Studies of Biometric Fusion
Appendix B

Effectiveness of Score-Level Fusion

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

[1] Mitretek Systems

[2] Independent consultant to Mitretek Systems

[3] National Institute of Standards and Technology

20 July 2006

Abstract

This three-part appendix contains the results of experiments measuring the effectiveness of different categories of fusion: multi-modal (finger and face), multi-instance (multiple finger positions), multi-matcher, and multi-sample (multiple enrollments).

Appendix B.1: Score-Level Fusion of Face and Multiple Fingerprints

This is an analysis of the effectiveness of multi-modal (finger and face) and multi-instance (multiple finger positions) score-level fusion, focusing on the extent to which different biometric modalities and instances are independent, and the effect of that independence on the accuracy of fusion. It includes detailed analyses of the effects of fusing scores from varying combinations of fingers, and the effect of fusing face and fingerprint scores. This paper provides large-scale empirical evidence that score-level fusion using multiple finger positions is highly effective, as is fusion of fingers and face: fusing two fingerprints or one fingerprint and face generally resulted in a 50-90% reduction in false reject rate (FRR) relative to the stronger of the two inputs at a constant false accept rate (FAR).

Appendix B.2: Score-Level Fusion of Multiple Matchers

This is an analysis of the effectiveness of score-level matcher fusion, in which multiple matchers produced scores from comparisons of the same pairs of images. Both face and fingerprint matchers were evaluated. 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. A 10-30% reduction in missed identifications (relative reduction in false rejection rate) was achieved. Due to data correlation, algorithm fusion is less effective than either instance or mode fusion, but can still improve accuracy given limited data.

Appendix B.3: Score-Level Fusion of Multiple Fingerprint Samples

This is an analysis of the effectiveness of score-level sample fusion, which uses more than one sample from each biometric instance, such as multiple fingerprint images from each of a person's fingers. Multi-sample fusion is of operational interest because it can improve matching accuracy without additional collection of data by retaining successfully matched probes in the gallery in addition to the originally enrolled sample. False reject rates were reduced by 45% to 73%.

Contents

Appendix B.1: Score-Level Fusion of Face and Multiple Fingerprints

Contents

1 Introduction: Combining Face and Multiple Fingerprints

Optimizing the design of multi-biometric systems requires understanding the extent to which the multiple inputs or methods contribute additional, complementary information to the decision making process. Much of the effectiveness of fusion depends on the extent to which different biometric modalities and instances are independent.

This paper investigates four questions:

Q1 To what extent are face and fingerprint scores independent?
Q2 To what extent are fingerprint scores from different fingers independent, and does a law of diminishing returns govern the number of fingers to use?
Q3 What are effective combinations of fingerprint and face biometrics to fuse?
Q4 Is it reasonable to train and/or evaluate biometric systems on chimeras?

Chimeras are composites of data representing virtual "subjects" that combine biometrics from multiple individuals. Chimeras are often used in evaluations that lack sufficient real data. For example, an evaluation that has fingerprint data from one set of subjects and face data from another set of subjects may choose to treat the data as if the faces and fingerprints came from the same individuals. The assumption behind the use of chimeras is that face and fingerprint data are fully independent.

In this study, we used chimeras for a different purpose: to measure the extent of data independence. We first measured fusion performance in the standard way, so that each set of face and fingerprints all originated from a single subject. We then determined what the effect of fusion would be if the fused biometric scores were independent by using chimeras, created by associating the face and fingerprints of different people. By comparing the chimera results to the actual results, we could determine the extent of data independence.

To answer the four questions, we conducted two studies: we first conducted a small-scale, exploratory study on the NIST BSSR1 public domain data set, then used the lessons learned to conduct a large-scale study on the NBDF06 dataset. We include the results of both studies because they involve different datasets and matchers: the BSSR1 data is publicly available; and the large NBDF06 dataset allowed much more precise measurements.

Datasets and experimental design are discussed in Appendix A. Unless otherwise noted, all experiments used Product of Likelihood Ratios as the fusion technique, which was the most accurate of the techniques we implemented; see Appendix C for details.

2 Data Independence and Score-Level Fusion

Biometric score independence has been identified as an issue in the published literature, but with little large-scale empirical analysis. Given the general dearth of data available, researchers frequently make assumptions regarding the independence of data. These assumptions may be largely inconsequential for small empirical studies where measurement precision is limited, but can be significant to the design and evaluation of highly accurate systems.

The correctness of an independence assumption has various implications:

- Whether predictions of the benefits of fusion are valid
- Whether the joint score data contains more information than any one score set alone (so that fusion has the potential to be beneficial)
- Whether the fusion technique is well-suited to the problem (so that the potential can be realized)

Corresponding to each of these implications is a distinct question regarding the validity of the independence assumption:

- Does sufficient dependence exist to invalidate predictions?
- Does sufficient independence exist to justify fusion (operational cost-benefit)?
- Is the fusion technique sufficiently robust with respect to dependence in the data to achieve good results?

In order to evaluate the validity of the independence assumption as it pertains to score-level fusion, we must consider that

- Dependence is essentially a characteristic of the ***score*** data that is to be fused. Such characteristics can vary greatly according to the operational scenario, according to the choice of biometrics, data collection procedures, accuracy of matchers, etc.
- The significance of any observed dependencies in the scores is determined by how those scores will be used. This includes the choice of fusion technique and operational objectives, such as accuracy and robustness.

These topics are not to be confused with how data is sampled for an evaluation. For instance, in sampling, dependence often results from reusing subjects, as when N scores are produced by comparing one probe subject to N gallery subjects [FRVT, FpVTE, SDK]. Nor are we discussing cross-class dependencies: higher scores are assigned to genuines than to imposters.

These topics of independence arise frequently in the literature on biometric fusion. Assumptions of independence are common in both empirical and theoretical works. Lacking access to large multi-modal (or multi-instance) datasets, many researchers have created chimerical datasets based on an assumption of independence [Jain-99b; Fierrez-Aguilar-03; Snelick-03; Indovina-03; Wang-03; Poh-05c; Poh-05e; Snelick-05]. In several cases, the researchers use chimeras while recognizing the uncertainty of the method. Many in the biometrics community have questioned the validity of using chimeras, e.g., [Poh-05e].

Theoreticians (often explicitly) [Dass-05; Griffin-05; Kittler-98; Jain-05; Poh-05d; Scott-05] and analysts (sometimes implicitly) make independence assumptions when they develop or select fusion techniques. A clear example of this is the "product rule" as applied to the ratio of posterior probabilities (Product of Likelihood Ratios, discussed in Appendix C).

2.1 Sources of Dependence

In fusion, biometric inputs derive from samples that belong to the same individual, were often collected at one encounter (except, e.g., multi-sample gallery images), and have undergone some common processing. Based on these factors, the extent of independence should be expected to vary among the different categories of fusion:

- ***Multi-modal*** data (such as faces and fingerprints) are generally believed to be fairly independent (Question 1 investigated by this study). If the different modalities are collected at a single encounter, they may have dependencies due to factors specific to that encounter such as quality problems due to a hurried subject, an uncooperative subject or an incompetent operator.
- ***Multi-instance*** data (such as multiple fingerprints) should not be expected to be as independent as multi-modal data. Multi-instance data will have the encounter-specific factors mentioned above as well as additional sources of dependencies specific to the subject, the genetic relationships between fingers, and the use of a single collection device. For slap fingerprints, dependencies will be particularly high because of simultaneous collection. Some of the correlations between fingers are shown in Figure 1. 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).

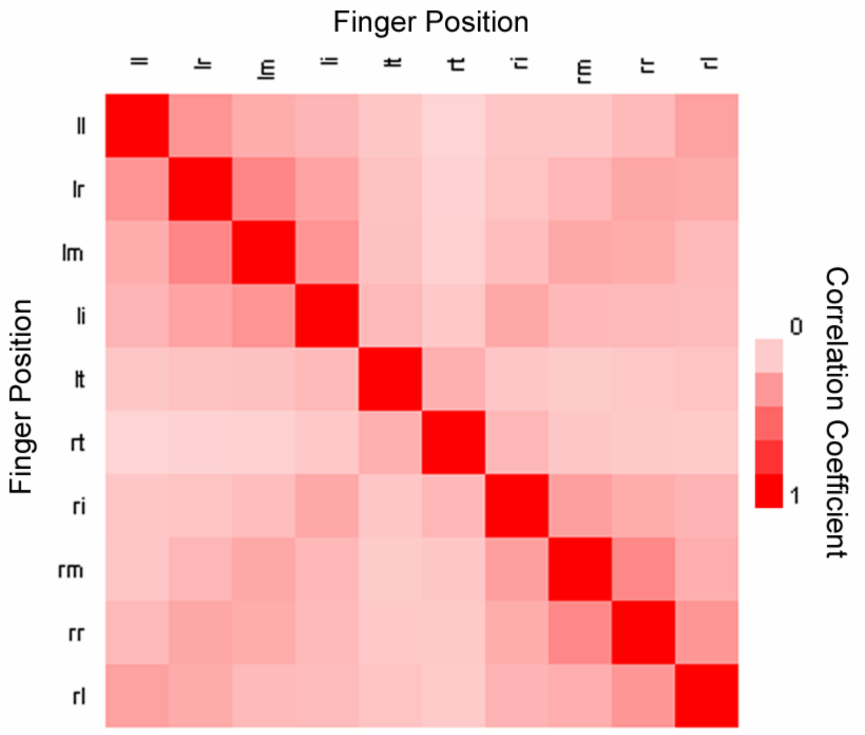

Figure 1: Correlations between genuine scores by finger position (left little (ll) to right little (rl)) for slap fingerprints from c. 65,000 subjects. Darker colors show higher correlations. Values range from 0.17 to 0.47 (ignoring the identity diagonal).[1] *[NBDF06 data; Matcher I]*

- ***Multi-sample*** data (such as face images from a video sequence) should be expected to be moderately correlated, because of sample-specific variability and subject-specific dependence. An analysis of multi-sample variance [Goats] showed that match scores "cannot generally be attributed to intrinsic characteristics of a person's fingerprints, but should be attributed to collection problems or other characteristics of the specific fingerprints used." Multi-sample analysis was not possible as part of this study at the time of writing.
- ***Multi-algorithm*** data (such as results from different matchers on common samples) should be expected to be highly dependent, because the algorithms are working with the same data. The effectiveness of algorithm fusion depends on the independence of several factors: the biometric features used, the feature extraction algorithms, and the matching algorithms. (Matcher fusion is discussed in Appendix B.2)

2.2 Examples

Figure 2 shows examples of two scatterplots of joint score distributions, purposely shown at small scale to accentuate the overall form and the effect of dependence. The plot on the left shows two modalities

[1] The upper right and lower left of the table are mirrors of each other: e.g. the RL column and RL row report the same results.

(BSSR1 data); the plot on the right shows two matchers on the same sample (NBDF06 data). In these plots, imposter scores are shown in red; genuine scores are shown in black.

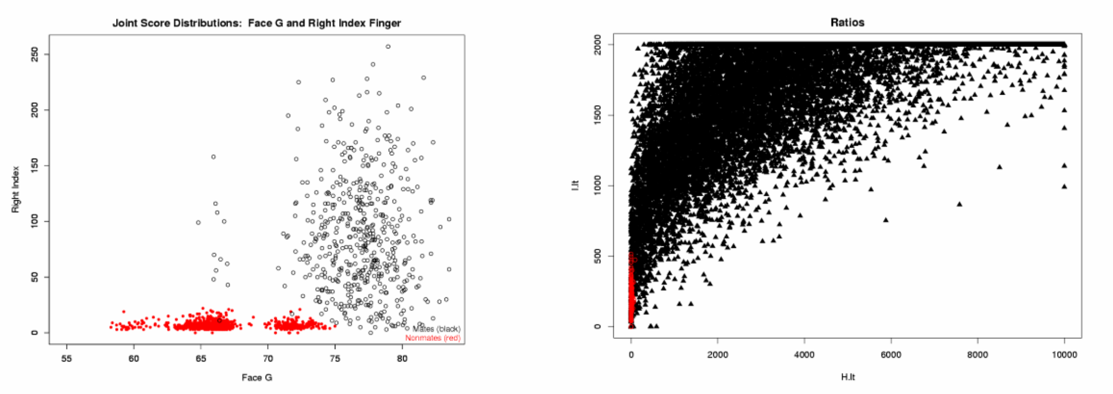

Figure 2: Effect of data independence on scatterplots: the left figure shows essentially independent data (face and finger), while the right figure shows much more dependent data (the same fingerprint images, two different matchers). Note how the dispersion of data points decreases as dependence increases. Completely dependent data would form a line or curve when plotted in this way.

[Left: BSSR1 dataset, matchers V & G; Right: NBDF06 dataset, matchers H & I, left thumbs]

3 Large-Scale Study of Face and Fingerprint Fusion using NBDF06 Data

For this study, pairs of face and fingerprint images from the NBDF06 dataset were compared using three face matchers (identified as A, B, and C) and three fingerprint matchers (identified as H, I, and Q). This data is described in *IV: Description of Datasets and Pre-Fusion Data Characteristics*. The resulting matcher scores were combined by the Product of Likelihood Ratios method as described in Appendix C. This method of fusion approximates Neyman-Pearson optimization to minimize the false reject rate (FRR) at a specific false accept rate (FAR); it was the most effective technique among several investigated. Throughout this paper, results are summarized by reporting FRR at FAR=10^{-4}.

3.1 Multimodal Dependencies: Face vs. Finger

This experiment tests the hypothesis that face and fingerprint scores are independent.

3.1.1 Experimental Design

Test subjects have one face and one finger. There are ten finger positions and nine different combinations of face and fingerprint matchers, so we run 90 trial sets. Each trial set contains two trials, one actual subject and one chimera.

Type of Trial	How score pairs are constructed	Subjects / Trial	Trials / Type
Actual subject	Pair the face score and one finger score of an individual .	64,867 genuines 122,000 imposters	1
Chimera	Pair the face score of one person with one finger score of another person .	64,867 genuines 122,000 imposters	1
Total number of trials (per type of input pair)			**2**

Table 1: Description of trials, and number of trials per set

3.1.2 Results

Figure 3 shows that for the fusion of one fingerprint with face, there is little difference in accuracy between actual subjects and chimeras. There is a slight advantage of chimeras over actual subjects, but that is small in comparison to the main effects of fusion. These results show that faces and fingerprints are nearly independent, and therefore, to a great extent validate the use of chimeras where the face image comes from one individual and the finger(s) from another for purposes of system design and evaluation.

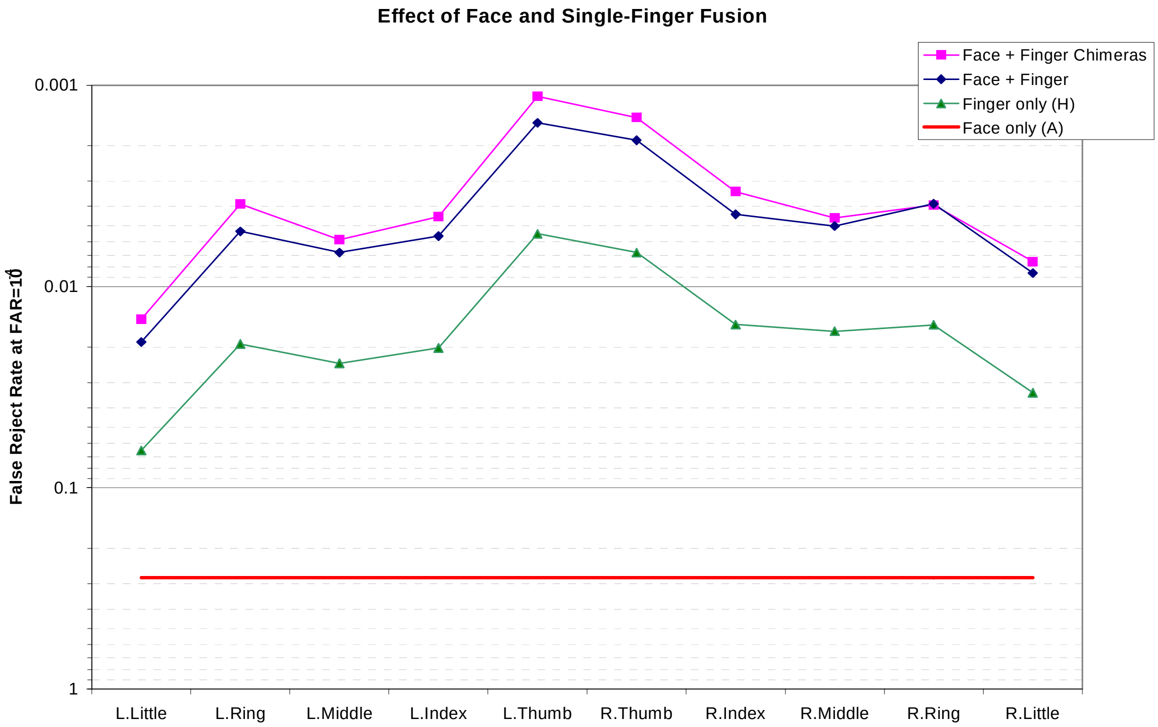

Figure 3: Finger and face fusion comparing actual finger+face fusion with finger+face chimera fusion: face and finger modalities are nearly independent. *[NBDF06 dataset, matchers H & A]*

Several additional observations can be made from Figure 3: the index, middle and ring fingers perform similarly, thumbs are much better, and little fingers are much worse. Note that the right fingers are slightly better than the left, but that the left thumbs are slightly better than the right thumbs. The contribution from the face modality can be seen to be essentially constant from finger to finger, as expected due to independence.

3.2 Between-Hand Dependency: Finger vs. Finger

This experiment investigates the dependencies between corresponding fingers of each hand, including all five possible finger pairs.

3.2.1 Experimental Design

Subjects have two corresponding fingers (e.g., two thumbs), one from each hand. For the chimeras, each finger is from a different person. There are five different types of subjects and three different matchers, so we run 15 trial sets. Each trial set contains two trials, one actual subject trial and one chimera trial.

Type of Trial	How score pairs are constructed	Subjects / Trial	Trials / Type
Actual subject	Pair the scores of corresponding right and left fingers of one individual .	64,867 genuines 122,000 imposters	1
Chimera	Pair the scores of corresponding right and left fingers of two distinct individuals .	64,867 genuines 122,000 imposters	1
Total number of trials (per type of input pair)			**2**

Table 2: Description of trials, and number of trials per set

3.2.2 Results

Figure 4 shows that correlations among fingerprint scores significantly limit the benefits of fusion: chimera finger pairs perform much better (by roughly an order of magnitude) than finger pairs from actual subjects. Fusing corresponding fingers from opposite hands is clearly better than using single fingers. Right-hand fingers yield better results than left hand fingers, except for the thumb: all three matchers achieve greater accuracy on left thumbs than on right thumbs on this dataset.

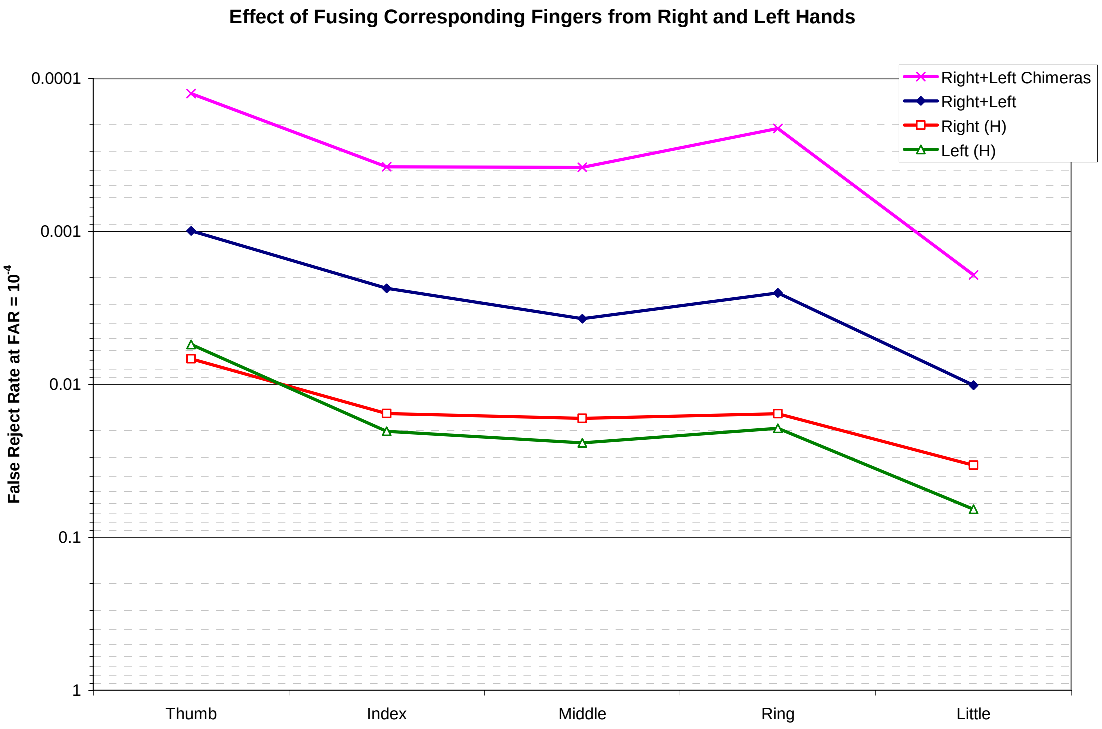

Figure 4: Effect of fusing two corresponding fingerprints from right and left hands, using fingerprint matcher H and Product of Likelihood Ratios fusion. The fusion (blue) line shows the effect of fusing right (red) and left (green) fingerprints. The chimera (pink) line shows what the effect would have been if the face and fingerprint data were independent. *[NBDF06 dataset, matcher H]*

Figure 5 compares the finger + face data of Figure 3 to the finger + finger data of Figure 4, and reveals very similar accuracies for these two combinations.[2] This can be explained by the fact that although the face is a weaker biometric than the second finger, the information it provides is nearly fully independent.

[2] One notable exception: the face is much more effective than the little finger.

Should face matcher technology continue to improve, the balance will tip in favor of using face as a second biometric for greater accuracy.

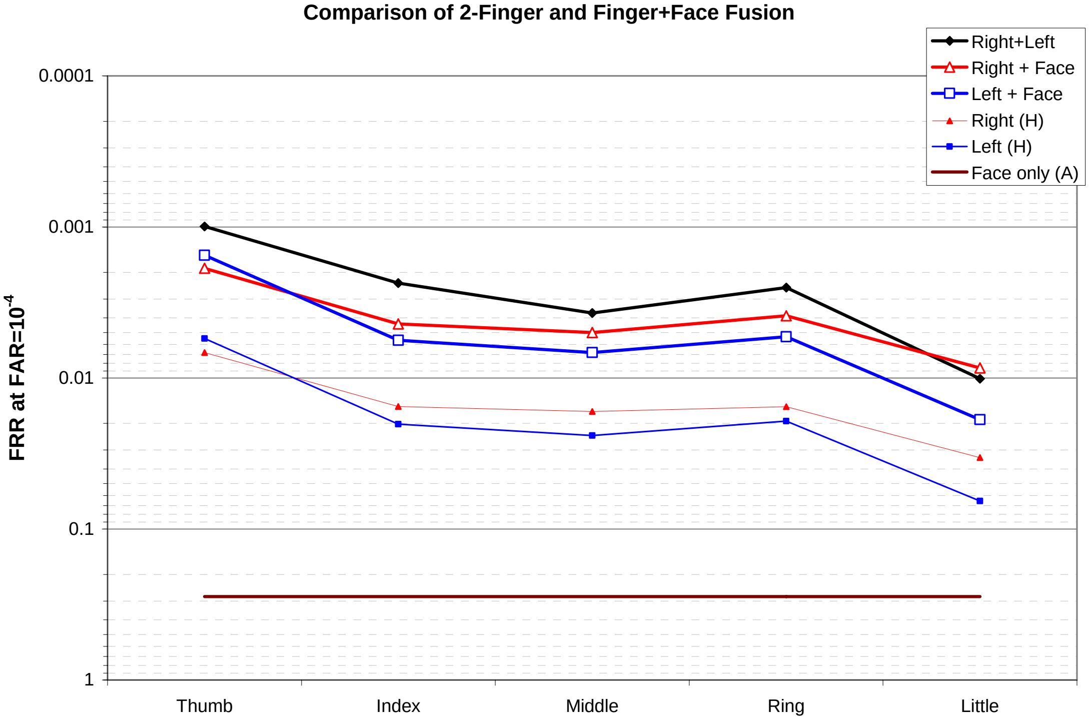

Figure 5: Comparison of 2-finger and finger + face fusion. *[NBDF06 dataset, matchers H & A]*

3.3 Between-Hand vs. Within-Hand Dependencies

This experiment tests whether between-hand dependencies (fingers from right and left hands) are weaker or stronger than within-hand dependencies (fingers from the same hand).

3.3.1 Experimental Design

If within-hand dependencies exceed between-hand dependencies then we expect the performance of a biometric system on fingers from the same hand (within-hand models) to be worse than its performance on fingers from different hands (between-hand models). As an example, within-hand dependencies might be caused by failure to place the hand properly on the collection device, resulting in low mate scores for multiple fingers. However, the question is complicated by the possibility that the result could differ for any pair of finger types (e.g. middle and index, or thumb and little).

In this experiment there are 10 different combinations of fingers and three different fingerprint matchers, so we run 30 trial sets. Each trial set contains two within-hand models and two between-hand models.

Type of Trial	How score pairs are constructed	Subjects / Trial	Trials / Type
Left-Right	Pair the scores of left finger x and right finger y from one individual.	64,867 genuines 122,000 imposters	1
Left-Left	Pair the scores of left finger x and left finger y from one individual.	64,867 genuines 122,000 imposters	1
Right-Left	Pair the scores of right finger x and left finger y from one individual.	64,867 genuines 122,000 imposters	1
Right-Right	Pair the scores of right finger x and right finger y from one individual.	64,867 genuines 122,000 imposters	1
		Total number of trials (per type of input pair)	2

Table 3: Description of trials, and number of trials per set. "x" and "y" are different fingers types (i.e. thumb, index, middle, ring, or little).

3.3.2 Results

These results show that dependencies between adjacent fingers on the same hand do substantially limit the benefits of fusing adjacent fingers. Nevertheless, fusing adjacent fingers is highly beneficial when compared to not using fusion.

Section 3.2 showed strong effects of dependencies between two fingers, one from each hand. These results show that dependencies are slightly greater when the two fingers are selected from the same hand. Notice that, although the finger combinations differ, Figure 4 and Figure 6 can be compared by examining the results for Index, Middle and Ring fingers which all perform similarly. See Table 6 for tabular results for all pairwise combinations.

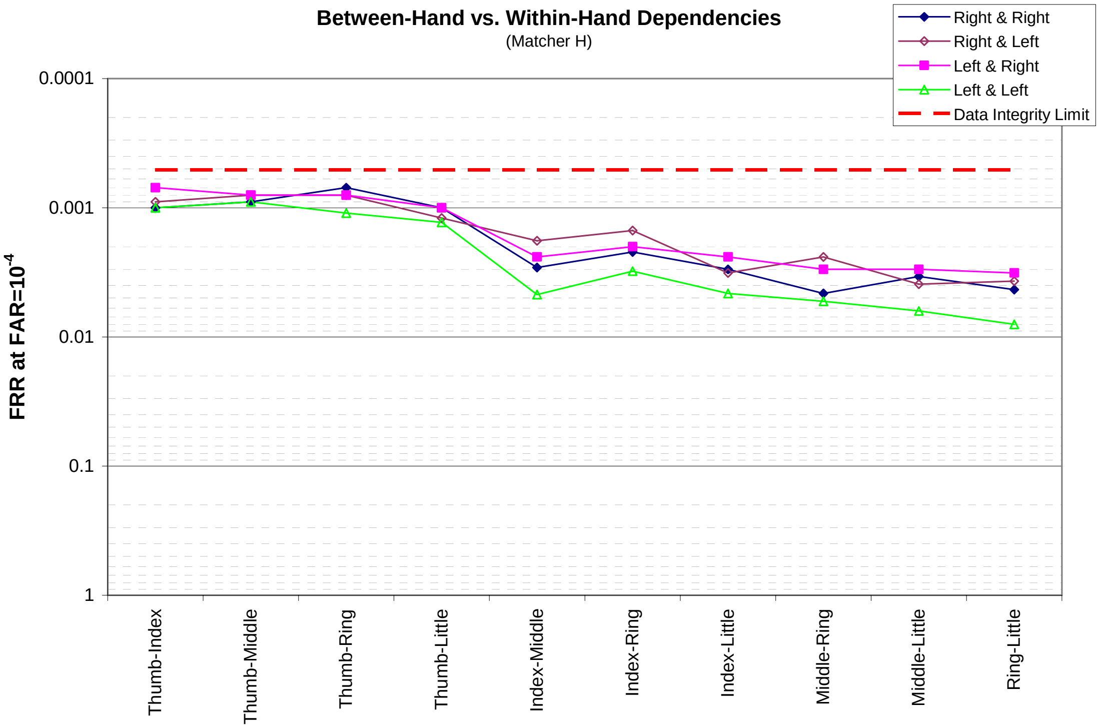

Figure 6: Hand to Hand Correlations: fusing data from opposite hands is beneficial.

[NBDF06 dataset, matcher H]

3.4 N-Way Fusion

3.4.1 Experimental Design

Test subjects have N-fingers, or N-fingers plus face. We use fourteen different combinations of (one or more) fingers and nine different combinations of matchers (3 fingerprint * 3 face). Each trial set contains four trials: one fingers-only trial of actual subjects, one fingers and face trial of actual subjects, one fingers and face chimera trial, and one "reference" trial (face only).

Type of Trial	How score pairs are constructed	Subjects / Trial	Trials / Type
Fingers Only	Combine scores of N different fingers from one individual.	64,867 genuines 122,000 imposters	1
Fingers+Face	Combine scores of N different fingers and face from one individual.	64,867 genuines 122,000 imposters	1
Reference	Face score from one individual.	64,867 genuines 122,000 imposters	1
Total number of trials (per finger combination)			**3**

Table 4: Description of trials, and number of trials per set

3.4.2 N-Finger Results

Figure 7 shows the effect of fusing various combinations of fingers. ***The accuracy of the fused combinations is limited primarily by database errors*** (due to misidentified subjects, or swapped or repeated fingers). The dashed lines indicate "data integrity limits," or the number of ***known*** database errors, which limit achievable performance to a TAR of approximately 0.9995. In the NBDF06 dataset, 33 subjects out of 64,867 (0.051%) were found to have some or all of their fingers misidentified, of whom 24 (0.037%) also have their faces misidentified: these are the two red lines. FRR can pass the 0.051% limit with some finger combinations, but not the 0.037% limit. Some additional data integrity errors may remain undetected, especially if they involve the face but not the fingerprints. The methods used to identify data integrity errors are described in Appendix A.

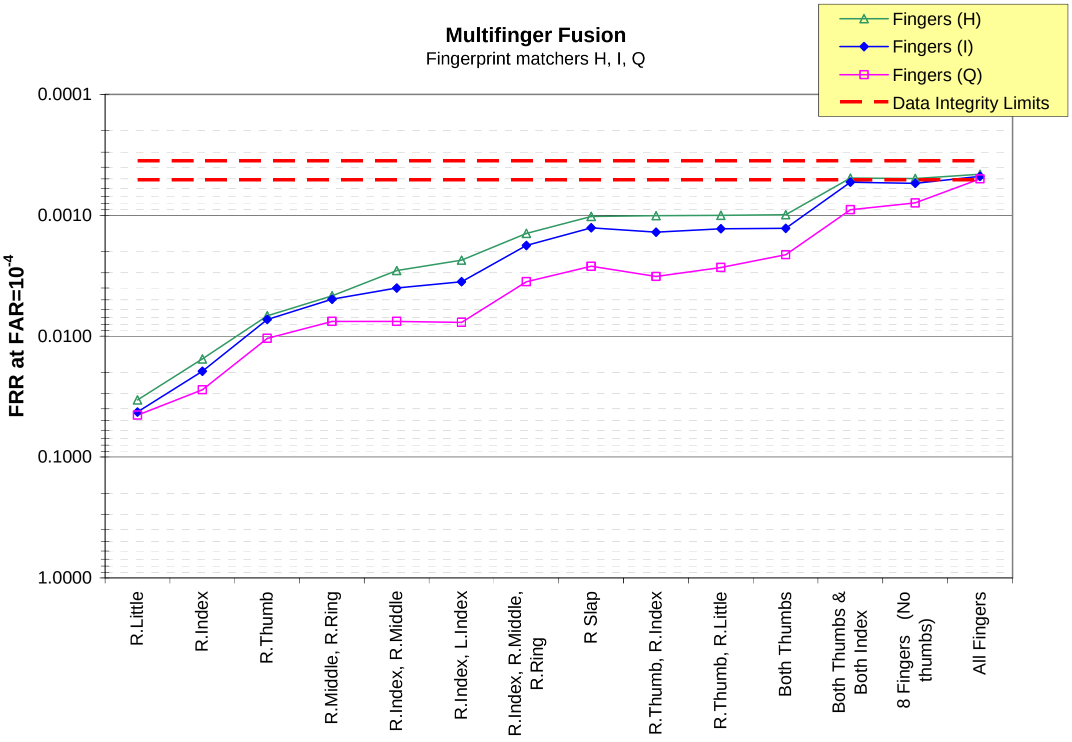

Figure 7: Fusion of multiple fingers, showing data integrity limits (based on the number of *known* database errors) — matchers H, I, & Q

Several comments and observations can be made based on 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 4-finger slap is approximately as effective as a thumb and one 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.

3.4.3 N-Finger + Face Results

Figure 8 and Figure 9 show the results of N-finger and face fusion for two pairs of matchers: the other combinations of matchers generally lie between these two.

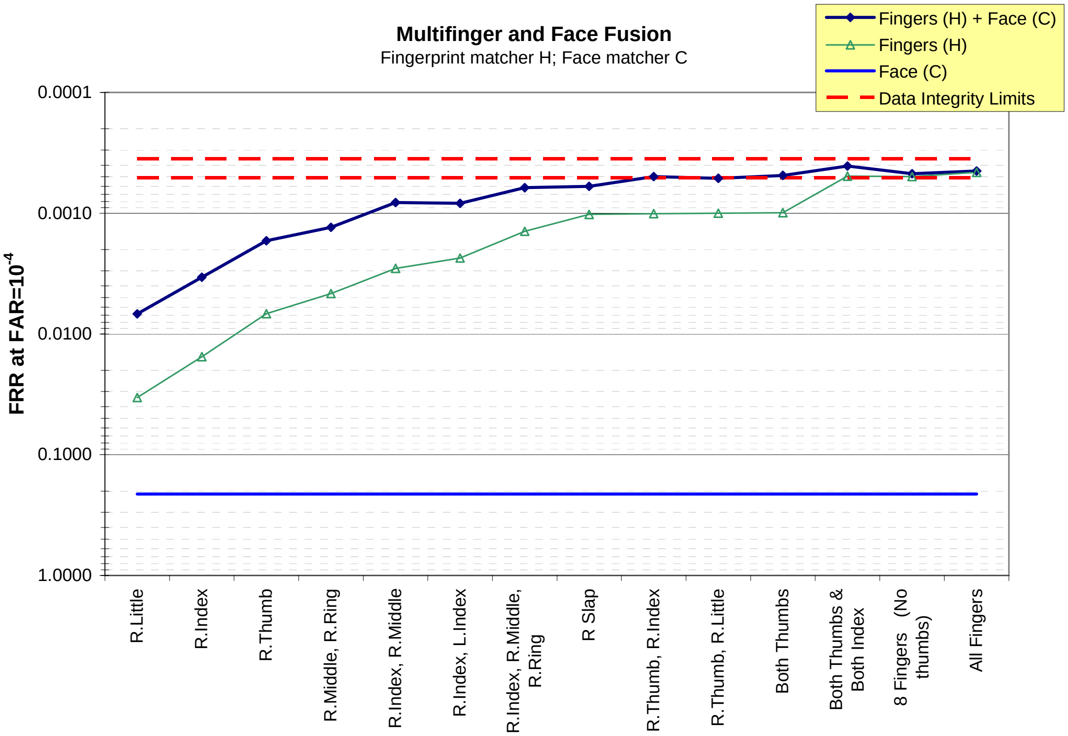

Figure 8: Fusion of multiple fingers and multiple fingers with face, showing data integrity limits (based on the number of *known* database errors) — matchers H & C

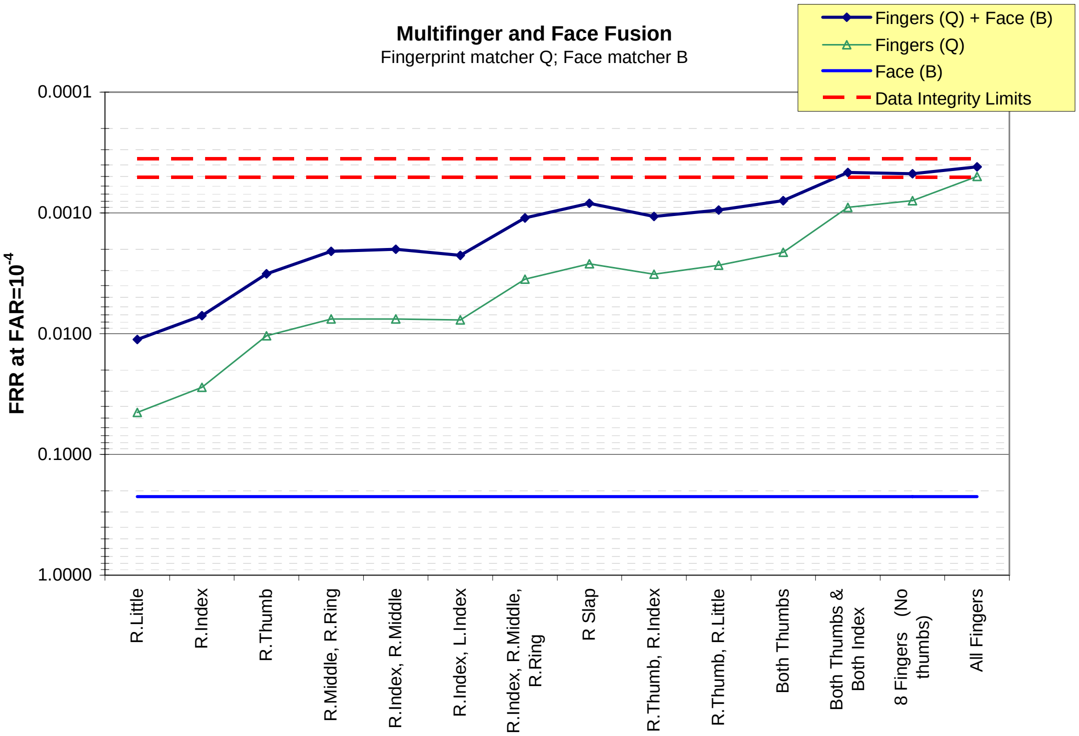

Figure 9 Fusion of multiple fingers and multiple fingers with face, showing data integrity limits (based on the number of *known* database errors) — matchers Q & B

Combining face with the fingers is beneficial in all cases:

- Adding face data to one or two fingers reduces FRR by nearly an order of magnitude.
- Index+Middle+Face (or L.Index+R.Index+Face) is more effective than a 4-finger slap.

3.5 Optimal combinations

3.5.1 Experimental Design

These results are similar to those of the previous subsection, but focus on the gains achieved through different combinations of biometrics. These results include all pairwise combinations of fingers and face, and some 3-way combinations. This analysis is intended to help guide the selection of specific combinations of biometrics.

Type of Trial	**How score pairs are constructed**	**Samples / Trial**	**Trials / Type**
One Face	Face score from one individual.	186,867	1
One Finger	Finger score for each of 10 different fingers from one individual.	186,867	10
Two Fingers	Combine scores of 2 different fingers from one individual.	186,867	45
One Finger+Face	Combine scores of one finger and face from one individual.	186,867	10
Two Fingers+Face	Combine scores of 2 different fingers and face from one individual.	186,867	45
		Total number of trials (per finger combination)	**111**

Table 5: Description of trials, and number of trials per set

3.5.2 Two-finger fusion, and one-finger + face fusion

Table 6 compares the effectiveness of all 45 pairwise combinations of two fingers (matcher H), and all 30 combinations of the three face matchers with one finger (matcher H). The top row and left column show the accuracy before fusion. Note the substantial improvement in every case, especially the lowest FRR values, which are all combinations of a thumb with another finger. Some of these results approach the data integrity limits (FRR≈0.0005 at FAR=10^{-4}), meaning that 2-finger fusion sometimes approaches the maximum measurable accuracy for this dataset.

		RT	RI	RM	RR	RL	LT	LI	LM	LR	LL
		0.0068	0.0155	0.0167	0.0155	0.0337	0.0055	0.0202	0.0241	0.0193	0.0653
R.Thumb	0.0068		0.0010 85%	0.0009 87%	0.0007 89%	0.0010 85%	0.0010 82%	0.0009 87%	0.0008 88%	0.0008 88%	0.0012 82%
R.Index	0.0155			0.0029 81%	0.0022 86%	0.0030 81%	0.0007 87%	0.0023 85%	0.0018 89%	0.0015 90%	0.0032 79%
R.Middle	0.0167				0.0046 70%	0.0034 80%	0.0008 86%	0.0024 86%	0.0037 78%	0.0024 86%	0.0039 76%
R.Ring	0.0155					0.0043 72%	0.0008 86%	0.0020 87%	0.0030 81%	0.0025 84%	0.0037 76%
R.Little	0.0337						0.0010 82%	0.0024 88%	0.0030 87%	0.0032 83%	0.0101 70%
L.Thumb	0.0055							0.0010 82%	0.0009 83%	0.0011 81%	0.0013 76%
L.Index	0.0202								0.0047 77%	0.0031 84%	0.0046 77%
L.Middle	0.0241									0.0053 72%	0.0063 74%
L.Ring	0.0193										0.0080 59%
L.Little	0.0653										
Face A	0.2800	0.0019 72%	0.0044 72%	0.0050 70%	0.0039 75%	0.0086 75%	0.0015 72%	0.0056 72%	0.0068 72%	0.0053 72%	0.0189 71%
Face B	0.2237	0.0021 68%	0.0040 74%	0.0046 72%	0.0039 75%	0.0076 77%	0.0016 71%	0.0060 70%	0.0060 75%	0.0052 73%	0.0170 74%
Face C	0.2119	0.0017 75%	0.0034 78%	0.0037 78%	0.0033 79%	0.0068 80%	0.0016 71%	0.0049 76%	0.0054 77%	0.0040 79%	0.0146 78%

0.0010 — FRR at FAR = 10^{-4}
85% — Reduction in FRR

Best values
Medium values
Worst values

Table 6: Comparison of FRR at FAR=10^{-4} for all 2-finger combinations, and combinations of all single fingers with the three face matchers. Percentage reduction in FRR is relative to the stronger input alone.[3] The lowest FRR values are in blue; the highest in red. The greatest FRR reductions are in blue; the least in red.

[NBDF06 dataset; matcher H and all face matchers, product of likelihood ratios fusion]

Note:

- Pairwise fusion always improves FRR substantially, with typical gains of 70-90%.
- 2-finger fusion usually outperforms finger+face, unless the left little finger is involved.
- The improvement in FRR shown for finger+face often approximates the TAR for the face matchers involved (A: 0.72; B: 0.78; C: 0.79). This is a result of the independence of the data.
- FRR improves most for left-right and thumb-slap combinations, and least for neighboring fingers and left-right little fingers.

These same results are summarized for all matchers in Table 7. Note that the comparative performance of two-finger vs. face+finger reflects the choice of matchers: matcher H yields substantially more accurate results on two fingers than finger+face; fingerprint matcher Q is slightly less accurate on two fingers than finger+face.

[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%

	Two fingers			Single finger + face		
	H fingers	I fingers	Q fingers	H+face	I+face	Q+face
Min	59%	48%	51%	68%	71%	64%
Median	83%	79%	72%	74%	76%	75%
Average	82%	78%	71%	74%	77%	74%
Max	90%	90%	84%	80%	84%	79%

Table 7: Reduction in FRR where FAR = 10^{-4}, relative to the stronger of the inputs. Two fingers summarizes all 45 pairwise combinations; Single finger + face summarizes fusion of each of the ten fingers against each of the three face matchers.

[NBDF06 dataset; all matchers, product of likelihood ratios fusion]

3.5.3 Two-finger + face fusion

Table 8 extends these results to 3-way fusion, showing all combinations of two fingers with face, for matchers H and C. Note that FRR continues to reduce substantially, but the benefits have begun to diminish. This effect is due at least in large part to data integrity limits. Table 8 also shows presents these same results as a percentage reduction in FRR when adding face to 2-finger fusion.[4]

		RT	RI	RM	RR	RL	LT	LI	LM	LR	LL
		0.0068	0.0155	0.0167	0.0155	0.0337	0.0055	0.0202	0.0241	0.0193	0.0653
R.Thumb	0.0068	0.0038	**0.0005**	**0.0004**	**0.0004**	**0.0005**	**0.0005**	**0.0004**	**0.0004**	**0.0005**	0.0006
		44%	44%	51%	50%	44%	49%	51%	50%	47%	42%
R.Index	0.0155	**0.0005**	0.0072	0.0008	0.0007	0.0009	**0.0004**	0.0008	0.0007	0.0006	0.0009
		51%	53%	**53%**	**72%**	**70%**	**71%**	40%	65%	61%	63%
R.Middle	0.0167	**0.0004**	0.0008	0.0102	0.0013	0.0009	**0.0005**	0.0009	0.0013	0.0007	0.0010
		50%	**72%**	**39%**	39%	**72%**	**73%**	40%	64%	66%	**70%**
R.Ring	0.0155	**0.0004**	0.0007	0.0013	0.0090	0.0011	**0.0004**	0.0008	0.0011	0.0009	0.0009
		44%	**70%**	**72%**	42%	42%	**74%**	43%	59%	64%	66%
R.Little	0.0337	**0.0005**	0.0009	0.0009	0.0011	0.0133	**0.0005**	0.0009	0.0008	0.0010	0.0024
		49%	**71%**	**73%**	**74%**	60%	60%	48%	62%	**72%**	69%
L.Thumb	0.0055	**0.0005**	**0.0004**	**0.0005**	**0.0004**	**0.0005**	0.0028	**0.0005**	0.0006	0.0006	0.0006
		51%	40%	40%	43%	48%	49%	49%	45%	42%	48%
L.Index	0.0202	**0.0004**	0.0008	0.0009	0.0008	0.0009	**0.0005**	0.0098	0.0016	0.0010	0.0013
		50%	65%	64%	59%	62%	45%	52%	52%	66%	67%
L.Middle	0.0241	**0.0004**	0.0007	0.0013	0.0011	0.0008	0.0006	0.0016	0.0138	0.0014	0.0017
		47%	61%	66%	64%	**72%**	42%	66%	43%	43%	**74%**
L.Ring	0.0193	**0.0005**	0.0006	0.0007	0.0009	0.0010	0.0006	0.0010	0.0014	0.0098	0.0018
		42%	63%	**70%**	66%	69%	48%	67%	**74%**	49%	49%
L.Little	0.0653	0.0006	0.0009	0.0010	0.0009	0.0024	0.0006	0.0013	0.0017	0.0018	0.0288
		49%	**72%**	**76%**	**76%**	**77%**	56%	**71%**	**73%**	**77%**	56%

Table 8: FRR at FAR=10^{-4} for fusion of all combinations of two fingers with face, and reduction in FRR (due to the contribution of face relative to two fingers only). The diagonal shows fusion of each single finger with face as a baseline. FRR values that reach the data integrity limit are in blue. The greatest FRR reductions are in blue; the least in red.

[NBDF06 dataset; matchers H and C, product of likelihood ratios fusion]

[4] For example, when adding face to Left Thumb and Right Index, the 2-finger FRR_{LT*RI}=0.000737 (from Table 6, but with increased precision) and the face+2-finger $FRR_{LT*RI*Face}$=0.000444 (from Table 8, increased precision), the improvement in FRR = (0.000737 – 0.000444)/ 0.000737 = 40%0.000740.0007 43

4 Exploratory Study of Face and Fingerprint Fusion using BSSR1 Data

Prior to analyzing NBDF06, an exploratory study was conducted using data from the NIST Biometric Scores Set, Release 1 [BSSR1], which is described in Appendix A. It is large enough for an exploratory study, and has the advantage of being in the public domain, permitting comparisons with other studies.

Figure 10 illustrates scatterplots of the genuine and imposter joint distributions of the Face(G) and R-index(V) scores. The striking separation is characteristic of all four sets. In these scatterplots, only a random subsample of the 133,386 imposter scores is shown.

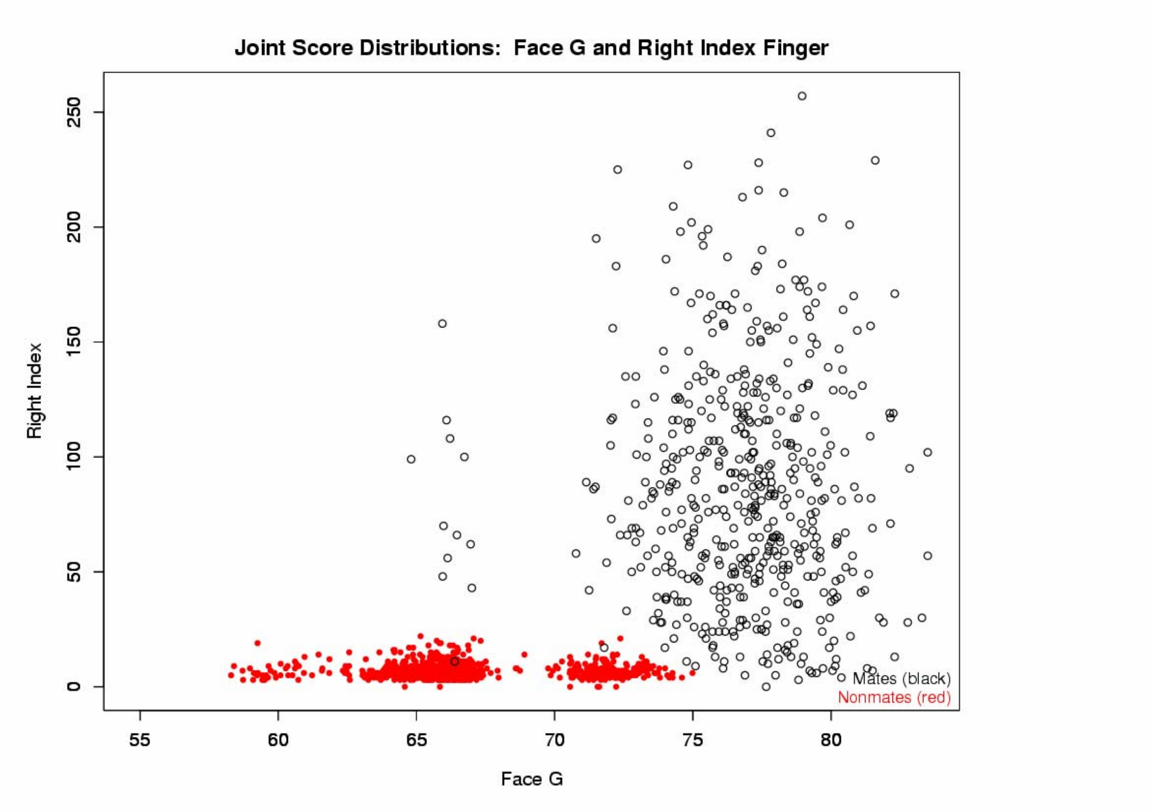

Figure 10: Joint distributions of genuine (black) and imposter (red) Face(G) and R-index(V) scores
[BSSR1 dataset; matchers G and V]

4.1 Multimodal Dependencies: Face vs. Finger

This experiment investigates the independence of two biometric modalities: faces and fingerprints.

4.1.1 Experimental Design

To test the hypothesis that faces and fingerprints are independent, we compare the performance of biometric fusion on chimeras constructed to have independent face and fingerprint data with performance on actual subjects. If the ROC curve for the chimeras is not noticeably different from the ROC curve for real individuals, then we have evidence that face and finger scores have little co-dependence.

The experiment consists of four sets of thirteen trials. Each trial involves computing the ROC curve for a fusion run on a test set of 517 "subjects", each of whom has one finger and one face. Thus the input to the fusion algorithm is a pair of matcher scores, one for the face and one for the finger. Since there are two kinds of fingers (right and left index fingers), one fingerprint matcher $_{(V)}$, and two face matchers (C, G), it is possible to construct four kinds of score pairs: R-index$_{(V)}$ ⊗ Face$_{(C)}$, R-index$_{(V)}$ ⊗ Face$_{(G)}$, L-index$_{(V)}$ ⊗ Face$_{(C)}$, L-index$_{(V)}$ ⊗ Face$_{(G)}$. Therefore we conduct four sets of trials, one set for each kind of score pair.

Each trial set consists of 13 trials (see Table 9). Given two sources of individual fingerprint scores (i.e. BSSR1 Set 1 and BSSR1 Set 2), we construct two kinds of chimeras for a total of three types of trials. Because there is multimodal data for only 517 individuals, the number of subjects in each trial is limited to 517, and there can only be one trial for real individuals. However, the number of potential chimeras is quite large,[5] so to better distinguish real effects from sample variance, six trials are run for each type of chimera, with each of the six trials employing a different set of chimeras.

In every trial, the ROC is calculated from a 517 x 517 similarity matrix of fusion scores. Face and fingerprint scores are fused using a weighted linear combination. Two weights are used, one for combining fingers with Face$_{(C)}$ data (266*Face + Finger), and another for Face$_{(G)}$ data (5*Face + Finger). These weights are selected to optimize TAR at FAR = 10^{-4}.

Type of Trial	How score pairs are constructed	Samples / Trial	Trials / Type
Actual subject	Pair the face and finger scores of a single individual from BSSR1 Set 1	517x517	1
Chimera 1	Pair the face and finger scores of two distinct individuals from BSSR1 Set 1.	517x517	6
Chimera 2	Pair the face score of an individual from BSSR1 Set 1 with the fingerprint score of an individual from BSSR1 Set 2.	517x517	6
Total number of trials (per type of input pair)			**13**

Table 9: Description of trials, and number of trials per set

4.1.2 Results

The results for this experiment are captured in four plots (one per set of trials). Each plot contains 13 fusion ROC curves, plus two benchmark (face and finger) ROC curves. Figure 11 shows results for R-index$_{(V)}$ ⊗ Face$_{(C)}$; the other three plots are very similar. As expected, the fused ROCs substantially outperform the ROCs of the individual face and fingerprint matchers. Once again, no significant difference in the results of the real and artificial subject sets confirms a negligible correlation between face and fingerprint data.

The validity of using chimeras may depend on the datasets involved, however, as interpretation of the charts does appear to vary by dataset. One possible explanation is that if a dataset consists of two subpopulations, each with distinct score distributions for fingers and faces, then that source of dependence would be removed through the use of chimeras.

Several caveats should be noted:

- These results are based on small sample sizes.
- These results are based on only one subject population
- These results are based on matchers that are not highly accurate: NIST VTB is far less accurate than the top performing fingerprint matchers [FpVTE]; the face matchers date from c. 2002.

[5] The total number of chimera pairs that can be built from BSSR1 Set 1 is 133,386. Our 6 sample sets used only 12,408 of these.

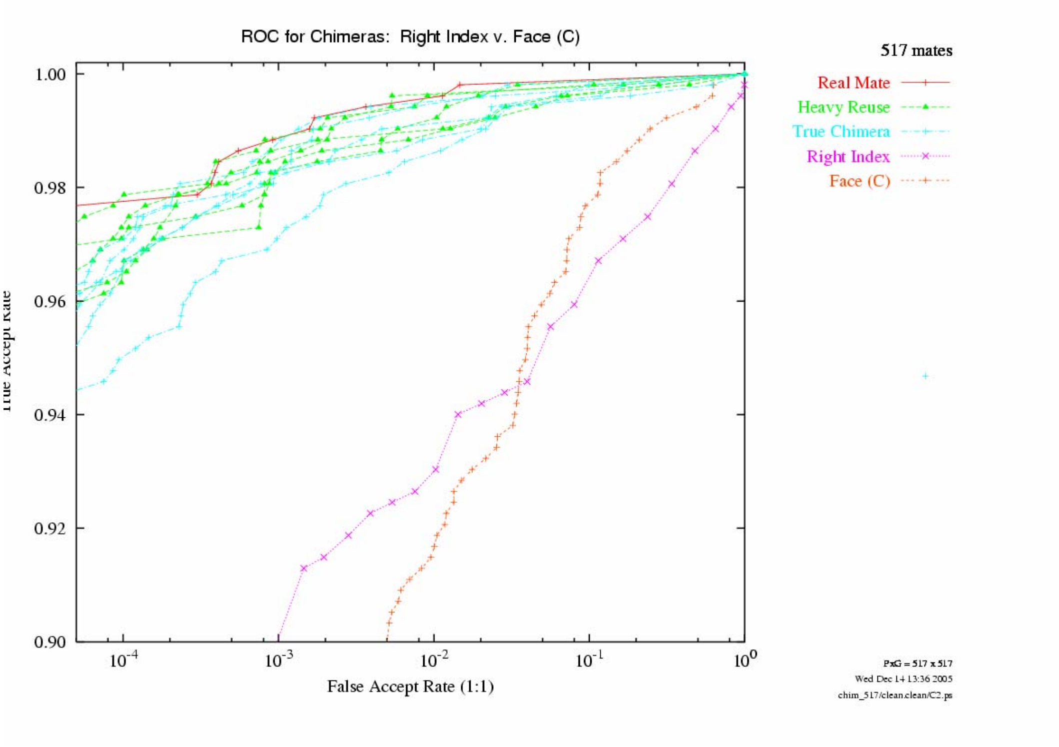

Figure 11: Fusion performance for Right Index fingers(VTB) and Face(C).

[BSSR1 dataset; matchers C and V]

4.2 Between-Hand Dependency: Finger vs. Finger

This experiment investigates the dependencies between the index fingers of each hand.

4.2.1 Experimental Design

For this experiment, we define a test "subject" as having one left index finger and one right index finger. With just one matcher available (fingerprint matcher V), only one kind of score pair is possible, and we run just one trial set.

In order to distinguish real effects from sample variance, multiple ROCs are generated by partitioning the available data. In every case, matrices of size 1000 x 1000 are used, and each probe image has one mated image in the gallery. The 6 trials of actual subjects were created by partitioning the original 6000 subjects into 6 equal-sized sets of 1000 subjects. The 3 trials of Chimeras were created by taking left fingers from one subset, right fingers from another (i.e., from 3 pairs of 1000 subjects).

Type of Trial	How score pairs are constructed	Samples / Trial	Trials / Type
Actual subject	Pair the left index finger and right index finger scores of a single individual from BSSR1 Set 2.	1000x1000	6
Chimera	Pair the left index finger and right index finger scores of distinct individuals from BSSR1 Set 2.	1000x1000	3
		Total number of trials	**9**

Table 10: Number of trials per type

For this experiment, the fusion algorithm is a simple sum of the raw scores. This is effective because the score distributions for the two fingers are quite similar.

4.2.2 Results

The results of this investigation (Figure 12) demonstrate that an independence assumption is overly optimistic: the two-finger chimeras result in substantially higher accuracy than those for actual subjects, indicating significant dependence between right and left index finger scores. This result is not surprising given several known sources of dependence, as discussed in Section 2.1. The improvement in FRR is approximately 50%, which is much less than was measured for the same fingers on the NBDF06 dataset. This difference is explained by differences in matcher accuracy.

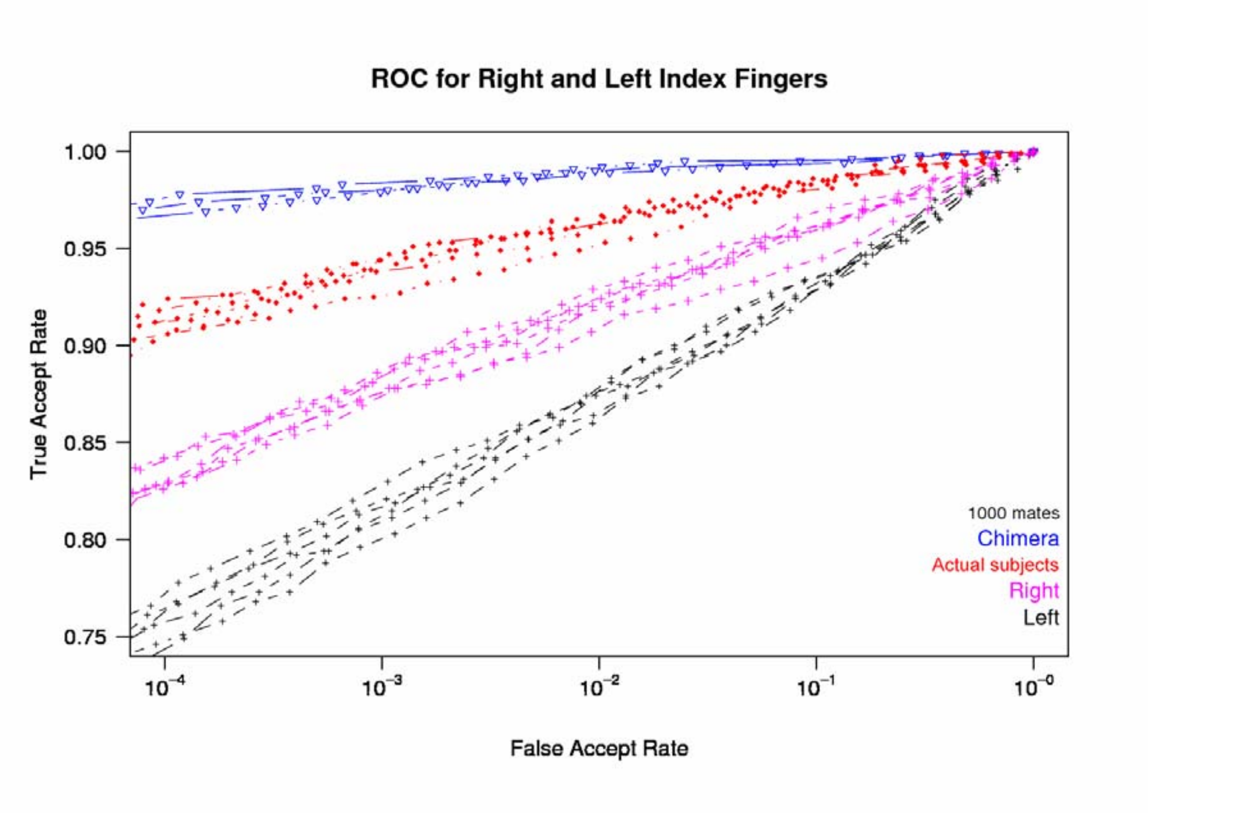

Figure 12: Comparison of accuracy between single Left and Right fingers, 2-finger fusion for actual subjects, and 2-finger fusion for Chimeras. The Chimera results show what fusion performance would have been if the fused fingers were fully independent.

[BSSR1 dataset; matcher V, Best Linear fusion]

5 Discussion & Conclusions

Fusion of fingers is highly effective, as is fusion of fingers and face. Fusing two fingerprints or one fingerprint and face generally resulted in a 50-90% reduction in false reject rate (FRR) relative to the stronger of the two inputs at a constant false accept rate (FAR).

Much of the effectiveness of fusion depends upon the extent to which different biometric modalities and instances are independent. In this study, chimeras were used, not of necessity due to lack of data, but as a tool for directly evaluating the effects of correlated data, and this work shows that they are extremely useful for that purpose.

When estimating the benefits to finger fusion, an independence assumption is overly optimistic: two-finger chimeras result in substantially higher accuracy than those for actual subjects. For the fusion of one fingerprint with face, however, there is little difference in accuracy between actual subjects and chimeras.

For this reason, fusing the face with fingers offers a distinct advantage. The choice of fusing two fingers vs. one finger and face depends on the choice of matchers. Using two fingers and face was more effective than a 4-finger slap. As the accuracy of matchers continues to improve, there will be an on-going need to reassess the benefits of specific combinations.

It is difficult to generalize the results of finger fusion because accuracy varies substantially by finger position, and correlations among scores vary greatly by finger positions. The much larger thumbprints are substantially more effective than the other fingers. The relative advantage of thumbs over index fingers is comparable to the relative advantage of index fingers over little fingers. In all cases, score correlations substantially limit the benefits to finger fusion. The strongest correlations occur among neighboring fingers within a slap, but all combinations show some correlation. The little finger was often surprisingly effective in fusion, possibly because quality problems such as incorrect placement affected the little fingers differently from the other fingers.

As more inputs are fused and accuracy approaches 100%, the maximum achievable accuracy is limited by data integrity problems (misidentifications, swapped prints, missing images). This limitation is most pronounced when working with the most accurate matchers, measuring FRR at a high FAR, and/or combining several scores. Data integrity had a substantial limiting effect on FRR (at FAR=10^{-4}) when combining as few as three or four scores.

Appendix B.2: Score-Level Fusion of Multiple Matchers

Contents

6 Introduction: Combining Multiple Matchers

This is a report of the effectiveness of algorithm fusion using face and fingerprint data from the NBDF06 dataset, as well as fingerprint data from the FpVTE MST dataset. These datasets are described in Appendix A.

The categories of biometric fusion include the use of multiple types of biometric data or multiple methods of processing. Fusion based on multiple types of data improves accuracy due to an increased amount of relatively independent data, but at the cost and complexity of collecting more data.

Algorithm fusion is a generic term for the combination of multiple methods of processing for each individual sample. Here the processes being fused are feature extraction and matching:

- Feature extractors generate templates for each sample, which are then fused at the template level. This is generally only appropriate with compatible templates.
- Matchers produce scores or decisions. In many cases (such as in this study) the feature extractor is an integral part of the "matchers" being evaluated.

Using multiple methods of processing on the same samples is generally expected to be less effective than either instance or mode fusion, but can be effective if the results from the processing algorithms are relatively independent. Matcher fusion has the advantage of maximizing the benefit of the data available. It should be noted that some feature extractors and matchers internally use multiple algorithms, unknown to the user or evaluator; the users of a commercially available matcher may already be the beneficiaries of matcher fusion.

The use of multiple algorithms to process common data is a classic problem in the pattern recognition and machine learning literature.[6] The effectiveness for biometric fusion depends on the degree of independence brought by the different algorithms. This hinges on two factors:

- ***Different types of algorithms***. "Maximum benefit (theoretically) would be derived from algorithms that are based on distinctly different and independent principles (such algorithms may be called 'orthogonal')." [SC37-24722, p. 6] Ideally, the matchers being fused would use

[6] See [Jain-99c] for a summary of references.

fundamentally different types of features, different methods of feature extraction, and different matcher algorithms.

- ***Complementary errors***. If two feature extractors or matchers duplicate each other's mistakes (or one is a subset of the other), no substantial gain can be expected from fusion. If errors differ due to limitations of specific implementations, the underlying algorithms, or data-specific characteristics, fusion can help fill in those differences.

This study seeks to empirically quantify the effect of matcher fusion, identify potential sources of complementary technology, and assess whether there exists significant potential for the most accurate matchers to be further improved by fusion with other matchers.

7 Previous Work

Score-level matcher fusion of correlated data has been discussed in several papers:

- Jain investigated matcher fusion using logistic regression [Jain-99c]. We found this technique to be highly effective (see Appendix C). They demonstrate a benefit to fusing algorithms, particularly with one pair of fingerprint matchers. As discussed in Appendix E, the difficulty lies in estimating the density distributions from sample data.
- [Grother-04], Section 6.4 investigated matcher fusion (termed "Multi-System Fusion"). Scores were normalized using the empirical cumulative distribution function of the genuine scores[7], then z-normed. Fusion was implemented as a simple sum of the normalized scores, or as a weighted sum, where the weight was selected to maximize the area under the curve (AUC)[8]. Grother observes that the results of maximizing AUC "are only moderately better;" however, note some matcher combinations are much better (e.g., fusing Visionsphere and Eyematic went from an averaging behavior to real benefits).
- [Fierrez-06] investigated the fusion of a minutiae-based matcher with a ridge-based matcher using an adaptive scheme based on an image quality metric. Scores were normalized using parametric transformations (tanh and exp), then fused by one of two methods: simple sum[9] and weighted sum. The weighting was based on the quality score, such that the weaker minutia-based matcher was progressively downweighted as image quality dropped[10]. On low quality images, a simple sum had an averaging effect, yielding performance midway between the two matchers; the adaptive method performed similarly to the ridge-based method.[11] On high quality images, fusion was clearly effective; and both techniques produced similar results.

[7] For comparison, the Product of FARs technique evaluated in this study normalizes based on a smooth model of the imposter cdf. Three major differences between these implementations included: the distribution, smoothing, and z-norm.

[8] This study did not investigate AUC optimization. However, note that computational methods of optimizing AUC are approximate. When evaluating TAR at FAR=10^{-4} (the standard point of comparison in this study), special care may be required to ensure that this region (0.01% of the range of the ROC) was in fact optimized. Notice, for instance, in Figure 20 of [Grother-02] how the ROC shows markedly greater downward curvature than those achieved in this study.

[9] Actually an average of the two scores, which is effectively not different from sum.

[10] The adaptive quality weighting scheme is defined such that the score combination ranges from equal weighting of matcher results to using only the (more accurate) ridge-based matcher results. It is not explained why greater weight is never given to the minutia-based matcher.

[11] These used small datasets, and no confidence intervals were shown.

- [Fierrez-05b] evaluated the Simple Sum of raw scores, SVM, and Dempster-Shafer methods. They found that none of these methods consistently outperformed the others. This may reflect assumptions made by these techniques that are not always valid: for example, although they cite Kittler for Simple Sum, they do not first ensure that scores are normalized as specified in that paper. They observed "the combination of the top performing individual systems can be outperformed by other combinations." While it is entirely possible for pairs other than the top two systems to exhibit greater independence, and hence offer greater potential for fusion, the results in [Fierrez-05b] more likely reflect a failure to achieve good fusion of the top two systems. The statistical significance of the relative rankings is not discussed, which is notable, given the small datasets, non-random sampling and the very large number of combinations tested: they combined up to 41 matchers on a dataset constructed from 100 subjects, 8 fingerprint samples per subject.

8 Analysis Methods

In matcher fusion, scores from pairs of matchers are highly correlated because the two matcher scores both derive from comparisons of the same pair of images. Thus, as shown in Figure 13, the usual method of estimating the joint density distributions as the product of univariate densities is clearly inaccurate (the genuine scores are clearly correlated); the independence assumption is not valid. However, detail of the low score range of the same graph (Figure 14) shows that in the critical region of the decision boundaries, not only are the scores sufficiently independent to warrant fusion, but also that this simplified method of joint density estimation might yield good results.[12]

[12] In these two figures, only a random subsample of imposters (shown in red) is plotted.

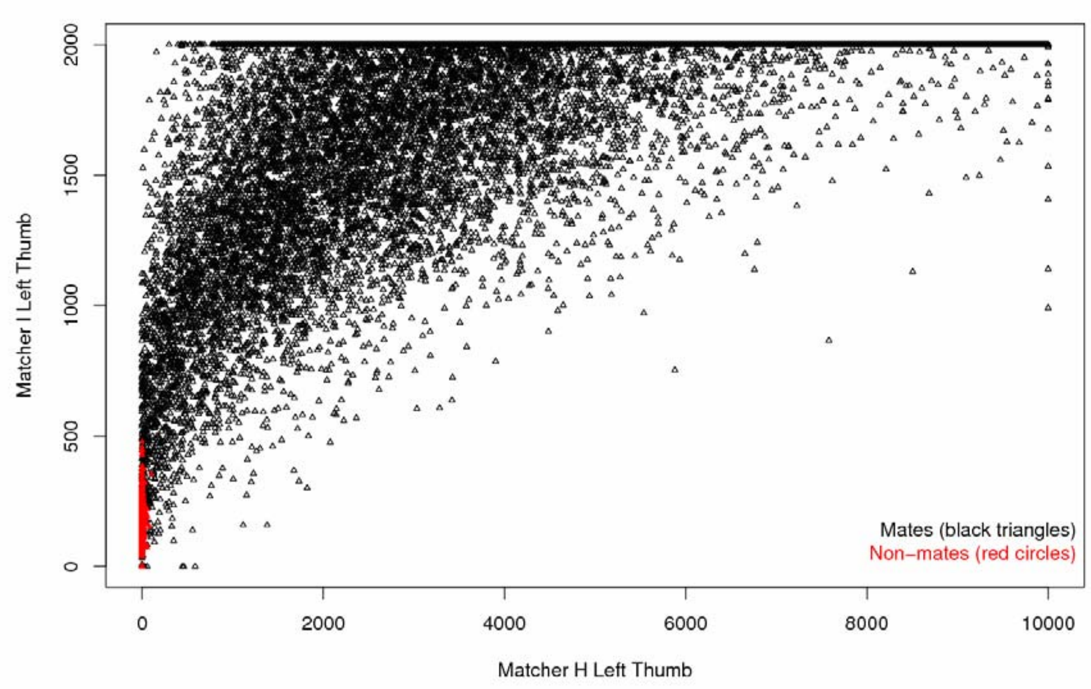

Figure 13. Scatterplot for NBDF06 dataset, left thumb scores, matchers H and I. Genuine scores are shown in black; imposters in red. Highly dependent data would result in results clustered around a line or curve; independent data would be broadly scattered.

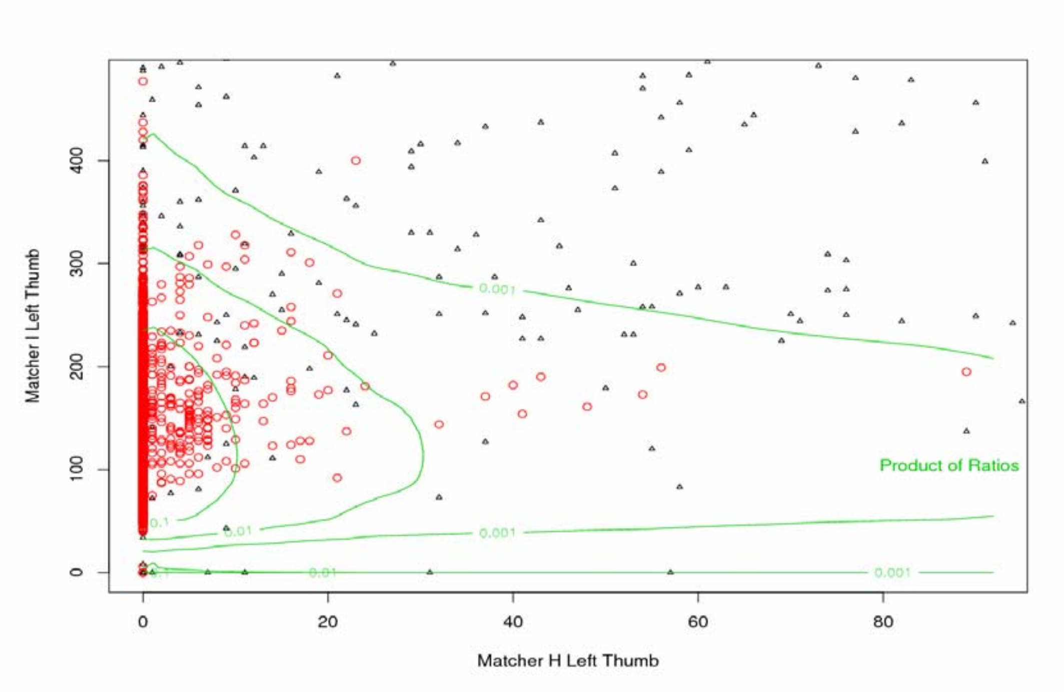

Figure 14. Detail of scatterplot from Figure 13 (NBDF06 dataset, left thumb scores, matchers H and I), with decision boundaries. Genuine scores are shown in black; imposters in red; Product of Likelihood Ratios decision boundaries in green. Note that only a random subsample of imposters (shown in red) is plotted.

A limited analysis was conducted to compare the efficacy of fusion techniques on dependent data (techniques are described in Appendix C). Figure 15 shows, for an example pairing of matchers from FpVTE MST, the ROCs obtained by three different fusion techniques. Note that despite the simplified density estimation (independence assumption), the relative performance of these techniques is quite similar to that achieved on highly independent data. That is, the three techniques perform similarly to one another, and the Product of Likelihood Ratios is slightly superior to the other two. This was found to be typical in several pairwise comparisons of FpVTE matchers, but was not systematically confirmed in the FpVTE data.

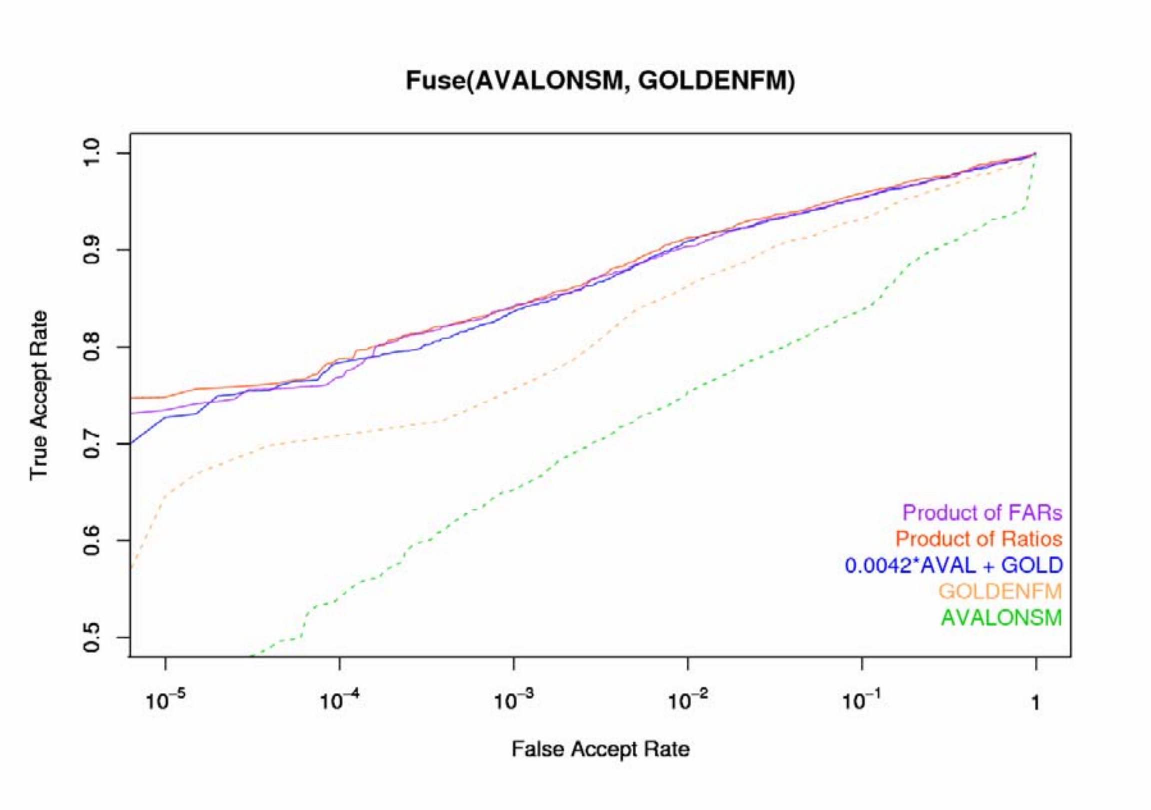

Figure 15: Comparison of three fusion techniques (combining Avalon and Golden Finger[13], FpVTE MST dataset)

The results in this multi-algorithm study used two methods of fusion. For the NBDF06 analysis (Section 9), the Product of Likelihood Ratios method was used. Because that approach requires painstaking curve fitting of both genuine and imposter score distributions for every matcher, it was not a practical option for this analysis of the fourteen FpVTE MST matchers. For the FpVTE analysis (Section 10), matcher scores were fused using the (fully automated) "best linear" combination. As discussed in Appendix C, linear combinations generally achieve very good fusion results but are not optimal. "Best linear" also makes an interesting choice for matcher fusion because it does not assume independence.

[13] Specific hardware and software products identified in this report do not imply recommendation or endorsement by the National Institute of Standards and Technology.

9 NBDF06 Results

Figure 14 (above) shows that there is a potential benefit to fusing multiple matcher scores even when using highly accurate fingerprint matchers. The following subsection presents ROCs that demonstrate this benefit on various combinations of matchers. Section 9.2 provides additional data on the joint score distributions to help explain these results. Matcher fusion was performed using the Product of Likelihood Ratios technique.

9.1 Results

Matcher fusion, using the Product of Likelihood Ratios technique, can reduce FRR substantially. Table 11 shows that for pairs of fingerprint matchers, FRR was generally reduced by about 10-30% (varies by finger position); for pairs of face matchers, FRR was reduced by 10-13%.[14]

	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 11: Reduction in FRR where FAR = 10^{-4}, for pair-wise matcher fusion. Fingerprint results are for all ten finger positions.

(All fusion was Product of Likelihood Ratios, using the NBDF06 dataset)

The following three charts (Figure 16 – Figure 18) show the effect of three-way matcher fusion.

[14] The percent improvement in FRR compares the number genuines missed when algorithms are fused to the number of genuines missed by the stronger algorithm alone. For example, for two matchers X and Y, if TAR_X=.9, TAR_Y=.95 and TAR_{X+Y}=.96, then the improvement in FRR is 100*(.05-.04)/.05 = 20%.

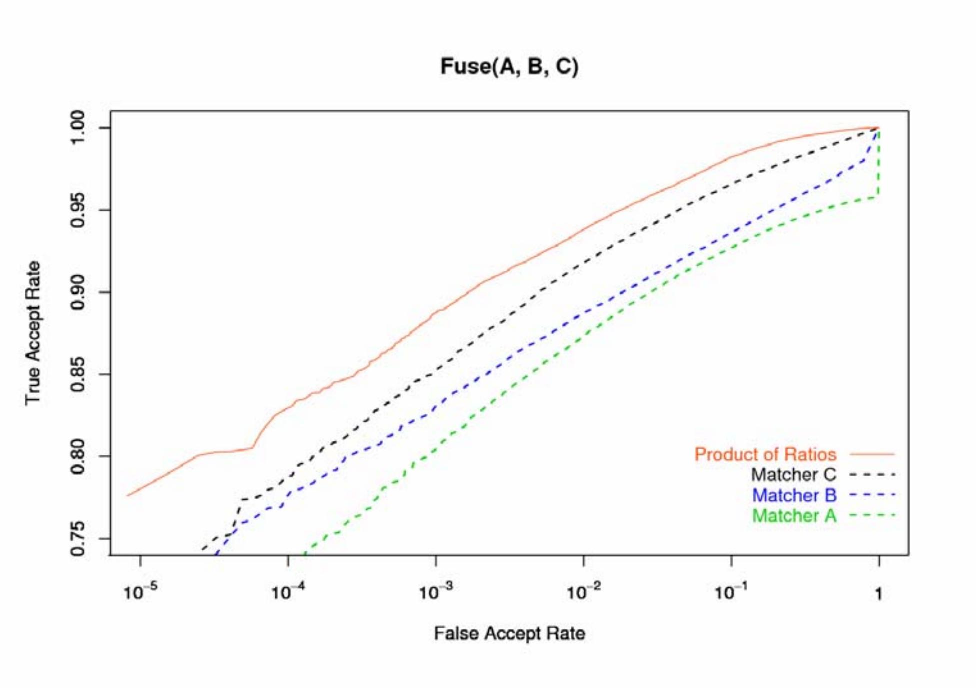

Figure 16: Fusing all three face matchers reduced FRR by 20% relative to Matcher C (at FAR=10^{-4}). Note that much of the benefit at high FAR (between 1 and 0.01) is directly due to Density Ratio normalization of Matcher A, not fusion.

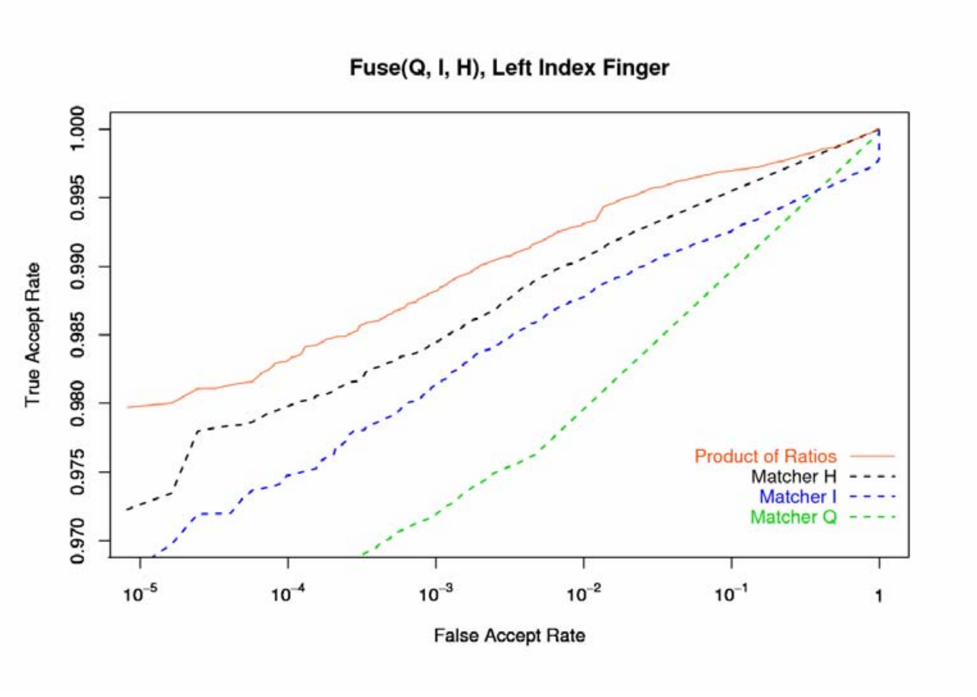

Figure 17: Fusing all three matchers on left index fingers reduced FRR by 17% relative to Matcher H (at FAR=10^{-4}).

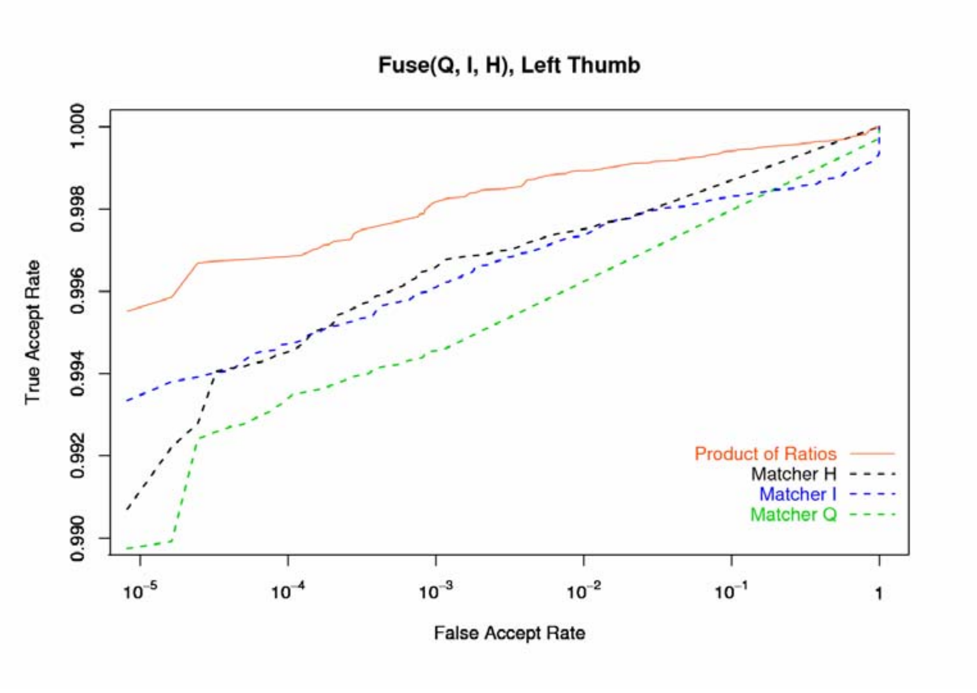

Figure 18. Fusing all three matchers on left thumbs reduced FRR by 40% relative to Matcher I (at FAR=10^{-4}).

9.2 Discussion of Data Dependence

Since single-instance multi-matcher fusion uses the same data for both (or all) algorithms, the independent information that is necessary for effective fusion must come from differences (if any) between algorithms. The H and I fingerprint matchers have been shown to be among the most accurate fingerprint matchers currently available [SDK], a review of the extent of independence of these algorithms is of note, because there is a concern that as accuracy increases, the orthogonality between algorithms may lessen.

This section provides additional information about the dependence of Matcher H and I scores on NBDF06 fingerprint data (as shown in Figure 13 and Figure 14 in Section 8) in order to help explain the results presented in Section 9.1.

Measures of Dependence

While Figure 13 shows that the overall joint score distribution of H and I has a high degree of dependence, Figure 14 shows substantial independence in the area of genuine and imposter overlap. The following charts show details of the degree of dependence between H and I. Figure 19 shows that the imposter score distribution for Matcher I is not highly dependent on Matcher H scores[15]. Figure 20 shows substantial dependence among the genuine scores, yet there is still considerable spread in Matcher I scores when Matcher H reports a score of zero (equivalent to a FAR of 1).

[15] Matcher I score distribution for those cases where matcher H score is zero cannot be discerned from the scatterplot; Matcher I score distribution for those cases where matcher H score is between FAR=10^{-4} and 10^{-2} is intermediate (roughly centered) between the two shown.

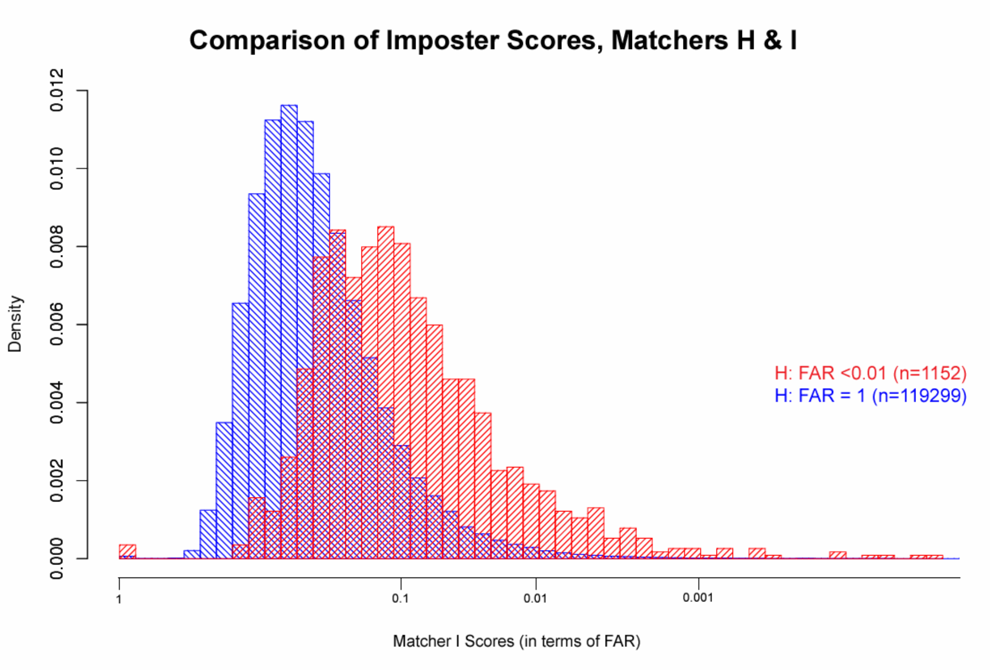

Figure 19: Matcher I imposter score distribution of groups selected based on scores from Matcher H.
Blue: H FAR=1 (minimum value); Red: H FAR<=10^{-2} *(Left thumbs; NBDF06 dataset)*

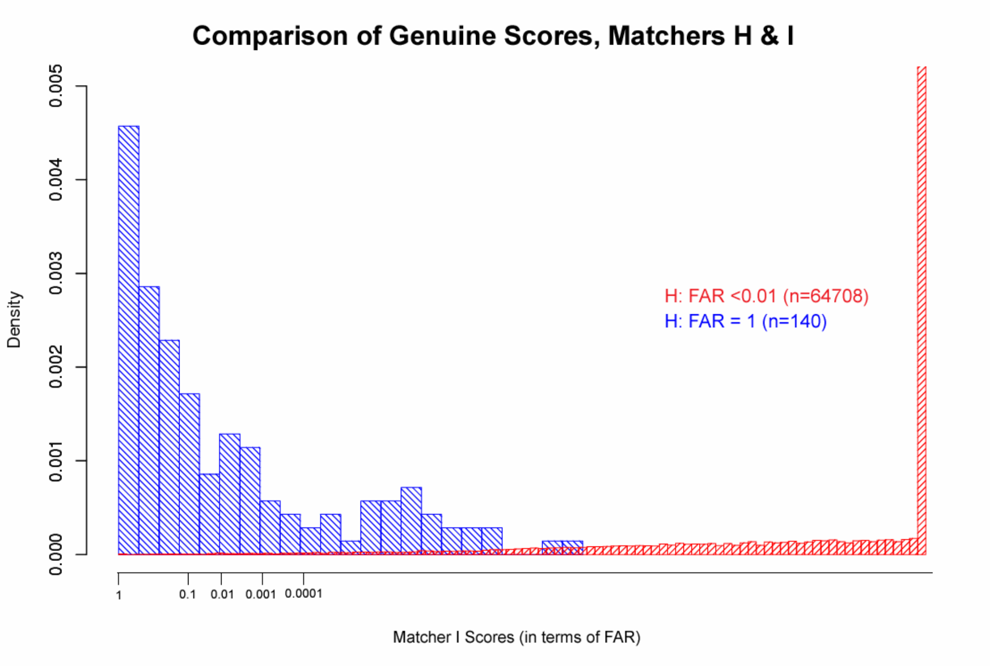

Figure 20: Matcher I genuine score distribution of groups selected based on scores from Matcher H.
Blue: H FAR=1 (minimum value); Red: H FAR<=10^{-2} *(Left thumbs; NBDF06 dataset)*

Table 12 summarizes the joint densities. As seen in Figure 14, these data show that much of the score dependence occurs among the high scoring genuines (definitive matches); lower scores, in the region of difficult discrimination, are not as strongly correlated.

Imposter (n=122,000)		I FAR				
		1 to 10^{-2}	10^{-2} to 10^{-4}	10^{-4} to Max Score	Max score	Total
H FAR	1	97.70%	0.08%	0.01%	0.00%	97.79%
	<1 to 10^{-2}	1.27%	0.00%	0.00%	0.00%	1.27%
	10^{-2} to 10^{-4}	0.93%	0.01%	0.00%	0.00%	0.93%
	10^{-4} to Max Score	0.01%	0.00%	0.00%	0.00%	0.01%
	Max score	0.00%	0.00%	0.00%	0.00%	0.00%
	Total	99.90%	0.09%	0.01%	0.00%	100.00%
Genuine (n=64,867)		I FAR				
		1 to 10^{-2}	10^{-2} to 10^{-4}	10^{-4} to Max Score	Max score	Total
H FAR	1	0.16%	0.01%	0.05%	0.00%	0.22%
	<1 to 10^{-2}	0.02%	0.00%	0.01%	0.00%	0.03%
	10^{-2} to 10^{-4}	0.11%	0.04%	0.15%	0.00%	0.30%
	10^{-4} to Max Score	0.10%	0.09%	12.88%	39.09%	52.17%
	Max score	0.00%	0.00%	0.02%	47.26%	47.29%
	Total	0.39%	0.14%	13.12%	86.36%	100.00%

Table 12: Correspondence between groups of imposter and genuine scores for matchers H and I

(Left thumbs; NBDF06 dataset)

Table 12 shows the ability of Matchers H and I to discriminate genuines from imposters in those cases where the other matcher returned a score equivalent to a FAR between 1 and 10^{-2}. For example, of the 0.22% of the genuines for which H returned a score equivalent to FAR=1, I returned scores equivalent to FAR<10^{-4} for 0.05% — nearly a quarter of those cases.

10 FpVTE MST Results

[FpVTE] included a brief discussion of the correlation of system results, the potential for fusion, and a single simplified example of fusion. This section provides a more rigorous examination of the effect of matcher fusion on the FpVTE MST matchers.

Matcher scores for each pair of matchers were fused using the Best Linear fusion technique. As noted there, best linear fusion is not optimal, but is effective when detailed score distribution analysis and curve fitting is not practical.

Table 13 reports the TAR at FAR=10^{-4} for each matcher alone (results along diagonal) and for the fused results of each matcher pair. Table 14 reports the corresponding improvement in FRR relative to the stronger matcher of each pair. Fusion is beneficial when the fused TAR is substantially higher than that of either contributing matcher alone. Notice that

- Fusion was beneficial in almost all cases:
  - In many cases, the relative improvement in FRR (over the stronger matcher) is 30-40%.
  - The few cases of degradations in performance are not statistically significant: these all involve the three most accurate matchers and a very small number of misclassifications.
- The fused results do not change the overall performance ranks of the more accurate systems:

  - For the three most accurate matchers (NEC, Cogent, Sagem)[16], no fusion combination results in a statistically significant improvement.
  - The only fused results that surpass the base NEC performance (99.5%) involve NEC.
  - The only fused results that surpass the base Cogent performance (99.2%) involve Cogent or NEC.
  - The only fused results that surpass the base Sagem performance (98.3%) involve Sagem, Cogent, or NEC.
- Matchers that benefit from fusion with the NIST matcher might be improved by incorporating technology from the NIST matcher, which is freely available in source code form [NFIS].

[16] Specific hardware and software products identified in this report do not imply recommendation or endorsement by the National Institute of Standards and Technology.

	123ID	Antheus	Avalon	Biolink	Cogent	GoldenF	Identix	Motorola	NEC	Neuro	NIST	Phoenix	Sagem	Techno	Ultrascan
123ID	0.333	**0.170**	0.289	**0.142**	0.008	**0.186**	0.094	0.064	0.007	0.058	**0.129**	0.219	0.018	0.281	**0.125**
Antheus	**0.170**	0.232	**0.195**	**0.120**	0.010	**0.148**	**0.082**	0.061	0.005	**0.047**	**0.102**	**0.163**	0.019	**0.191**	**0.088**
Avalon	**0.289**	**0.195**	0.459	0.162	0.009	**0.216**	0.100	0.064	0.005	0.060	0.142	**0.228**	0.019	**0.281**	**0.118**
Biolink	**0.142**	**0.120**	0.162	0.166	0.008	**0.123**	**0.086**	0.061	0.005	0.056	**0.103**	**0.135**	0.018	0.162	**0.097**
Cogent	0.008	0.010	0.009	0.008	0.009	0.009	0.009	0.008	0.003	0.008	0.009	0.009	0.006	0.009	0.008
Goldenfinger	**0.186**	**0.148**	**0.216**	**0.123**	0.009	0.291	**0.080**	0.066	0.005	**0.052**	**0.115**	**0.182**	0.017	**0.208**	**0.110**
Identix	0.094	**0.082**	0.100	**0.086**	0.009	**0.080**	0.101	**0.053**	0.005	**0.043**	**0.076**	**0.087**	0.017	0.096	**0.074**
Motorola	0.064	0.061	0.064	0.061	0.008	0.066	**0.053**	0.066	0.005	**0.035**	0.059	0.065	0.017	0.066	0.058
NEC	0.007	0.005	0.005	0.005	0.003	0.005	0.005	0.005	0.005	0.004	0.005	0.005	0.004	0.005	0.005
Neuro	0.058	**0.047**	0.060	0.056	0.008	**0.052**	**0.043**	**0.035**	0.004	0.060	**0.052**	**0.052**	0.016	0.060	**0.051**
NIST_VTB	**0.129**	**0.102**	0.142	**0.103**	0.009	**0.115**	**0.076**	0.059	0.005	**0.052**	0.154	**0.128**	0.018	0.145	**0.091**
Phoenix	**0.219**	**0.163**	**0.228**	**0.135**	0.009	**0.182**	**0.087**	0.065	0.005	**0.052**	**0.128**	0.281	0.019	**0.216**	**0.106**
Sagem	0.018	0.019	0.019	0.018	0.006	0.017	0.017	0.017	0.004	0.016	0.018	0.019	0.019	0.019	0.017
Techno	**0.281**	**0.191**	**0.281**	0.162	0.009	**0.208**	0.096	0.066	0.005	0.060	0.145	**0.216**	0.019	0.523	**0.124**
Ultrascan	**0.125**	**0.088**	**0.118**	**0.097**	0.008	**0.110**	**0.074**	0.058	0.005	**0.051**	**0.091**	**0.106**	0.017	**0.124**	0.137

Table 13: FRR at FAR = 10^{-4} for fused FpVTE MST matchers. Pre-fusion performance is shown along the diagonal. Results in blue indicate cases in which fusion is beneficial and statistically significant; in no cases were degradations statistically significant.

	123ID	Antheus	Avalon	Biolink	Cogent	GoldenF	Identix	Motorola	NEC	Neuro	NIST	Phoenix	Sagem	Techno	Ultrascan
123ID		**27%**	13%	**15%**	8%	**36%**	7%	3%	-39%	4%	**16%**	22%	5%	16%	**9%**
Antheus	**27%**		**16%**	**28%**	-14%	**36%**	**19%**	7%	0%	**22%**	**34%**	**30%**	1%	**18%**	**36%**
Avalon	**13%**	**16%**		2%	0%	**26%**	2%	3%	-1%	0%	8%	**19%**	2%	**39%**	**14%**
Biolink	**15%**	**28%**	2%		11%	**26%**	**15%**	8%	-4%	7%	**33%**	**19%**	5%	2%	**29%**
Cogent	8%	-14%	0%	11%		1%	-1%	8%	34%	5%	0%	0%	29%	0%	4%
Goldenfinger	**36%**	**36%**	**26%**	**26%**	1%		**21%**	0%	0%	**13%**	**25%**	**35%**	11%	**29%**	**20%**
Identix	7%	**19%**	2%	**15%**	-1%	**21%**		**20%**	0%	**29%**	**24%**	**14%**	9%	5%	**27%**
Motorola	3%	7%	3%	8%	8%	0%	**20%**		0%	**42%**	11%	2%	8%	1%	12%
NEC	-39%	0%	-1%	-4%	34%	0%	0%	0%		12%	-4%	0%	16%	-5%	-6%
Neuro	4%	**22%**	0%	7%	5%	**13%**	**29%**	**42%**	12%		**15%**	**14%**	14%	0%	**16%**
NIST_VTB	**16%**	**34%**	8%	**33%**	0%	**25%**	**24%**	11%	-4%	**15%**		**17%**	6%	6%	**34%**
Phoenix	**22%**	**30%**	**19%**	**19%**	0%	**35%**	**14%**	2%	0%	**14%**	**17%**		0%	**23%**	**22%**
Sagem	5%	1%	2%	5%	29%	11%	9%	8%	16%	14%	6%	0%		0%	10%
Techno	**16%**	**18%**	**39%**	2%	0%	**29%**	5%	1%	-5%	0%	6%	**23%**	0%		**9%**
Ultrascan	**9%**	**36%**	**14%**	**29%**	4%	**20%**	**27%**	12%	-6%	**16%**	**34%**	**22%**	10%	**9%**	

Table 14: Percent FRR improvement over the more accurate of the two matchers; compare to the previous table. Statistically significant improvements are in blue. In no cases were degradations statistically significant.

Figure 21 shows example ROCs for pairwise fusion of FpVTE MST matchers with the NIST VTB matcher.

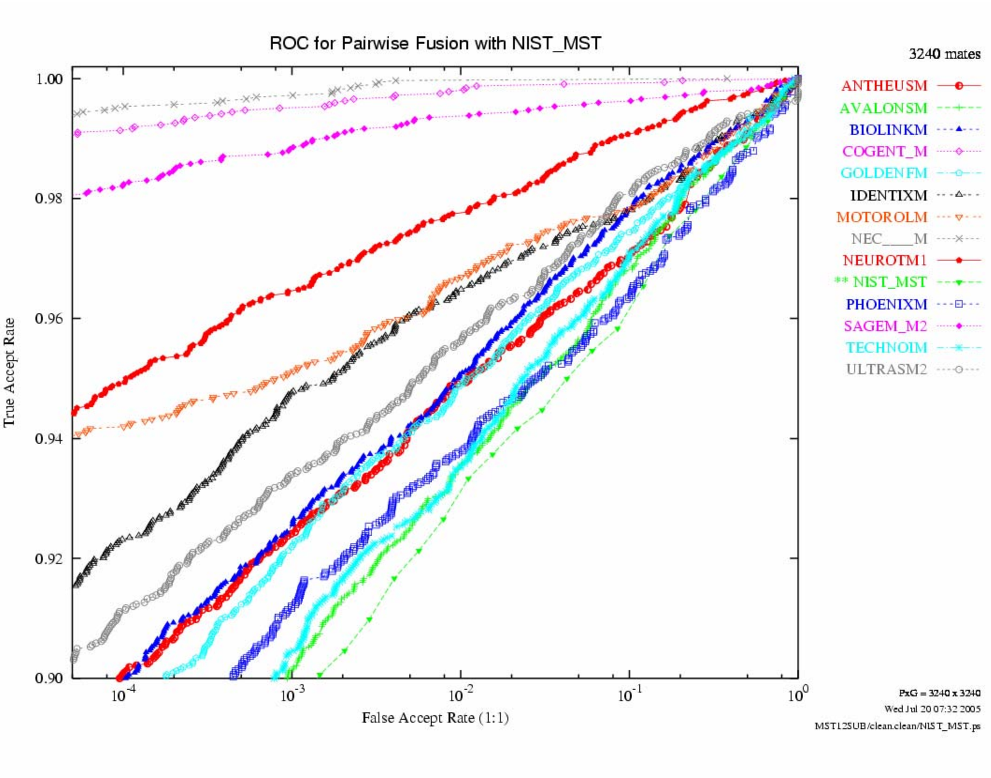

Figure 21: "Best Linear" fusion of FpVTE MST matchers with NIST VTB

11 Comparison of Multi-Matcher, Multi-Instance, and Multi-Modal Performance

Although matcher fusion clearly improves accuracy, that improvement is much less than what can be achieved with instance or mode fusion. For example, Figure 22 compares instance fusion of right and left index fingers (solid lines) against matcher fusion (green lines) using matchers H (red) and I (black), on the NBDF06 dataset. Single-instance multi-matcher fusion showed moderate improvement in accuracy (about 15-25% reduction in FRR where FAR=10^{-4}), where single-algorithm multi-instance fusion showed a dramatic improvement (about 80-85% reduction in FRR where FAR=10^{-4}).

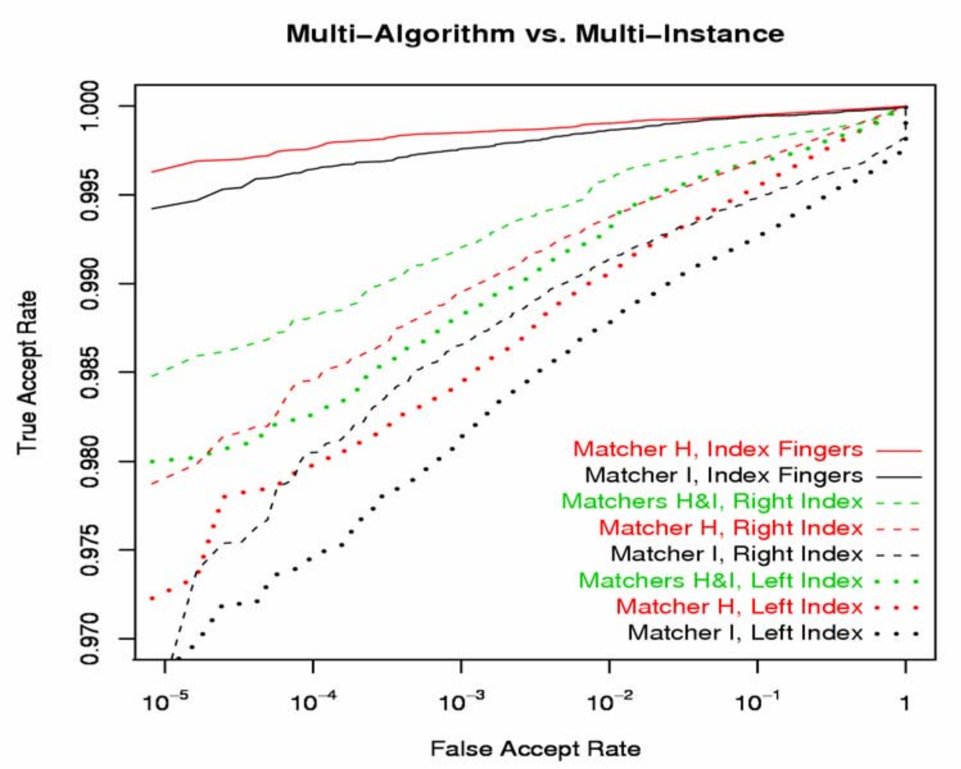

Figure 22: Comparison of matcher (algorithm) and instance fusion. (All fusion was Product of Ratios, using the NBDF06 dataset)

Table 7 summarizes the variation of performance improvement for the three categories of fusion (measured at FAR=10^{-4}).[17]

	Multi-Instance			Multi-Modal			Multi-Matcher			
		N=45			N=30		N=3		N=10	
	H fingers	I fingers	Q fingers	H+face	I+face	Q+face	Faces	H+I	H+Q	I+Q
Min	59%	48%	51%	68%	71%	64%	10%	14%	8%[18]	9%
Median	83%	79%	72%	74%	76%	75%	10%	25%	20%	20%
Average	82%	78%	71%	74%	77%	74%	11%	25%	16%	20%
Max	90%	90%	84%	80%	84%	79%	13%	33%	32%	32%

Table 15: Reduction in FRR where FAR = 10^{-4}, relative to the stronger of the inputs.[19] (All fusion was Product of Ratios, using the NBDF06 dataset)

Two key points can be derived from these results:

[17] This is a combination of Table 7 and Table 11.

[18] One H+Q data run that was not well-optimized resulted in an anomalous performance degradation (–7%), but subsequent optimization yielded an FRR improvement of 8%. This highlights the fact that results are sensitive to optimization procedures. No other runs were revised.

[19] For example, fusion of inputs that separately have FRRs of 87% and 90% (at 10^{-4}), resulting in a FRR of 92%, is a 20% relative improvement in FRR.

- Matcher fusion generally results in a 10-30% reduction in FRR, which is much less than for instance or mode fusion. This should be expected due to the limited independence in matcher fusion.
- Instance fusion generally results in a 50-90% reduction in FRR, whereas mode fusion generally results in a 65-85% reduction in FRR. In general, there was about the same level of improvement when fusing a fingerprint with a face as when fusing two fingerprints. As discussed in Appendix B.1, once multiple fingerprint scores have been fused, face scores become more effective than another finger instance due to their independence.

12 Conclusions

This study found that

- Matcher fusion often does produce a substantial increase in accuracy, although typically much less than that for instance or mode fusion.
  - Matcher fusion generally results in a 10-30% reduction in FRR, while instance and mode fusion generally result in a 50-90% reduction in FRR.
  - This should be expected due to the greater degree of data independence for instance and mode fusion.
- Some combinations of matchers are more effective than others. Substantial accuracy improvements were achieved even when combining some of the most accurate matchers.
- It should be noted that accuracy is not the only consideration in system design. Fusing two weaker systems in parallel might have cost or resource advantages over a single more accurate system.
- Although the scores being fused clearly are not independent, the Product of Likelihood Ratios fusion technique was consistently effective on the NBDF06 dataset. In this technique, the joint densities are estimated using an independence assumption. This result demonstrates that the large training sets required to model joint distributions are not required in order to implement effective optimization per the Neyman-Pearson Lemma.
- The findings of this analysis are based on the fusion of single-finger matcher scores, and frontal image face scores. The extent to which these results generalize to other biometric modalities is not known.

Appendix B.3: Score-Level Fusion of Multiple Fingerprint Samples

Contents

13 Introduction: Multi-Sample Score-Level Fusion

Multi-sample data refers to samples acquired from the same source, such as multiple images of a single fingerprint, images of the same face, or recordings of a speaker. This report describes experiments assessing the effectiveness of one use of multi-sample data: retaining multiple samples in the gallery, such as when an additional enrollment is added for every encounter with a subject.

Many systems encounter multi-sample data but do not take full advantage of it. Such systems explicitly or implicitly involve design decisions whether to use multiple samples for selection or fusion:

- On the gallery side after a successful match, a matching system may
  - Always retain the original enrolled sample in the gallery,
  - Retain the "best" sample yet encountered for this subject, or
  - Retain the samples from each encounter in the gallery.
- On the probe side, a collection process may
  - Always collect a single sample,
  - Use image quality metrics to determine whether image recapture is appropriate,
  - Use image quality metrics to select the "best" image from a series of images, or
  - Collect multiple samples to use as probes.

It is tempting to assume symmetric benefits on the probe and gallery sides, that similar results might be expected from having a second gallery sample as a second probe sample. However, multiple probe samples can be expected to have highly correlated sample quality problems, as both are likely to come from one encounter with the subject. Multiple gallery samples collected at different times can be expected to have a greater degree of independence and therefore should be effective in fusion.

13.1 Background

[Grother-04] included an evaluation of multi-sample fusion using the face recognition data from [FRVT2003], showing a clear benefit:

> *Performance improves substantially with K [K=number of samples] Only Visionsphere shows an anomalous decline in performance. For the leading systems much of the improvement is realized for K = 2. Thereafter*

> *returns typically diminish. For example, looking at the Identix system for summed raw scores, for the five image population, the false non-match rate (i.e. 1 − PTA) decreases from 8% (K = 1) to 6% (K = 2) with only a further reduction of 0.7% for two more scores. [Grother-04, p. 27]*

The [Goats] study was an evaluation of multi-sample data: ten encounters each for 6,000 subjects. That study showed a great deal of variation between samples from each subject, in terms of scores and image quality metrics. That report concluded

> *A hard to match fingerprint, therefore, is indicative of problems with that specific fingerprint image, and does not mean another fingerprint from the same finger would be hard to match. [Goats, p. 23]*

Since matcher scores vary substantially between collection encounters, then the risk of a failure to match (low matcher score) should decrease if the gallery contains samples from more than one collection encounter.

[Hopper-05] discussed a study that was conducted to determine the efficacy of searching latent fingerprints against a gallery containing both rolled and flat (segmented slap) images, with these results:

- 59.8% of successful searches matched against both rolls and flats
- 27.8% matched only against rolls
- 12.4% matched only against flats

The improvement that resulted in using flats in addition to rolls could be attributed either to the use of two different types of fingerprint capture (roll vs. slap) or to the use of multiple samples in the gallery.

14 Approach

The method used in this study was to measure how accuracy would be affected if a gallery retained two samples per subject rather than just one.

14.1 Data and Matchers

As discussed in Appendix A, the NBDF06 dataset used throughout these studies had 64,867 mated subjects, with two face/fingerprint sets each. Of these, 4,015 subjects had 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 4,015 * 4,014 off-diagonal scores of the similarity matrices.[20] The matchers used were the H, I, and Q fingerprint matchers.

14.2 Scores

The scores were generated as if two of the samples (Sample_1 and Sample_2) were enrolled in the gallery, and the third sample was used as the probe. Therefore, for each pair of subjects compared, Score_A =

[20] The background set was not fully randomized, but was approximately randomized using a block design where each probe was compared to a gallery of approximately 100 subjects. The gallery subjects used varied between probes.

match(Sample1, Sample3) and ScoreB = match(Sample2, Sample3). For each genuine or imposter comparison, 60 scores were produced (10 fingers x 3 matchers x 2 scores).

14.3 Fusion Techniques

As the scores to be fused have the same score distribution (same finger, same matcher), they are fused by Simple Sum of Raw Scores, which is shown in Appendix C to be highly effective when the scores come from a single matcher and corresponding fingers. Scores were also fused by Maximum of Raw Scores, *Max(ScoreA, ScoreB)*, which implements OR decision level fusion.

For the 30 combinations of finger positions and matchers,

- Two baseline (unfused) ROCs were produced for ScoreA = match(Sample1, Sample3) and ScoreB = match(Sample2, Sample3)
- Two fused ROCs were produced for the two techniques

15 Results

The 30 sets of ROCs show very consistent behavior. Figure 23 and Figure 24 show characteristic results. In every case, the fused results result in a substantial improvement in accuracy. Note that there is no substantial difference in performance between the two fusion techniques.

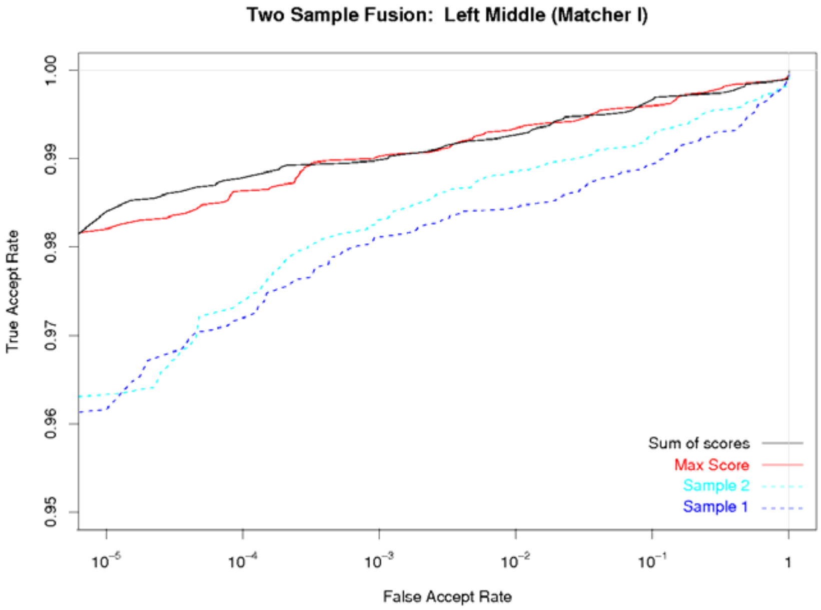

Figure 23: Effects of multi-sample fusion

(fingerprint matcher I, left middle finger, NBDF06 multi-sample data)

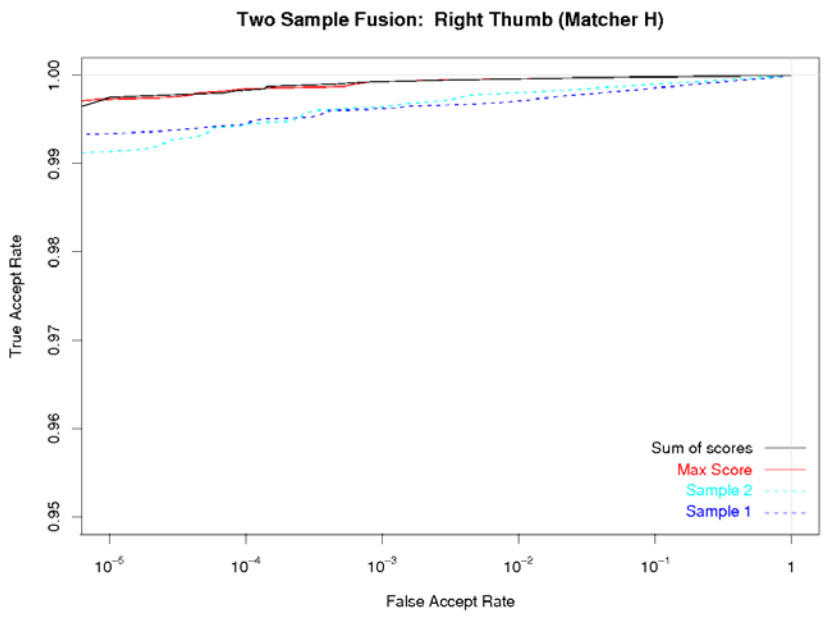

Figure 24: Effects of multi-sample fusion

(fingerprint matcher H, right thumb, NBDF06 multi-sample data)

Table 16 shows summary results for all finger positions, broken down by matcher and fusion technique. The improvement ranged from 45% to 73%.

Matcher	Technique	Min	Median	Max	Std. Dev.
H	Max of raw scores	52%	58%	73%	6%
H	Sum of raw scores	53%	57%	70%	5%
I	Max of raw scores	49%	55%	66%	6%
I	Sum of raw scores	49%	56%	72%	8%
Q	Max of raw scores	45%	51%	57%	3%
Q	Sum of raw scores	45%	52%	57%	4%

Table 16: Reduction in FRR at FAR=10^{-4}, relative to the stronger input. Results are computed over all ten finger positions.

16 Conclusions

This study showed that there is a clear benefit from multi-sample, score-level fusion: false reject rates were reduced from 45% to 73%. These benefits are due to sample-specific variability, which can be largely be attributed to quality problems present in one but not all samples, and the overlap of areas contained in each sample. Because data quality is a problem in every biometric capture process, multi-sample fusion can be expected to be effective for other biometric modalities.

Based on the assumption that the multiple samples share common score distributions, the simplest of fusion techniques, Simple Sum and Max of Raw Scores were used. These techniques require no training data and are expected to scale reliably.

Multi-sample fusion from the use of multiple enrollments is likely to be of interest since it is based von the retention of existing data rather than 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.

17 References

[Dass-05] Dass, Nandakumar, and Jain; “A Principled Approach to Score Level Fusion in Multimodal Biometric Systems”.

[Fierrez-03] J. Fierrez-Aguilar, J. Ortega-Garcia, D. Garcia-Romero and J. Gonzalez-Rodriguez, “A comparative evaluation of fusion strategies for multimodal biometric verification”, in *Proc. 4th IAPR Intl. Conf. on Audio- and Video-based Biometric Person Authentication, AVBPA*, Springer LNCS-2688, pp. 830-837;

[Fierrez-05b] Fierrez-Aguilar, Nanni, Ortega-Garcia, Cappelli, Maltoni; “Combining Multiple Matchers for Fingerprint Verification: A Case Study in FVC2004”; *ICIAP* 2005, LNCS 3617, pp. 1035–1042, 2005.

[Fierrez-06] J. Fierrez-Aguilar, Y. Chen, J. Ortega-Garcia and A. K. Jain; “Incorporating image quality in multi-algorithm fingerprint verification”; *Proc. of International Conference on Biometrics* (ICB); Hong Kong, January 5-7, 2006.

[FpVTE] C. Wilson, R.A. Hicklin, H. Korves, B. Ulery, M. Zoepfl, M. Bone, P. Grother, R. Micheals, S. Otto, C. Watson; “Fingerprint Vendor Technology Evaluation 2003”; *NIST Interagency Report* 7123. June 2004.

[FRVT2002] Phillips, Grother, Micheals, Blackburn, Tabassi, Bone; “Face Recognition Vendor Test 2002”; *NIST Interagency Report* 6965; March 2003.

[Goats] A. Hicklin, C. Watson, B. Ulery; “The Myth of Goats: How many people have fingerprints that are hard to match?”; *NIST Interagency Report* 7271, Sept. 2005.

[Griffin-05] P. Griffin; “Optimal Fusion for Multi-Biometric Identity Verification”; Identix Research Preprint RDNJ-05-0001, Jan. 2005.

[Grother-04] P. Grother, “Face Recognition Vendor Test 2002 - Supplemental Report”, *NIST Interagency Report 7083*, 02 February 2004.

[Hopper-05] T. Hopper; “Identification Flats”; ANSI/NIST Fingerprint Standard Workshop I; April 2005.

[Indovina-03] M. Indovina, U. Uludag, R. Snelick, A. Mink, A. Jain; “Multimodal Biometric Authentication Methods: A COTS Approach”; *Proc. MMUA 2003, Workshop on Multimodal User Authentication*.

[Jain-99b] A.K. Jain, L.Hong, and Y. Kulkarni; “A Multimodal Biometric System using Fingerprints, Face and Speech”; *2nd Int'l Conference on Audio- and Video-based Biometric Person Authentication*, Washington D.C., pp. 182-187, March 22-24.

[Jain-99c] A. K. Jain, S. Prabhakar and S. Chen; “Combining Multiple Matchers for a High Security Fingerprint Verification System”; *Pattern Recognition Letters*, Vol 20, No. 11-13, pp. 1371-1379, 1999.

[Jain-05] Jain, Nandakumar, Ross; “Score Normalization in Multimodal Biometric Systems”; *Pattern Recognition* 38 (2005) 2270-2285; Oct. 2004.

[Kittler-98] J. Kittler, M. Hatef, R. Duin, and J. Matas; “On Combining Classifiers”; *IEEE Trans on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 3, March 1998

[NFIS] C. Watson, et al; "NIST Fingerprint Image Software“.

[Poh-05c] Norman Poh and Samy Bengio; “Can Chimeric Persons Be Used in Multimodal Biometric Authentication Experiments?”; *Institut Dalle Molle d'Intelligence Artificielle Perceptive*, IDIAP-RR 05-20, 2005.

[Poh-05d] Norman Poh and Samy Bengio; “Towards Explaining the Success (Or Failure) of Fusion in Biometric Authentication”; *Institut Dalle Molle d'Intelligence Artificielle Perceptive*, IDIAP-RR 05-43, 2005.

[Poh-05e] Poh & Bengio “A Score-Level Fusion Benchmark Database For Biometric Authentication”.

[SC37-24722] SC37 Working Draft Technical Report 24722 on “Multimodal and Other Multibiometric Fusion”; 2006-02-14 [ISO/IEC JTC 1/SC 37 N1506]

[Scott-05] C. Scott and R. Nowak; “A Neyman-Pearson Approach to Statistical Learning”; 2005.

[SDK] Craig Watson, Charles Wilson, Karen Marshall, Mike Indovina, & Rob Snelick; “Studies of One-to-One Fingerprint Matching with Vendor SDK Matchers”; *NIST Interagency Report* 7221; April 2005.

[Snelick-03] R. Snelick, M. Indovina, J. Yen, and A. Mink, “Multimodal Biometrics: Issues in Design and Testing”; *Proceedings of Fifth International Conference on Multimodal Interfaces*, (Vancouver, Canada), November 2003.

[Snelick-05] R. Snelick, U. Uludag, A. Mink, M. Indovina, and A. Jain, “ Large Scale Evaluation of Multimodal Biometric Authentication Using State-of-the-Art Systems”, *IEEE Transactions on Pattern Analysis and Machine Intelligence* , Vol. 27, No. 3, pp. 450-455, March 2005.

[Wang-03] Wang, Tan, Jain; “Combining Face and Iris Biometrics for Identity Verification”; 2003.