NISTIR 7346

Studies of Biometric Fusion

Brad Ulery[1]
Austin Hicklin[1]
Craig Watson[2]
William Fellner[1]
Peter Hallinan[3]
[1] Mitretek Systems
[2] National Institute of Standards and Technology
[3] Mitretek Systems Consultant

National Institute of Standards and Technology
Technology Administration, U.S. Department of Commerce

NISTIR 7346

Studies of Biometric Fusion

Brad Ulery
Austin Hicklin
Mitretek Systems

Craig Watson
Image Group, Information Access Division
Information Technology Laboratory

William Fellner
Mitretek Systems

Peter Hallinan
Mitretek Systems Consultant

September 2006

U.S. DEPARTMENT OF COMMERCE
Carlos M. Gutierrez, Secretary
TECHNOLOGY ADMINISTRATION
Michelle O'Neill, Acting Under Secretary of Commerce for Technology
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
William A. Jeffrey, Director

Studies of Biometric Fusion

Brad Ulery,[1] Austin Hicklin,[1] Craig Watson[2], William Fellner,[1] Peter Hallinan[3]

[1] Mitretek Systems
[2] National Institute of Standards and Technology
[3] Consultant to Mitretek Systems

20 July 2006

Abstract

This is an evaluation of different score-level fusion techniques, and the results of a variety of fusion experiments using face and fingerprint data from 187,000 individuals, with matcher scores from three fingerprint and three face recognition systems.

Eight score-level fusion techniques were implemented and evaluated. These differed in effectiveness, in the types of training data required, and in the complexity of modeling of genuine and imposter distributions. The most effective fusion techniques were product of likelihood ratios and logistic regression. Techniques that were nearly as effective were product of False Accept Rates (FARs) and an optimized linear method.

Multi-modal fusion is highly effective: fusing one fingerprint and face resulted in a 64-85% reduction in false reject rate at a constant false accept rate of 0.0001. Multi-instance fusion using fingerprints from multiple fingers is also highly effective: fusing two fingerprints resulted in a 48-90% reduction in false reject rate. Multi-sample fusion using the enrollment of two samples rather than one resulted in a 45-72% reduction in false reject rate. Multi-algorithm fusion using different matchers on the same data resulted in an 8-33% reduction in false reject rate.

Contents

Acknowledgements

This work was performed for the National Institute of Standards and Technology, under contract through the U.S. Department of Interior, Contract NBCH-D-02-0039, Delivery Order D0200390057. Portions of this work draw on Mitretek Sponsored Research on Biometric Fusion conducted in 2004-2005.

We are grateful to a number of colleagues for their reviews and comments on earlier drafts of these documents, including Elham Tabassi and Patrick Grother (NIST); George Kiebuzinski, Larry Nadel, Harold Korves, and Shahram Orandi (Mitretek); and Harris Ulery.

1 Introduction

Biometric fusion is the use of multiple types of biometric data, or methods of processing, to improve the performance of biometric systems. One type of fusion is score-level fusion, which is the combination of matcher scores to improve accuracy. The scores used in fusion can be obtained through the use of multiple types of data for each subject (such as face and fingerprint, or fingerprints from different fingers), multiple samples from each subject, multiple matchers on a single type of data, or combinations of these.

The studies described in this report had two key objectives: to evaluate the effectiveness of different score-level fusion techniques, and to measure the benefits of different categories of fusion on a large quantity of operational face and fingerprint data. The categories of fusion evaluated included multi-modal (finger and face), multi-instance (multiple finger positions), multi-matcher, and multi-sample (multiple enrollments).

This report summarizes the studies we conducted: the reports contained in the Appendices include much more complete background information, details of techniques, and details of the experiments.

2 Data

The primary test dataset was NBDF06 (NIST Biometric Data Fusion 2006). NBDF06 contains operationally collected law enforcement data from approximately 187,000 subjects. At each encounter with a subject, one set of segmented slap fingerprints from all fingers and one face image (mugshot) was collected. NBDF06 includes data from one encounter with 122,000 subjects, two encounters with 60,852 subjects, and three encounters with 4,015 subjects. We refer to the 122,000 subjects without mated faces or fingerprints as imposters, and the 64,867 subjects with mates as genuine subjects. Rolled fingerprints were not used in this study. Appendix A describes in detail the derivation of genuine and imposter scores.

2.1 Face Images

The face images used are frontal, 24-bit color JPEG images compliant with ANSI/NIST-ITL 2000 image format specifications; the images are compliant with Best Practice Application Level 30 requirements [MugshotBP] except for size: image size is typically 384 x 480 pixels, smaller than the 480 x 600 minimum mandated by the Best Practices document. The images have controlled 3-point lighting, 18% gray backgrounds (with some exceptions), and uniform full frontal pose. The face occupies approximately 50% of the width of each image. The expressions are not controlled. A visual review of a sample of the face images shows that the images are fairly typical of recent mugshot photographs, and better than some databases such as BCC or US-VISIT (POE).

2.2 Fingerprints

NBDF06 contains slap livescan fingerprints from all ten fingers that were collected on FBI-certified livescan devices [FBI-Cert]. The devices used 2 to 2.5 inch high platens, smaller than the 3 inch platens required by the identification slaps standard [EFTS 7.1: Appendix F]. The thumbprints were collected as separate images. The four-finger slap images were segmented into individual fingerprint images using the NIST segmenter [NFIS]. Automated measures were used to identify probable segmentation failures, notably cases in which four fingerprints were not present or segmentation boxes touched or overlapped;

these cases, comprising approximately 5% of the source data, were excluded from the dataset. It would not be quite correct to regard these exclusions as failures to enroll (FTE) for several reasons:

- These were probable segmentation failures, which are distinct from failures to enroll: failure to segment indicates a problem with the association of an individual fingerprint image with its finger position, and does not imply anything about the quality or content of the fingerprint features.
- Today, slap segmentation is required at the time of collection in order to be compliant with FBI standards [EFTS 7.1: Appendix N] for identification slaps, which mandates larger scanner platen sizes and segmentation at the time of capture. The cases excluded in this evaluation presumably would have been flagged for recapture at the time of collection if the collection process included slap segmentation.

The dataset as tested should be expected to contain some incorrectly segmented fingerprints, although small samples of manually inspected fingerprints (ones that resulted in low matcher scores) did not reveal segmentation failures; see [SlapSeg] for an evaluation of segmentation accuracy and a discussion of issues in slap segmentation.

Figure 1 shows the distribution of fingerprint quality for the NBDF06 fingerprint data, using the NFIQ metric [NFIQ; NFIS]. 1.3% of all fingerprints were NFIQ 5; 4.8% were NFIQ 4 or 5.

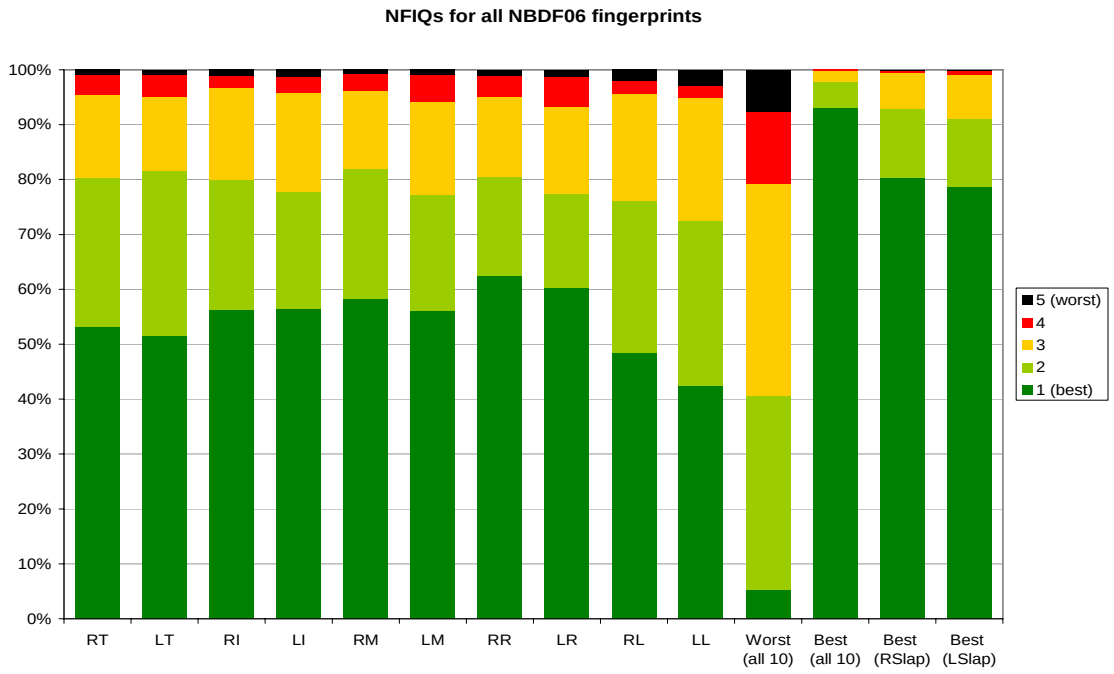

Figure 1: NFIQ results for all genuine and imposter fingerprints, including the best and worst from each subject, and best for each slap from each subject.

2.3 Matchers

Three fingerprint and three face matchers were used:

- The fingerprint matchers were three of the more accurate fingerprint matchers identified in the NIST SDK single finger and two-finger tests [SDK; SDK2]: matcher H, matcher I, and matcher Q. Single-

finger 1:1 matching accuracy for the fingerprint matchers ranged from a true accept rate (TAR) of 93 to 99.5% (depending on the finger position and matcher) at a false accept rate (FAR) of 10^{-4}.

- The face matchers were three recent (c. 2004-5) commercially available face recognition systems, referred to here as A, B, and C. One-to-one matching accuracy for the face matchers ranged from a TAR of 72 to 78% (depending on the matcher) at a FAR of 10^{-4}.

2.4 Data Integrity

Precise measurement of very small error rates on large datasets requires a detailed analysis of the data for potential data integrity issues, such as unconsolidated records (one person with records under different identifiers), misidentified records (records from different people using the same identifier), and swapped and repeated fingerprints (positionally mislabeled). Such issues are often due to collection problems or administrative errors; see [DataQuality] for a full discussion.

In [FpVTE] and [SlapSeg] we found that fusion can be used effectively to detect possible data integrity errors, a process that we continued here. Redundant data (faces and multiple fingers) and multiple accurate matchers provide a basis for locating problems which can then be reviewed manually. To automate detection, metrics were developed based on two methods: score-level fusion and Mahalanobis analysis (which identifies outliers in the score data). Both methods were effective at identifying subject pairs that were misidentified as genuines. Mahalanobis analysis readily revealed partial errors, such as swapped thumbs of subjects whose fused scores were high because the face or other fingerprints matched. Manual review was conducted of cases flagged as potential data integrity errors.

Among the 64,867 genuine subjects, we found 33 data integrity problems (0.051%): 24 had face and fingerprints misidentified (0.037%), and 9 had some but not all fingerprints swapped or repeated (0.014%). This means that a false reject rate (FRR) smaller than 0.037% is not generally possible (except when FAR approaches 1), and that FRR less than 0.051% is possible only when fusion tolerates these errors due to using additional fingerprints or face images that were not misidentified. Among all 186,867 subjects, no unconsolidated records were detected in our analysis of this dataset. This is not surprising, as each non-mate (impostor) was compared to only one subject in the gallery — i.e., a full similarity matrix was not produced. Among the very low-scoring genuines manually reviewed, an additional 46 subjects (0.071%) had some or all fingerprints noted as egregiously poor quality.

The detected data integrity errors were ***not*** removed from the data shown in this report.

3 Score-Level Fusion Techniques

Score-level fusion involves the combination of scores from two or more sources. Figure 2 shows an example of the univariate matcher score distributions (at bottom and at left), and how the joint distribution (scatterplot) enables greater discrimination between genuines and imposters. Each orange line corresponds to a decision threshold based on the unfused scores, which in this case fails to discriminate between the imposters and many genuines. This example shows linear fusion boundaries as well a polynomial boundary, each of which is superior to the univariate decisions. More elaborate — and accurate — decision boundaries follow the topology of the intersecting genuine and imposter distributions. Note that each decision boundary produces a single point on an ROC; each fusion technique implicitly defines a family of such boundaries from the sample data.

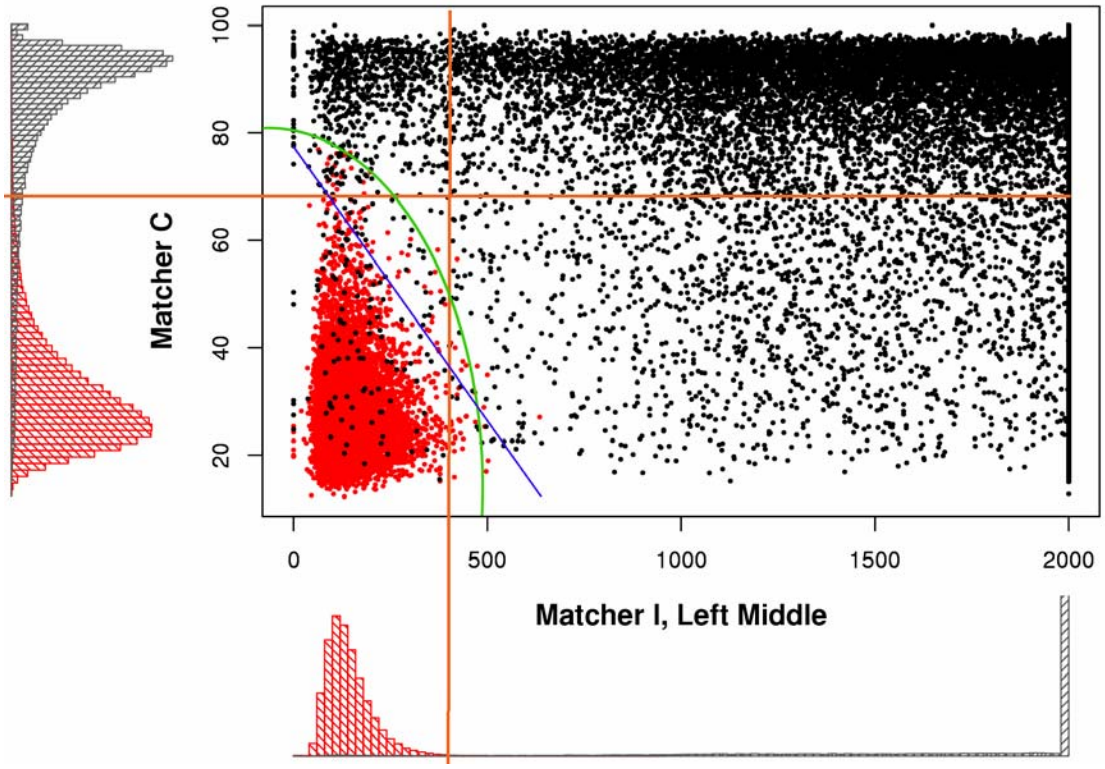

Figure 2: Example of matcher score distributions, with genuines in black and imposters in red. The lines show examples of decision boundaries separating genuines from imposters: each such boundary corresponds to a point on an ROC. Note that in this chart 86% of all fingerprint genuine scores are at the maximum (rightmost) value.

The presence of genuines in the upper left and lower right quadrants reveals that fusion succeeds primarily due to the fact that genuines are sometimes missed using one input, but can be identified using another input. More definitive exclusion of impostors is not the primary mechanism. This general rule appears to hold true for all the data studied.

This distribution in Figure 2 should not be taken as characteristic of all of the data: distributions vary substantially depending on the data being fused, as shown in Figure 3.

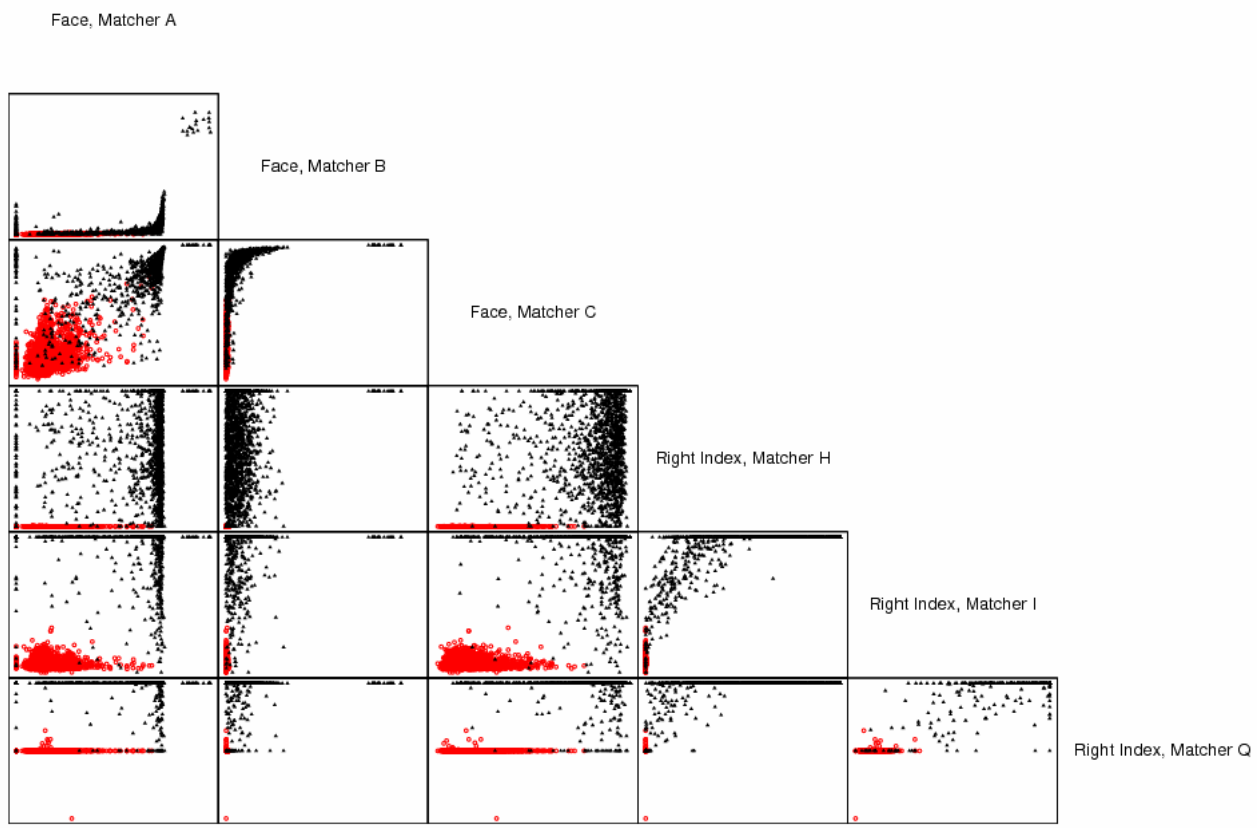

Figure 3: Examples of joint distributions comparing all face and fingerprint matchers, with genuines in black and imposters in red. Scale has been reduced to accentuate overall form of the distributions.

3.1 Implemented Techniques

In this study, we surveyed a variety of proposed methods of fusion, and selected eight for evaluation:

- ***Likelihood ratio-based methods***. The Neyman-Pearson (NP) lemma [Neyman-33] defines a criterion for the optimization of an ROC based on likelihood ratios.[1] While this approach is theoretically optimal, implementation assumes knowledge of joint genuine and imposter distributions. In practice, the accuracy of an NP implementation depends on accurate modeling of the distributions. Two likelihood ratio-based methods were implemented:
  - Product of Likelihood Ratios, as implemented here, is based on a multi-stage modeling process: probability density functions were separately modeled for each genuine and impostor distribution, using variable bandwidth kernels, log-linear tail tapering, and specific handling of spikes in the distributions; likelihood ratios were computed from these models for each matcher; scores were transformed to their likelihood ratios, then simply multiplied. Note that all of the

[1] The Neyman-Pearson lemma states that optimal decision boundaries are defined by equal likelihood contours. These can be visualized as analogous to elevation contour lines on a topographic map. If you 1) take an X,Y scatterplot of genuine and imposter scores, 2) map each point in the scatterplot to the ratio of genuines to imposters (likelihood ratios), and 3) plot those ratios in the Z dimension, then the Neyman-Pearson lemma states that the topological contours that follow a given "altitude" (a fixed likelihood ratio) correspond to optimal decision boundaries.

complexity in this implementation is in the modeling of distributions, rather than fusion per se, and that use of the univariate (not joint) distributions makes the simplifying assumption that fused scores are independent.

  - Logistic Regression is a standard statistical technique that directly models the likelihood ratio. Successful implementation requires accurate curve fitting of density distributions, which requires statistical expertise, but is supported by typical statistical packages: the log of the genuine/imposter density ratio is modeled (e.g. as a low-order polynomial function), then estimated from the training data by principle of maximum likelihood; density ratios are modeled independently for each matcher, and fusion is performed by adding the normalized scores (log likelihoods). Logistic Regression generally uses joint sample data. In this analysis, however, the independence assumption was highly effective, and various trial models showed that coefficients for carriers involving multiple scores were almost always statistically insignificant, so our models relied exclusively on univariate distributions.

- ***FAR-based methods***. In cases when the genuine distribution is unavailable or unreliable, scores can be normalized by transformation to False Accept Rates (using only the univariate imposter distributions) before fusion. However, correctly modeling the imposter distribution usually remains a complex process. In our implementations, FARs were modeled using the same processes used in Product of Likelihood Ratios. The techniques we implemented were Product of FARs, Min of FARs, and Max of FARs. Min and Max methods are decision-level fusion: Min is AND rule decision-level fusion, and Max is OR rule decision-level fusion (where decision thresholds are set to result in equal FARs).
- ***Linear methods***. These involve addition of weighted scores. The methods we implemented do not require modeling of score distributions.
  - Simple Sum of Raw Scores assumes the inputs have comparable scale, distribution, and strength — which is only a valid assumption in some cases, such as fusion of left and right index fingers scored by one matcher.
  - Simple Sum of Z-Normalized Scores requires only small samples of univariate imposter distributions to perform normalization. Scores are normalized to a mean of 0 and standard deviation of 1, then added without weighting.
  - Best Linear is a weighted sum of z-normalized scores. The weights are based on an optimal slope (hyperplane) determined empirically on joint training data. This solution entails iteratively rotating the decision boundary and evaluating TAR at a fixed FAR. This is conceptually simple and does not require modeling, but does require joint training data.

It is critical to note that the effectiveness of Product of Likelihoods, Logistic Regression, and the FAR-based techniques depends greatly on how well the distributions are modeled.

3.2 Evaluation of Techniques

Figure 4 compares the performance of these techniques on several fusion tasks.

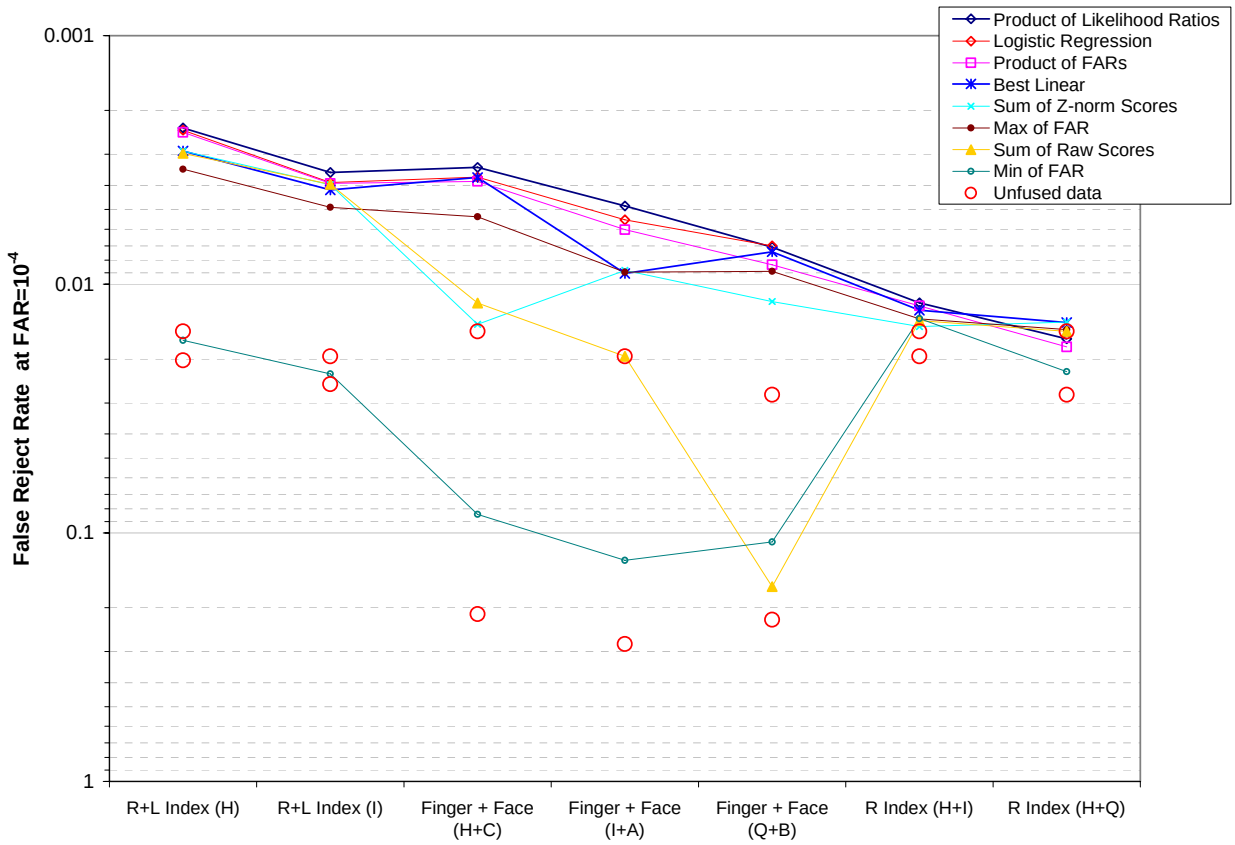

Figure 4: Eight techniques compared at FAR=10^{-4} on a variety of fusion tasks. The legend shows techniques in order of average performance for this data.

In general (summarizing a much broader set of results than shown on this chart), we found:

- Product of Likelihood Ratios was consistently most accurate, but most complex to implement.
- Logistic Regression, Product of FARs, and Best Linear usually performed nearly as well as Product of Likelihood Ratios.
- Sum of Z-Normalized Scores performs well in the multi-instance tasks.
- Max of FARs performs near the top. This is notable because this is essentially decision-level fusion, countering expectations that decision-level fusion would not be effective.
- Sum of Raw Scores makes simplifying assumptions that rarely hold except in the case of single matcher, multi-instance tasks.
- Min of FARs performs poorly.
- The distinctions between techniques were most pronounced for multi-modal tests (finger and face), where the potential improvement was great and the distributions highly dissimilar.

Figure 5 and Figure 6 show examples of ROCs for sample tests.

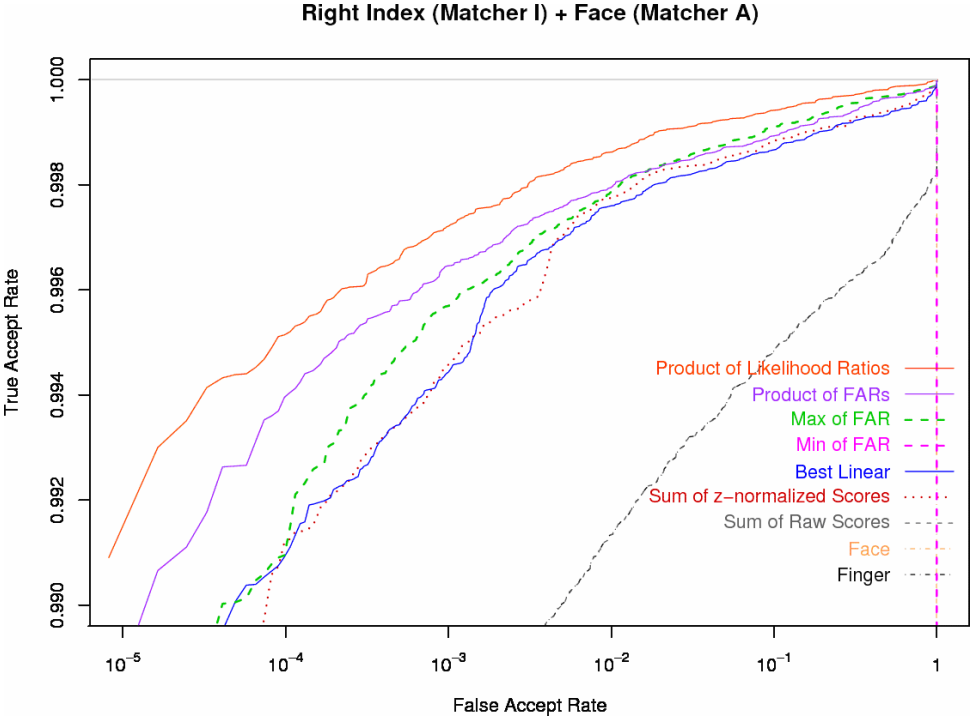

Figure 5: Comparison of ROCs for the implemented techniques, for an example test of multi-modal fusion. Logistic Regression was not included in this series of tests. The Face and Min of FAR graphs are superimposed. Due to the scale, little of the unfused face and fingerprint ROCs are visible on this chart: face (A) had a TAR=0.720 at FAR=10^{-4}; right index (I) had a TAR=0.981 at FAR=10^{-4}.

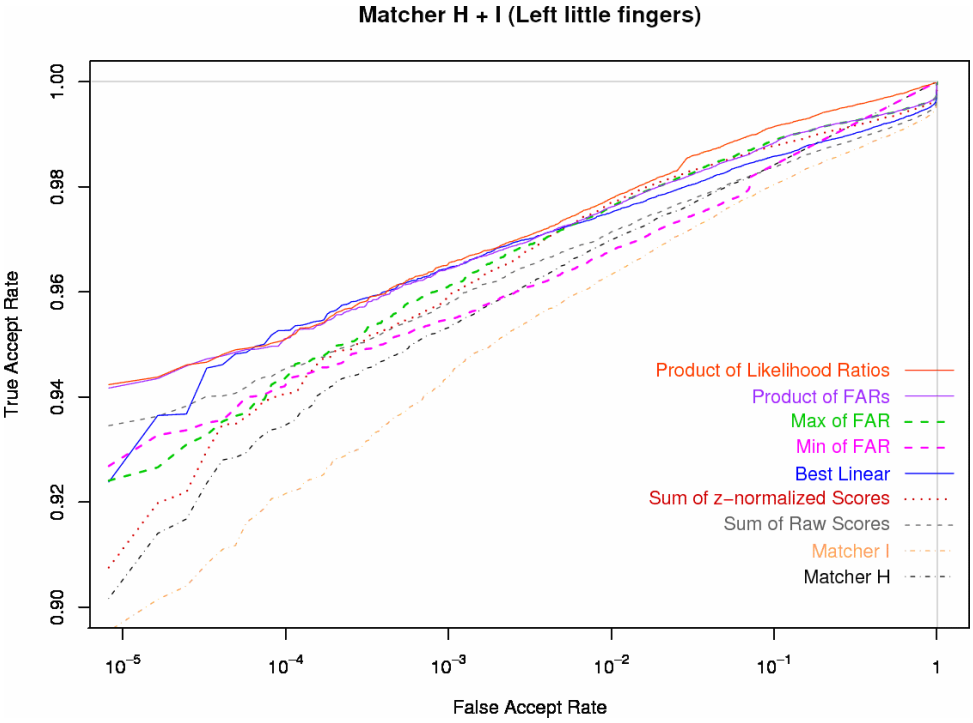

Figure 6: Comparison of ROCs for the implemented techniques, for an example test of multi-modal fusion. Logistic Regression was not included in this series of tests.

3.3 Considerations in the Choice of Techniques

The choice of fusion techniques cannot just be governed by overall accuracy. The different fusion techniques — as implemented — have substantially different requirements in the quantity and type of training data, the expertise required, and the complexity of modeling. Table 1 summarizes these requirements.

	Data required	Statistical expertise required	Complexity (As implemented)
Product of Likelihood Ratios	Univariate Genuine and Imposter	Yes	High
Logistic Regression	Univariate Genuine and Imposter	Yes	Medium
FAR-based methods	Univariate Imposter	Yes	High
Best Linear	Joint Genuine and Imposter	No	Medium
Sum of Z-Normalized Scores	Univariate Imposter (small amount)	No	Low
Sum of Raw Scores	None	No	None

Table 1: Summary of data and modeling requirements for fusion techniques

4 Results of Experiments

We conducted experiments to measure the benefits of different categories of fusion on the NBDF06 dataset. The categories of fusion evaluated included multi-modal (finger and face), multi-instance (multiple finger positions), multi-matcher, and multi-sample (multiple enrollments). Product of

Likelihood Ratios was used as the fusion technique unless otherwise noted. Except for the multi-sample experiments, the results described here all use 64,867 genuines and 122,000 imposters.

4.1 Multi-Modal: Face and Fingerprint Results

Score-level fusion of face and a single fingerprint is consistently very effective, as shown in Figure 7. Face and fingerprint data are nearly independent: the fused line is very close to — but still separable from — the independent chimera[2] line. Some evaluations use chimeras for test data, when it is not possible to use face and fingerprint data collected from the same individuals; these results show that the independence assumption implicit in the use of chimeras will often be a valid approximation.

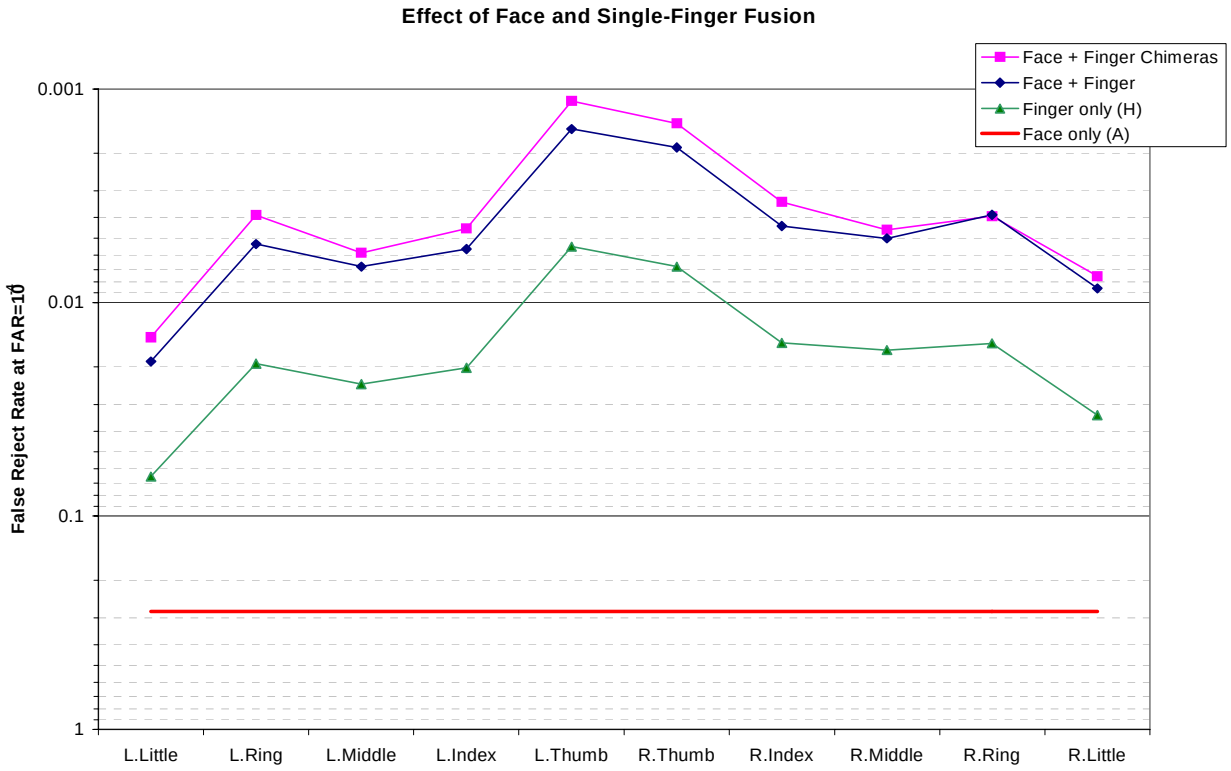

Figure 7: Effect of fusing face and single fingerprint scores for each finger position, using face matcher A, fingerprint matcher H and Product of Likelihood Ratios fusion. The fusion (blue) line shows the effect of fusing face (red) and fingerprint (green). The chimera (pink) line shows the effect of fusing face scores from one subject and fingerprint scores from another — this result indicates what the effect would have been if the face and fingerprint scores were independent.

Table 2 summarizes the effect of fusion for all single fingerprint and face fusion experiments, given all ten finger positions, three fingerprint matchers, and three face matchers. The results are measured as reduction of false reject rate for a fixed false accept rate. Note the uniformity of results: fusing face and single fingerprint scores always reduced FRR by more than half, and usually by 75%.

[2] Chimeras are composites of data representing virtual "subjects" that combine biometrics from multiple individuals (selected at random).

	Single finger + face		
	H+face	I+face	Q+face
Min	68%	71%	64%
Median	74%	76%	75%
Average	74%	77%	74%
Max	80%	84%	79%

Table 2: Summary of fusion of each of the ten fingers against each of the three face matchers, in terms of reduction in FRR where FAR = 10^{-4}, relative to the stronger of the inputs. [3]

4.2 Multi-Instance: 2-Fingerprint Results

Score-level fusion of two fingerprints is very effective, as shown in Figure 8. Note however, that fingerprint scores from different fingers are not independent: the distance between the fused and chimera lines shows that dependence substantially limits the benefits of fusion.

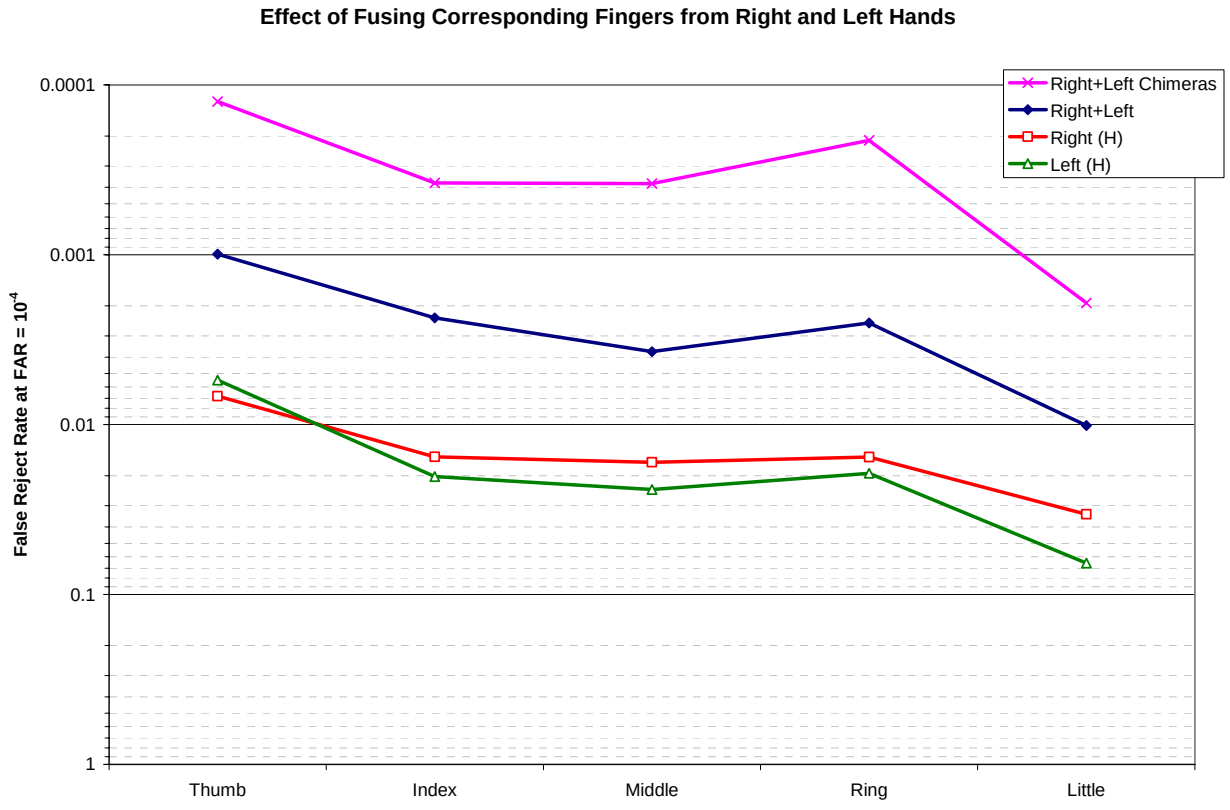

Figure 8: Effect of fusing scores for two corresponding fingerprints from right and left hands (matcher H, Product of Likelihood Ratios fusion). The fusion (blue) line shows the effect of fusing right (red)

[3] For example, when fusing Left Thumb and Right Index, FRR_{LT}=0.0055, FRR_{RI}=0.0155 and the fused FRR_{LT*RI}=0.0007, then the improvement in FRR = (min(0.0055, 0.0155) – 0.0007)/ min(0.0055, 0.0155) = (0.0055-0.0007)/0.0055 = 87%. This metric was computed at FAR = 10^{-4}, but was not highly sensitive to this operating point.

and left (green) fingerprint scores. The chimera (pink) line shows what the effect would have been if the face and fingerprint scores were independent.

Table 3 summarizes the effect of fusion for all combinations of two distinct fingerprints, given all ten finger positions, and three fingerprint matchers. The results are measured as reduction of false reject rate for a fixed false accept rate. On average, fusing two fingerprint scores results in about the same improvement in accuracy as fusing face and fingerprint scores: 77% rather than 75%. However, these results have more variability than the face + fingerprint results shown above in Table 2: 48-90% rather than 64-84%. This variability reflects matcher accuracy and varying levels of score correlation among pairs of finger positions.

	Two fingers		
	H fingers	I fingers	Q fingers
Min	59%	48%	51%
Median	83%	79%	72%
Average	82%	78%	71%
Max	90%	90%	84%

Table 3: Summary of fusion for all 45 pairwise combinations of fingers, in terms of reduction in FRR where FAR = 10^{-4}, relative to the stronger of the inputs.

Table 4 summarizes the results of [SDK2] using the same metrics. Those results correspond closely to ours.

2-finger SDK Report	
Min	58%
Median	80%
Average	77%
Max	89%

Table 4: Summary of two-finger fusion results from [SDK2][4] , in terms of reduction in FRR where FAR = 10^{-4}, relative to the stronger of the inputs. (Based on index finger results from ten matchers and four datasets, using simple sum of raw scores.)

4.3 Multi-Instance: N-Fingerprint Results

Figure 9 shows the results of fusing fingerprint scores in various combinations, using each of the three fingerprint matchers. Since these results exceed an FRR of 0.001, we show the data integrity limits discussed in Section 2.4: the upper line is the 0.037% limit due to the 24 subjects with face and fingerprints misidentified (0.037%), and the lower line is the 0.051% limit including the additional 9 subjects with some swapped or repeated fingers. The upper line is a hard limit; the lower line may be a limit depending on which fingers are used.

[4] Derived from Tables 3 and 6 in [SDK2].

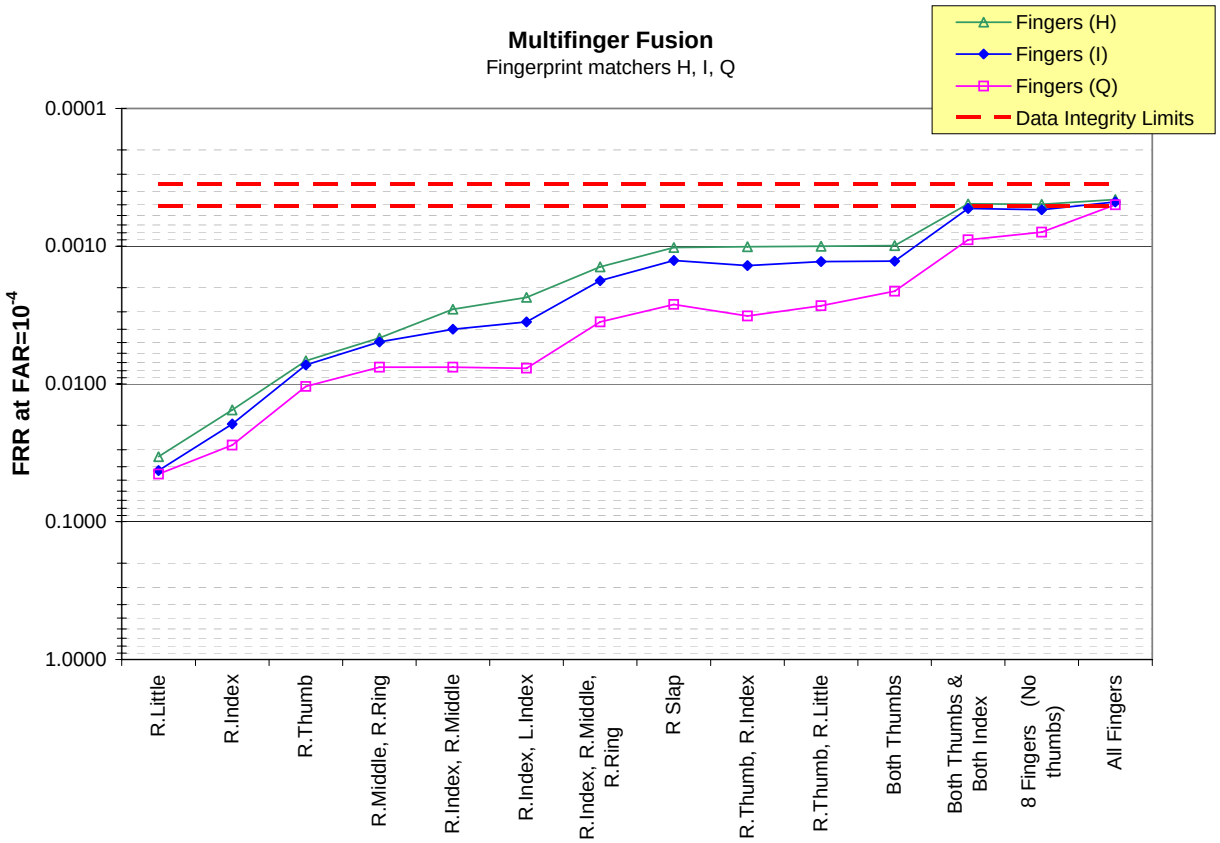

Figure 9: Fusion of various groupings of multiple fingers, showing data integrity limits (based on the number of known database errors) for fingerprint matchers H, I, & Q and Product of Likelihood Ratios fusion.

A number of conclusions may be drawn from these results:

- It is an oversimplification to say that fusing more fingers improves accuracy. The combinations of fingers used are at least as important as the number of fingers used.
- Thumbs are substantially more effective than the other fingers:
  - Thumbs offer as much performance advantage over index fingers as index fingers offer over little fingers. Two thumbs are much more accurate than two index fingers.
  - A four-finger slap is approximately as effective as a thumb and any other finger.
- For Matchers H and I, the combination of both thumbs and both index fingers reaches the data integrity limit. Note this combination has one fingerprint from each of the four images captured in a full set of slap fingerprints. In some experiments, the combination of both thumbs and one other finger reached the data integrity limit.
- As more inputs are fused and accuracy approaches 100%, the maximum achievable accuracy is limited by data integrity problems (misidentifications, swapped prints, missing images).

The primary reason that accuracy is sensitive to specific combinations of fingers is because of correlations between fingerprint scores. Figure 10 shows correlations between genuine scores for one fingerprint matcher. Note the correlations between neighboring fingers, among the four fingers collected in each slap, and between corresponding fingers on right and left hands (faint diagonal from top right to bottom left). Some of these correlations are specific to the way that these fingerprints were collected: if both

thumbs were captured in a single image, they would be more correlated; if the fingers were collected separately rather than in a single slap image, they would be less correlated.

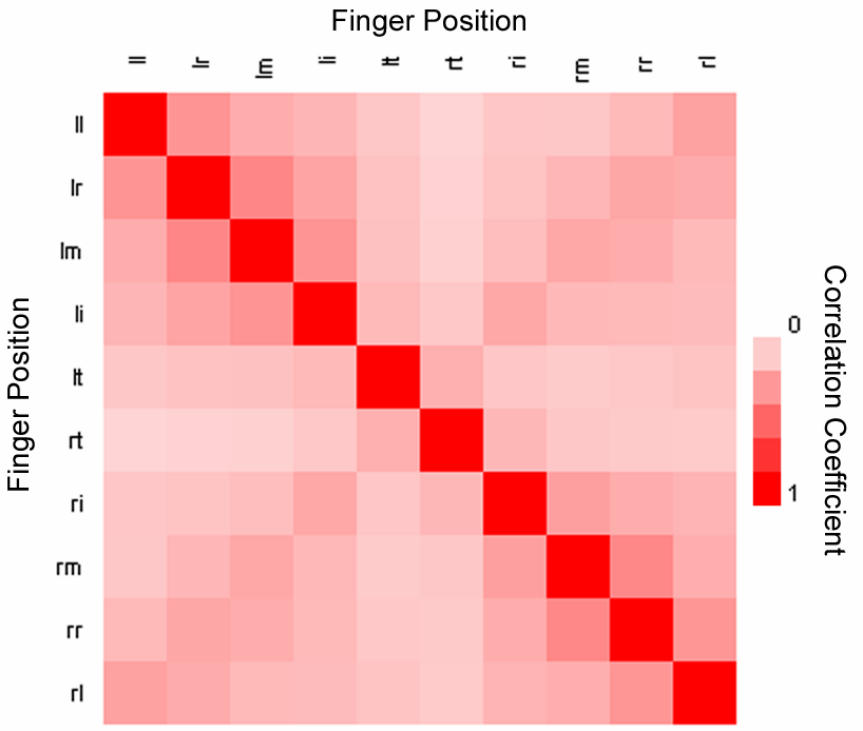

Figure 10: Correlations between genuine scores for Matcher I by finger position — left little (ll) to right little (rl). Darker colors represent higher correlations. Values range from 0.17 to 0.47 (ignoring the identity diagonal).

4.4 Multi-Instance/Modal: N-Fingerprint and Face Results

Figure 11 shows the effect of adding face to the results shown above in Figure 9. Combining face with fingers is beneficial in all cases, although the relative benefit of adding face decreases as FRR approaches the data integrity limit.

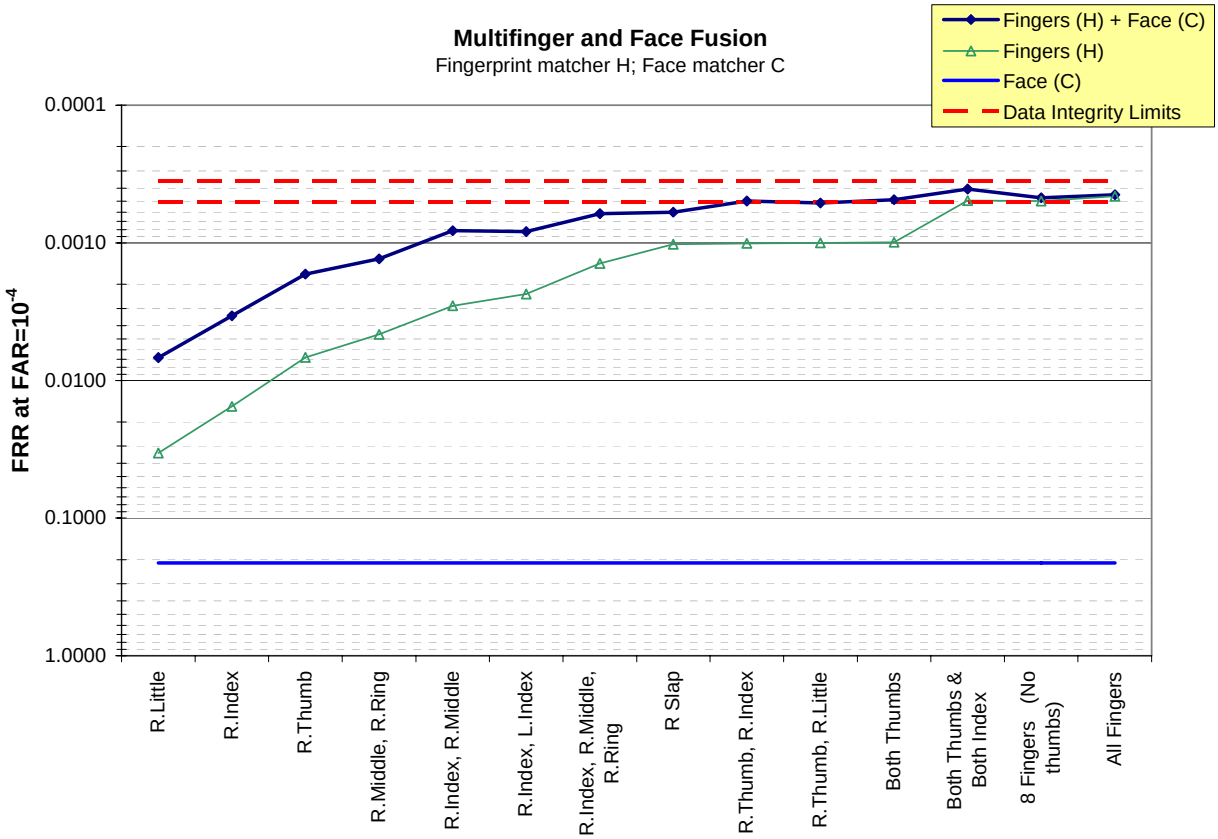

Figure 11: Fusion of various groupings of fingers, with and without face. This uses fingerprint matcher H and face matcher C, the most effective combination. (Product of Likelihood Ratios fusion)

4.5 Multi-Matcher Results

We evaluated the effect of fusing scores from multiple matchers, given the same input samples. Table 5 shows that for pairs of fingerprint matchers at fixed FAR=0.0001, FRR was generally reduced by 8-33%, varying by finger position; for pairs of face matchers, FRR was reduced by 10-13%.

	Face	Fingerprint		
		H+I	H+Q	I+Q
Min	10%	14%	8%	9%
Median	10%	25%	20%	20%
Average	11%	25%	16%	20%
Max	13%	33%	32%	32%

Table 5: Reduction in FRR where FAR = 10^{-4}, for pairwise matcher fusion using the Product of Likelihood Ratios technique. Fingerprint results are for all ten finger positions. Face results summarize the three pairwise combinations of matchers.

Figure 12 and Figure 13 show the effect of three-way matcher fusion.

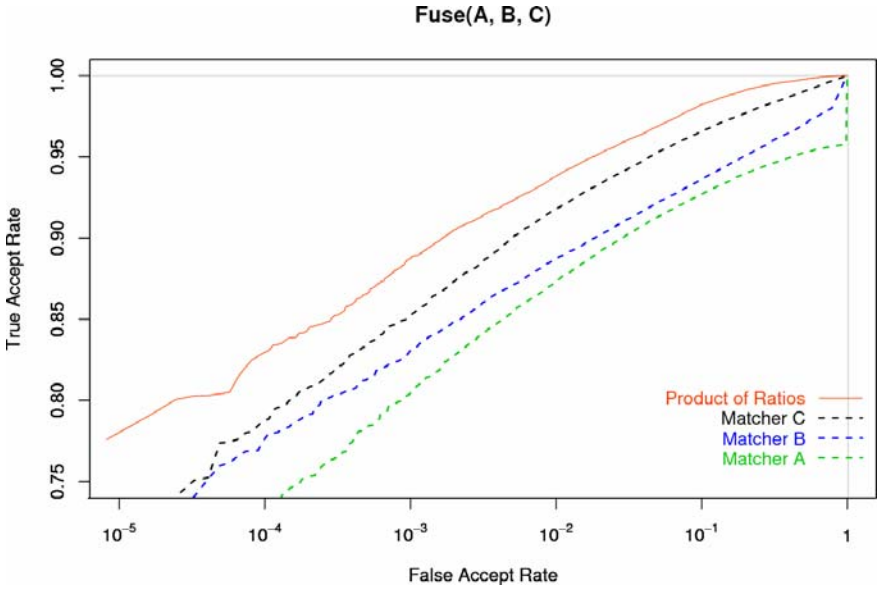

Figure 12: Fusing all three face matchers reduced FRR by 20% relative to Matcher C (at FAR=10^{-4}). Note that much of the benefit at high FAR (between 1 and 10^{-1}) is directly due to Likelihood Ratio normalization of Matcher A, not fusion.[5]

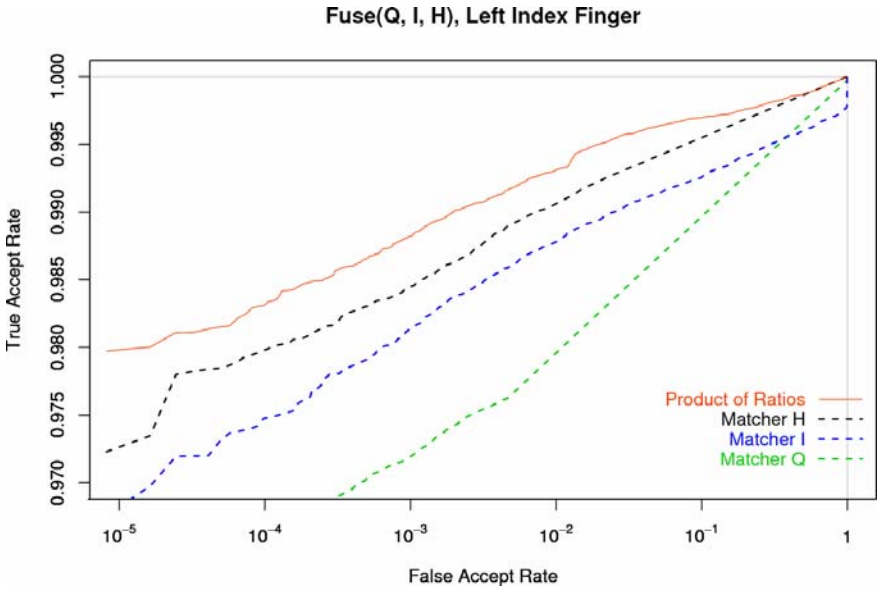

Figure 13: Fusing all three matchers on left index fingers reduced FRR by 17% relative to Matcher H (at FAR=10^{-4}).

[5] In some cases, normalization of scores by transformation to likelihood ratios improves ROC performance even without fusion. Face matcher A is a case in point, due to a spike in the genuine distribution at the minimum score value, which translates into a vertical drop at FAR=1. A similar effect but to a lesser extent occurs for fingerprint matcher I in Figure 13.

Matcher fusion often does produce a substantial increase in accuracy, although typically much less than that for multi-instance or multi-modal fusion. This should be expected due to the greater degree of data independence for instance and modal fusion. Since single-instance multi-matcher fusion uses the same data for both (or all) matchers, the independent information that is necessary for effective fusion must come from differences (if any) between matchers. Thus any improvements in accuracy reflect differences in the matchers that might be exploited either through score-level fusion or further improvement of existing matcher technology.

4.6 Multi-Sample Results

A multi-sample biometric system uses more than one sample from each biometric instance, such as multiple fingerprint images from each of a person's fingers. Multi-sample fusion from the use of multiple enrollments is likely to be of interest since it leverages existing data rather than requiring the collection of additional data. Therefore the cost and complexity of implementing this form of multi-sample fusion is likely to be much less than that of multi-modal or multi-instance fusion. The method used in this study was to measure how accuracy would be affected if a gallery retained two fingerprint samples per finger position per subject rather than just one.

NBDF06 includes 4,015 subjects with three fingerprint sets each. The fingerprint sets were captured in different collection encounters, on different dates. The mated (genuine) multi-sample data used in this analysis was comprised of the three segmented slap fingerprints per finger position for each of these 4,015 subjects. The non-mated (imposter) multi-sample data consists of 396,210 subject pairs selected from among the off-diagonal scores of the similarity matrices. The matchers used were the H, I, and Q fingerprint matchers. As the fused scores have the same score distribution (same finger, same matcher), they were fused by Simple Sum of Raw Scores (and by Max of Raw Scores, which produced essentially the same results).

Figure 14 shows the effect of multi-sample fusion: FRR was reduced by about half.

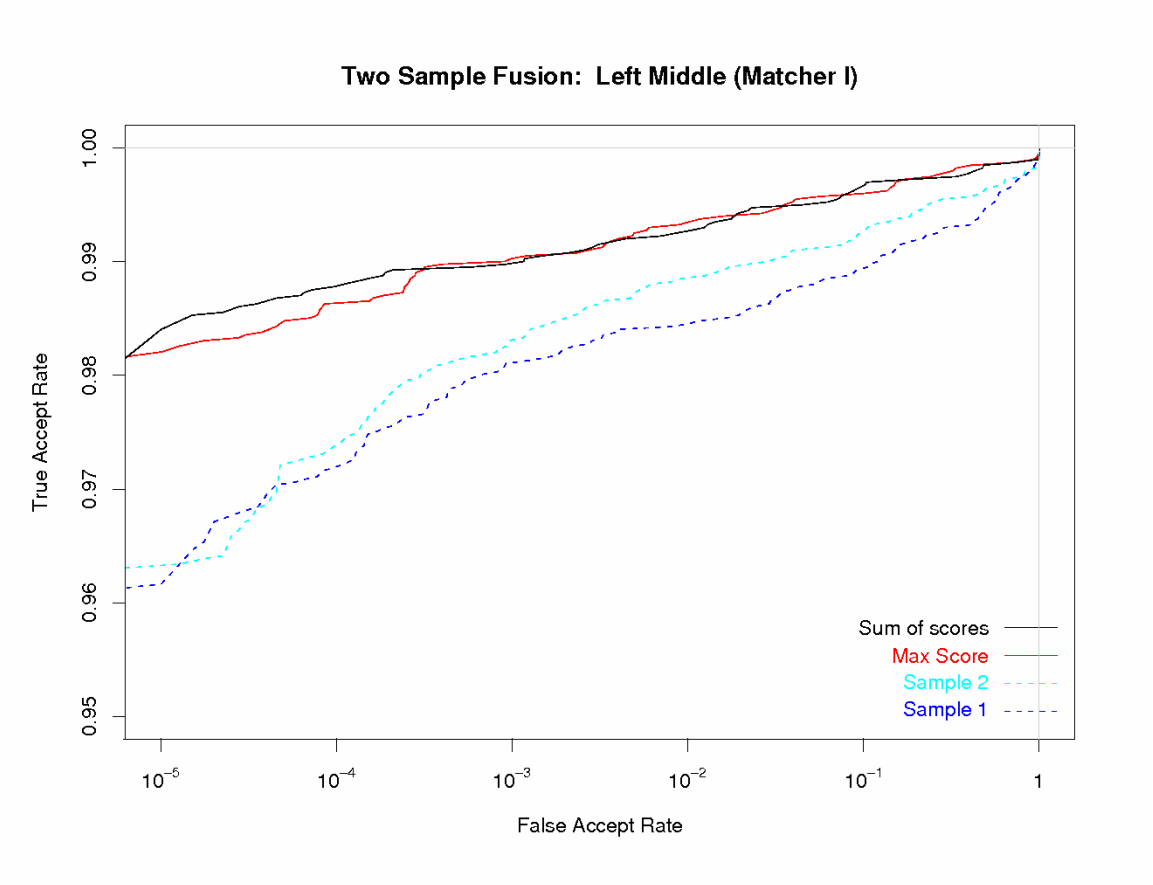

Figure 14: Example of effects of multi-sample fusion, for matcher I, left middle fingers.

Table 6 shows summary results of multi-sample fusion for all finger positions, by matcher. The improvement in FRR ranged from 45% to 72%.

	H	I	Q
Min	53%	49%	45%
Median	57%	56%	52%
Max	70%	72%	57%

Table 6: Effect of multi-sample fusion, using Raw Sum of Scores technique, in terms of reduction in FRR at FAR=10^{-4}, relative to the stronger input. Results for each matcher are computed over all ten finger positions.

5 Conclusions

Eight score-level fusion techniques were implemented and evaluated. These differed in effectiveness, in the types of training data required, and in their requirements for modeling genuine and imposter distributions.

- The most effective fusion techniques were Product of Likelihood Ratios and Logistic Regression, which are implementations of the theoretically optimal Neyman-Pearson Lemma. Product of Likelihood Ratios involved complex, detailed modeling of score distributions. Logistic Regression

achieved similar results using a standard statistical technique. Both techniques require statistical tools, training, and a substantial amount of training data.

- Techniques that were nearly as effective were product of FARs and Best Linear. Product of FARs requires modeling the non-mated (imposter) distribution, but does not require mated (genuine) data. Best Linear is a conceptually simple technique that requires joint training data, but does not require modeling of distributions.
- For cases in which the input scores are of similar strengths and distributions, such as fusing two index fingers using a single matcher, the choice of fusion technique had minimal effect on accuracy.

A variety of fusion experiments using face and fingerprint data were conducted, using the most accurate of the techniques we implemented, Product of Likelihood Ratios. The baseline performance of biometric systems is largely determined by matcher accuracy and sample quality. The number of fused scores, the extent to which those scores are correlated, and the fusion techniques used determine the additional benefits of fusion. As more inputs are fused and accuracy approaches 100%, the maximum achievable accuracy is limited by data integrity problems (such as misidentifications or swapped or repeated images). Fusing scores from at least three or four separately collected samples (finger instances and/or face) largely eliminates the effect of poor image quality.

- Multi-modal score-level fusion (face and fingerprint) was consistently highly effective, because the face and fingerprint matching scores are nearly independent. Fusing matcher scores from one fingerprint and face resulted in a 64-85% reduction in 1:1 false reject rate at a constant false accept rate of 0.0001. For example, an improvement of false reject rates from 1.0% (using fingerprints) to 0.25% (using face in addition) is a 75% reduction in false reject rates. Improved face image quality should be expected to result in further accuracy improvements.
- Multi-instance score-level fusion (using fingerprints from multiple fingers) was consistently highly effective. Fusing two fingerprints resulted in a 48-90% reduction in false reject rate at a constant false accept rate of 0.0001. The accuracy of multi-fingerprint fusion is dependent on which fingers are used, the number of fingers, and the correlations of the fingerprint scores. Example findings: fusing of scores from a thumb and any other finger was as effective as fusing the scores from four fingers on one hand (slap); the maximum possible accuracy on this dataset (0.05% false reject rate, at a false reject rate of 10^{-4}) could be achieved using both thumbs and one other finger. Fusion of scores from two fingers was about as accurate as fusion of scores from face and one fingerprint; fusion of scores from a four-finger slap was not as accurate as fusion of scores from face and two fingerprints.
- Multi-sample score-level fusion, such as from the use of multiple enrollments, was consistently effective. The use of single-finger gallery samples from two enrollments rather than a single enrollment resulted in a 45-72% reduction in false reject rate at a constant false accept rate of 0.0001, using Sum of Raw Scores fusion. Multi-sample fusion may be of interest since it leverages existing data rather than requiring the collection of additional data.
- Multi-matcher score-level fusion, using two matchers on the same data, was much less effective than the other methods, but may be of interest since it does not require the collection of additional data. A 10-13% reduction in false reject rate was achieved at FAR= 0.0001 by fusing scores from two face matchers; an 8-33% reduction was achieved using fingerprint matchers.

Although score-level fusion has clearly been shown to be effective, this does not necessarily mean that fusion can be successfully implemented in every situation. The extent to which the benefits of fusion can be realized in practice depends on

- The availability of multi-biometric data and/or multiple matchers
- The accuracy of the matchers
- The correlation of the scores

- The representativeness and quantity of training data
- A detailed understanding of score distributions
- How fusion is implemented (e.g. the choice of fusion technique, details of its implementation, and its role in the system architecture)

6 References

[DataQuality] A. Hicklin and R. Khanna; "The Role of Data Quality in Biometric Systems"; February 2006; http://www.mitretek.org/Role_of_Data_Quality_Final.pdf

[EFTS-7.1] Federal Bureau of Investigation, Criminal Justice Information Services (CJIS); Electronic Fingerprint Transmission Specification (EFTS); IAFIS-DOC-01078-7.1; May 2, 2005. http://www.fbi.gov/hq/cjisd/iafis/efts71/cover.htm

[FBICert] "Products Certified For Compliance with the FBI's Integrated Automated Fingerprint Identification System Image Quality Specifications"; http://www.fbi.gov/hq/cjisd/iafis/cert.htm

[FpVTE] C. Wilson, A. Hicklin, H. Korves, B. Ulery, M. Zoepfl, M. Bone, P. Grother, R. Micheals, S. Otto and C. Watson; "Fingerprint Vendor Technology Evaluation 2003"; NIST Interagency Report 7123. June 2004. http://fpvte.nist.gov/

[MugshotBP] "Best Practice Recommendation for the Capture OF Mugshots"; Version 2.0; September 23, 1997. (http://www.itl.nist.gov/iad/894.03/face/bpr_mug3.html) Note: this is identical to "Best Practice Application Level 30" in the forthcoming ANSI/NIST ITL-1 2006 standard.

[Neyman-33] J. Neyman and E. S. Pearson; "On the problem of the most efficient tests of statistical hypotheses"; Philosophical Transactions of the Royal Society, Series A, Containing Papers of a Mathematical or Physical Character, 231, p. 289-337; 1933.

[NFIQ] E. Tabassi, C. Wilson, and C. Watson; "Fingerprint Image Quality"; NIST Interagency Report 7151, August 2004. (http://fingerprint.nist.gov/NFIS/ir_7151.pdf)

[NFIS] C. Watson, M. Garris; "NIST Fingerprint Image Software". http://fingerprint.nist.gov/NFIS/

[SDK] C. Watson, C. Wilson, K. Marshall, M. Indovina and R. Snelick; "Studies of One-to-One Fingerprint Matching with Vendor SDK Matchers"; NIST Interagency Report 7221; April 2005. http://fingerprint.nist.gov/SDK/ir_7221.pdf

[SDK2] C. Watson, C. Wilson, M. Indovina and B. Cochran; "Two Finger Matching With Vendor SDK Matchers"; NIST Interagency Report 7249; July 2005. http://fingerprint.nist.gov/SDK/ir_7249.pdf

[SlapSeg] B. Ulery, A. Hicklin, C. Watson, K. Kwong; "Slap Segmentation Evaluation 2004"; NIST Interagency Report 7209. March 2005. http://fingerprint.nist.gov/SlapSeg04/ir_7209.pdf