O(n \log n + 3M)$ for fusion after SVM classification, where an extra pass over the SVM marginal distances is needed. This linear complexity is well suited for real time systems.

3. EXPERIMENTAL RESULTS

The first set of experiments evaluates the performance of the fusion techniques over the baseline features for failure prediction. The expectation was that the fused prediction techniques would perform no worse than the original features, and in most cases, outperform them. Table 1 shows the data sets used for experimentation. The NIST-multimodal BSSR1 data set [17] was used for all experiments. The subset of this data (fing_x_face) set that provides true multi-modal results is relatively small, providing match scores for 517 unique probes across two face (labeled C & G) recognition algorithms, and scores for two fingers (labeled li & ri) for one fingerprint recognition algorithm. In order to gather enough negative data for training and testing, negative examples for each score set were generated by removing the top score for matching examples. In order to address the limited nature of the multi-modal BSSR1 set, we created a "chimera" data set from the larger face and finger subsets provided by BSSR1, which are not inherently consistent across scores for a single user.

Results for a selection of data across both sets, all algorithms, are presented as FPROC curves in figures 3 - 7. Individual threshold fusion and multiple thresholds fusion (figures 3 and 4), as well as algorithm blending fusion across

Data Set	Training Samples	Test Samples	Face Algo.	Finger Algo.
BSSR1	600	200	2	1
BSSR1 "chimera"	6000	1000	2	1

Table 1. The data breakdown for machine learning with the NIST BSSR1 and BSSR1 "chimera" multimodal data sets.

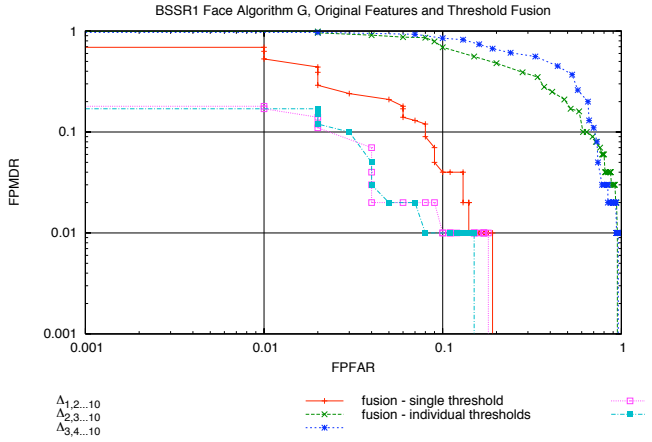


Fig. 3. FPROC curve depicting enhanced failure prediction with single threshold and multiple threshold fusion failure prediction results on BSSR1 face algorithm G. Baseline features provided for comparison.

modalities (figures 5 and 6) improve the performance of failure prediction, compared with the baseline features. Feature blending fusion (figure 7) produced results as good as the best performing feature, but never significantly better. Different combinations of blending were attempted, including mixing all features together (mixed 5), combinations of $\Delta_{1,2,...10}$, $\Delta_{2,3,...10}$, and $\Delta_{3,4,...10}$ (deltas 2 and 3), $\Delta_{2,3,...10}$, $\Delta_{3,4,...10}$, and DCT (deltas-DCT), $\Delta_{1,2}$ and DCT (1,2-DCT). While not improving failure prediction performance, this fusion technique does automatically select the potentially best performing features.

The second set of experiments was designed to evaluate the recognition system's performance after processing by the failure prediction fusion-based system. Figures 8 - 11 show recognition results for single threshold fusion on BSSR1 and the BSSR1 "chimera" set, multiple threshold fusion on BSSR1, and multi-modal algorithm fusion for the BSSR1 "chimera" set in the same ROC format as the results presented in [10] and [11]. All results for fused prediction outperform the original results for each set. Our results on BSSR1 are comparable with the results reported in [10] and [11]. Though due to the small size of BSSR1, this is not as meaningful as a comparison with a much larger set (such as our Chimera set) would be. In the case of multiple threshold fusion for face algorithm C,

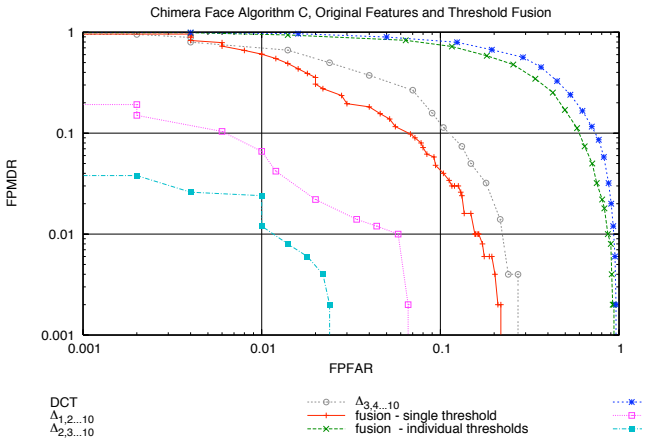


Fig. 4. FPROC curve depicting enhanced failure prediction with single threshold and multiple threshold fusion failure prediction results on BSSR1 "chimera" face algorithm C. Baseline features provided for comparison.

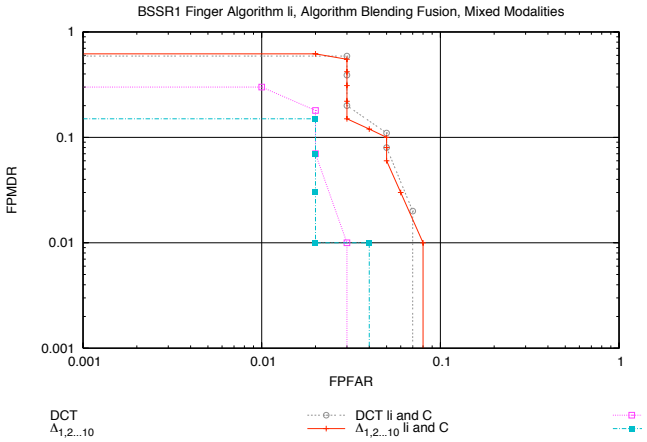


Fig. 5. FPROC curve depicting enhanced failure prediction with mixed modality algorithm blending fusion on BSSR1 finger li's best performing features.

our results are clearly better than [10] and [11]. Moreover, the best results of [10] and [11] were achieved by fusing across all modalities and all algorithms. We can achieve nearly the same results with a single algorithm for a single modality, leaving us more potential for handling larger data, and an excellent option if one or more modalities fail in a multi-modal system. As is shown In figure 10, simply applying decision fusion over the face C and G sets produces a result that is close to MSU's product fusion, and our weaker face G FP fusion results.

4. CONCLUSION

In this paper, we have developed a full multi-modal recognition system integrating fusion-based failure prediction that is suitable for real-time use. With respect to failure prediction,

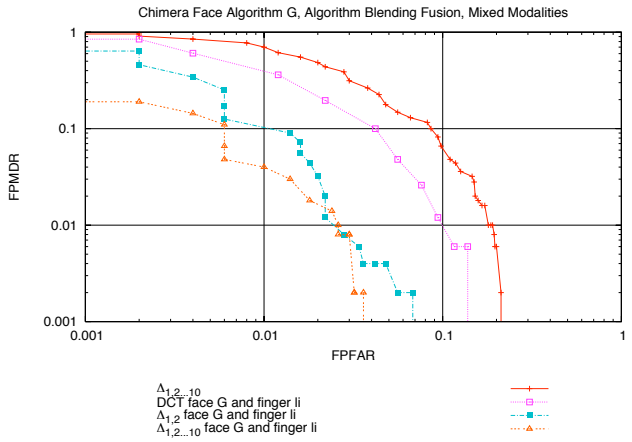


Fig. 6. FPROC curve depicting enhanced failure prediction with mixed modality algorithm blending fusion on BSSR1 "chimera" face algorithm G.

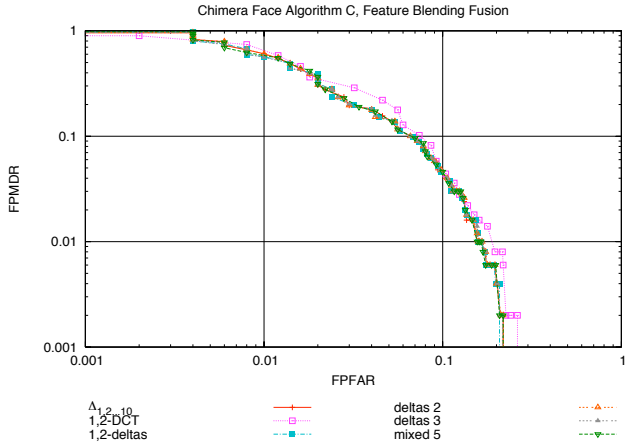


Fig. 7. FPROC curve depicting mixed feature fusion on BSSR1 "chimera" face algorithm C. Fused features perform as well as $\Delta_{1,2...10}$, which allows us to automatically isolate it as the best feature.

four different fusion techniques were introduced and evaluated, using FPROC curves, for this system using two data sets derived from the NIST BSSR1 set. This is the first published use of fusion to improve failure prediction. As the experimental results show, three of the four techniques are able to improve failure prediction, while one only achieves results as good as the best performing baseline feature - but even that allows for the automatic selection of the best performing feature.

This paper shows that using a failure prediction paradigm provides a uniquely effective core for fusion of multiple algorithms or modalities, and significantly enhances recognition system performance. This fusion via failure prediction is comparable to, or better than, previously published fusion

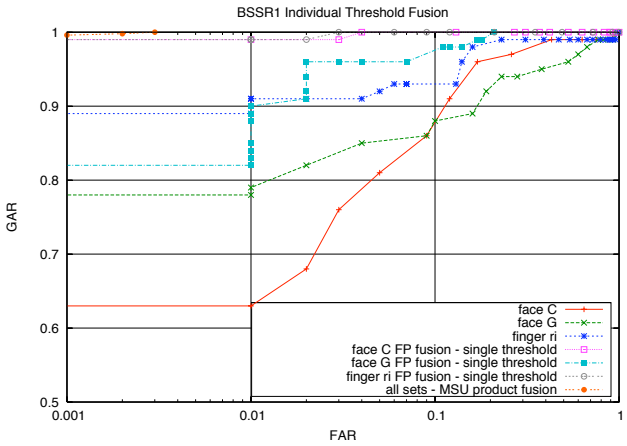


Fig. 8. Performance on all of BSSR1 before and after single threshold fused failure prediction. Our approach, when fusing with finger li, produced perfect performance (due to the small nature of this data set) and is not plotted. Product fusion results, including li, from [11] provided for comparison.

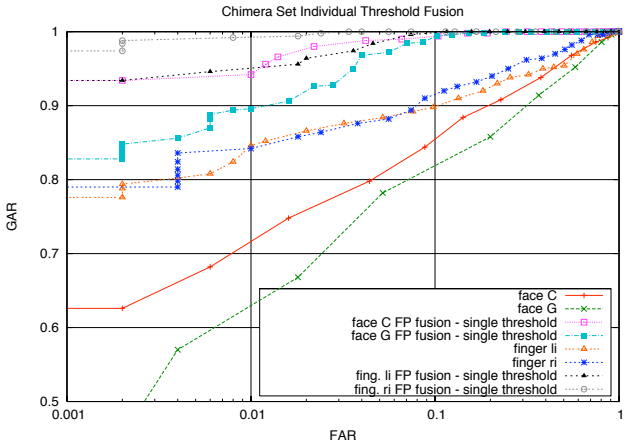


Fig. 9. Performance on all of BSSR1 "chimera" before and after single threshold fused failure prediction.

techniques, while utilizing less data to achieve its results. The utility of multi-modal failure prediction is clear - if one modality has failed, we can fuse information from another modality that has succeeded, and achieve good recognition performance. From a pure quality assessment for fusion, the failure condition might not be evident. We encourage researchers to consider the alternative to quality assessment presented in this paper, especially for systems and problems that can benefit from a per instance prediction of recognition failure.

5. REFERENCES

[1] E. Tabassi, C.L. Wilson, and C.I. Watson, "Fingerprint Image Quality, NFIQ," in *National Institute of Stan-*

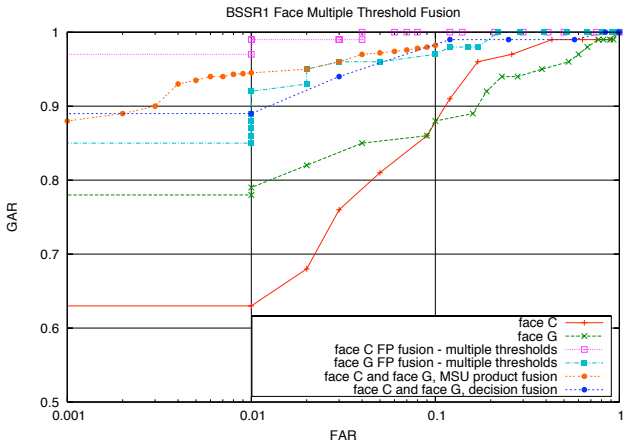


Fig. 10. Performance on BSSR1 face algorithms before and after multiple thresholds fused failure prediction. Product fusion results from [11] and threshold fusion for original scores provided for comparison; our face C prediction clearly outperforms [11].

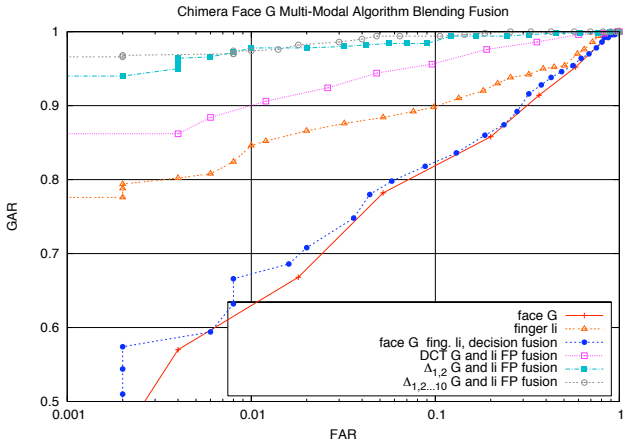


Fig. 11. Performance on BSSR1 "chimera" face algorithm G data before and after multi-modal algorithm blending fusion with finger li. Raw threshold fusion provided for comparison

dards 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] P. Flynn, "Ice Mining: Quality and Demographic Investigations of Ice 2006 Performance Results," 2008, Presentation at the NIST MBGC Kick-off Workshop.

[4] R. Beveridge, "Face Recognition Vendor Test 2006 Experiment 4 Covariate Study," 2008, Presentation at the NIST MBGC Kick-off Workshop.

[5] W. Scheirer, A. Bendale, and T. Boult, "Predicting Biometric Facial Recognition Failure With Similarity Surfaces and Support Vector Machines," in *In Proc. of the IEEE Computer Society Workshop on Biometrics*, 2008.

[6] 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.

[7] 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.

[8] T. Riopka and T. 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.

[9] B. Ulery, A. Hicklin, C. Watson, W. Fellner, and P. Hallinan, "Studies of Biometric Fusion," in *National Institute of Standards and Technology, NISTIR 7346*, 2006.

[10] K. Nandakumar, Y. Chen, S. Dass, and A. Jain, "Likelihood Ratio Based Biometric Score Fusion," *IEEE TPAMI*, vol. 30, no. 2, pp. 342–347, 2008.

[11] S. Dass, K. Nandakumar, and A. Jain, "A Principled Approach to Score Level Fusion in Multimodal Biometric Systems," in *Proceedings of Fifth International Conference on Audio and Video-based Biometric Person Authentication*, 2005, pp. 1049–1058.

[12] K. Nandakumar, Y. Chen, S. Dass, and A. Jain, "Quality-based Score Level Fusion in Multibiometric Systems," in *Proc. of International Conference on Pattern Recognition (ICPR)*, 2006, vol. 4, pp. 473–476.

[13] S. Prabhakar and A. Jain, "Decision-level Fusion in Fingerprint Verification," *Pattern Recognition*, vol. 35, no. 4, pp. 861–874, 2002.

[14] J. Kittler, N. Poh, O. Fatukasi, K. Messer, K. Kryszczuk, J. Richiardi, and A. Drygajlo, "Quality Dependent Fusion of Intramodal and Multimodal Biometric Experts," in *SPIE Biometric Technology for Human Identification IV*, 2007, vol. 6539, pp. 473–476.

[15] M. Vatsa, R. Singh, and A. Noore, "Svm Fusion of Multimodal Biometric Match Scores with Image Quality Metric," *International Journal of Neural Systems*, vol. 17, no. 5, pp. 880–893, 2007.

[16] J. Fierrez-Aguilar, J. Ortega-Garcia, and J. Gonzales-Rodriguez, "Discriminative Multimodal Biometric Authentication Based on Quality Measures," *Pattern Recognition*, vol. 38, no. 5, pp. 777–779, 2005.

[17] National Institute of Standards and Technology, "NIST Biometric Scores Set," 2004, http://www.itl.nist.gov/iad/894.03/biometricscores/.