Studies of Biometric Fusion
Appendix C

Evaluation of Selected Biometric Fusion Techniques

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

[1] Mitretek Systems

[2] Independent consultant to Mitretek Systems

[3] National Institute of Standards and Technology

20 July 2006

Abstract

This report summarizes and evaluates several common score-level biometric fusion techniques. The literature contains numerous proposals for score-level biometric fusion algorithms. Selecting the most effective fusion techniques depends on operational issues such as accuracy requirements, availability of training data, and the validity of simplifying assumptions. Of the techniques evaluated, product of likelihood ratios and logistic regression were found to be highly effective.

Contents

1 Introduction

In the course of this study, several score-level fusion techniques were implemented and applied to a variety of multi-biometric problems. This paper discusses what techniques were investigated, why those techniques were selected, and evaluates how the techniques compare in terms of accuracy. The paper also offers guidance for selecting among those techniques, as they are not all equally suited to different problems. Techniques for modeling score distributions are discussed in Appendix E. The set of techniques discussed are not fully representative of the breadth of alternatives, but are varied and instructive.

The purpose of this paper is to convey how common biometric fusion strategies compare with respect to performance, real-world applicability, and ease of implementation to persons selecting, designing or calibrating fusion algorithms.

2 Previous Work

One finds in the literature a plethora of techniques for fusing biometric scores. One also finds other communities of researchers solving similar problems in dissimilar ways: much of the growing body of research that supports data fusion is not specifically targeted to the narrow application area of biometric score fusion and its special needs. The biometric community has done a great deal of work in adapting techniques from a variety of specialized fields (e.g., statistics, pattern recognition, artificial intelligence, medicine) for biometric-specific purposes, but this is still a fertile area for further work.

A strong theoretical basis exists for biometric fusion (e.g., [Kittler-98; Verlinde-00; Poh-05d]). Many researchers have demonstrated that fusion is effective in the sense that the fused scores provide much better discrimination than the individual scores (e.g., [Ben-Yacoub-99; Dass-05; Fierrez-03; Griffin-05; Grother-02; Indovina-03; Jain-99b; Korves-05; Ross-01; Snelick-05; Wang-03]). Such results have been achieved using a variety of fusion techniques.

Several recent papers have compared various techniques on empirical data. Selections of recent results that help define "conventional wisdom" include:

- Kittler *et al* [Kittler-98] evaluated several classifier combination rules on frontal face, face profile, and voice biometrics (using a database of 37 subjects). They found that the "sum of *a posteriori* probabilities" rule outperformed the product, min, max, median, and majority of *a posteriori* probability rules (at EER) due to its resilience to errors in the estimation of the densities.
- Ben-Yacoub *et al* [Ben-Yacoub-99] evaluated five binary classifiers on combinations of three face and voice modalities (database of 295 subjects). They found that (a) a support vector machine and Bayesian classifier achieved almost the same performances; and (b) both outperformed Fisher's linear discriminant, a C4.5 decision tree, and a multilayer perceptron.
- Fierrez-Aguilar *et al* [Fierrez-03] found that a support vector machine outperformed (at EER) the sum of normalized scores when fusing face, fingerprint and signature biometrics (database of 100 subjects and 50 chimeras[1]).
- Jain *et al* [Jain-05] applied the sum of scores, max-score, and min-score fusion methods to normalized scores of face, fingerprint and hand geometry biometrics (database of 100 users, based on a fixed TAR). The normalized scores were obtained by using one of the following

[1] Chimeras are composites of data representing virtual "subjects" that combine biometrics from multiple individuals.

techniques: simple distance-to-similarity transformation with no change in scale (STrans), min–max, z-score, median-MAD, double sigmoid, tanh, and Parzen. They found that (a) the min–max, z-score, and tanh normalization schemes followed by a simple sum of scores outperformed other methods; (b) tanh is better than min-max and z-score when densities are unknown; and (c) optimizing the weighting of each biometric on a user-by-user basis outperforms generic weightings of biometrics.

- Snelick *et al* [Snelick-05] compared combinations of z-score, min-max, tanh and adaptive (two-quadrics, logistic and quadric-line-quadric) normalization methods and simple sum, min score, max score, matcher weighting, and user weighting fusion methods (database of about 1000 users, at a fixed FAR). They found that (a) fusing COTS fingerprint and face biometrics does outperform unimodal COTS systems, but the high performance of unimodal COTS systems limits the magnitude of the performance gain; (b) for open-population applications (e.g., airports) with unknown posterior densities, min-max normalization and simple-sum fusion are effective; (c) for closed-population applications (e.g. an office), where repeated user samples and their statistics can be accumulated, QLQ adaptive normalization and user weighting fusion methods are effective.
- Korves *et al* compared various parametric techniques on the BSSR1 dataset [Korves-05]. That study showed that the Best Linear technique performed consistently well, in sharp contrast to many alternative parametric techniques, including simple sum of z-scores, Fisher's linear discriminant analysis, and an implementation of sum of probabilities based on a normal (Gaussian) assumption.

These studies are illuminating, but many suffer from limited available data: the datasets are too small to evaluate performance at low FAR, and often are not representative of operational data. Some of these studies also suffer from other problems:

- Conclusions are based on simplifying assumptions, such as independence or normal score distributions, often with no investigation into the validity of those assumptions.
- Techniques are often not fully defined, so that it is not clear if results are due to the underlying concept behind an algorithm or the decisions made in the implementation. In general, comparisons of techniques are in fact comparisons of implementations of those techniques, so that negative results could be the result either of the underlying technique or an imperfect implementation — this is compounded if the details are not fully defined.
- The performance of many techniques depends upon information that must be supplied but may not be available, such as prior probabilities or score distributions. The results of the evaluation may hinge on the validity of assumptions (or equivalently, the quality of available information and the accuracy with which it is modeled).
- Confidence intervals on results are not discussed (and often large). Several papers discuss the computation of confidence intervals on ROCS [Wayman; Bolle-00b; Macskassy-04; Scott-06; Micheals-03]
- Some studies are based on approaches to classification that seek to minimize the probability of error or expected Bayes' cost (see [Scott-06]). Such approaches may have limited applicability in biometrics.

Thus, despite the progress that has been made in this field, there remains a clear need for large-scale empirical comparisons of fusion techniques. There is also a need for guidance on the implementation and selection of techniques.

3 Fusion Techniques Evaluated

We selected eight techniques based on performance in previous work. We implemented and systematically compared these techniques on a wide range of fusion problems, involving different matchers, modalities and instances. Table 1 briefly describes each technique.

Technique	Description	Rationale for Choice
Simple Sum of Raw Scores	Matcher scores are simply added, with no prior normalization. Scores are neither rescaled, nor weighted to account for differences in matcher accuracy.	Included largely to demonstrate its limited applicability, which includes situations where scores have comparable distributions, such as two fingers scored by one matcher, as in the NIST 2-finger SDK evaluations [SDK2].
Simple Sum of Z-normalized Scores	1. The mean and standard deviation of the imposter score distribution is estimated from sample data. 2. Scores are normalized by subtracting the mean of the imposter distribution, then dividing by the standard deviation of the imposter distribution. 3. The normalized scores are simply added without weighting.	Simple linear fusion technique.
Best Linear	Linear fusion (weighted sum), with an optimal slope (hyperplane) determined empirically on joint sample distributions. Solution entails iteratively rotating decision boundary and evaluating TAR at a fixed FAR.	Many linear techniques are discussed in the literature, but each determines slope from prior distributional assumptions. Previously identified as a top performer [Korves-05], Best Linear is often far superior to alternative linear techniques.
Product of Likelihood Ratios	1. Probability density functions are separately modeled for each genuine and impostor distribution, using variable bandwidth kernels, log-linear tail tapering, and spike handling. 2. Likelihood ratios are computed from these models for each matcher. 3. Scores are normalized by transformation to their likelihood ratios 4. Normalized scores are simply multiplied	This multi-stage modeling process attempts to estimate, very directly, the information required for application of the (theoretically optimal) Neyman-Pearson Lemma, with minimal simplifying modeling assumptions.
Logistic Regression	1. The log of the density ratio is modeled (e.g. as a low-order polynomial function), then estimated from the training data by principal of maximum likelihood 2. Density ratios were modeled independently for each matcher; fusion was performed by adding the normalized scores (log densities). Alternatively, the joint density ratio may be modeled directly, in which case normalization and fusion are not distinct steps.	A standard statistical technique that closely approximates the theoretically optimal Neyman-Pearson Lemma.

Technique	Description	Rationale for Choice
Product of FARs	1. Scores are transformed to the estimated right-tail integral of the imposter density distribution. 2. Distribution modeling is identical to that used for product of likelihood ratios (but only for the imposter distribution).	Fusion technique proposed in [Griffin-04b] for use with the BioAPI standard. The estimation technique as implemented was developed here. FAR normalization is a practical alternative to ratio normalization when little is known about the genuine distribution.
Min of FARs Max of FARs	1. Scores are transformed to FAR, as with product of FARs. 2. As implied by the names, simply the minimum (or maximum) of the unfused FARs. Min and max are decision-level fusion techniques, but calibration requires knowledge of score distributions.	Fusion techniques proposed in [Griffin-04b] for use with the BioAPI standard.

Table 1: Selected Techniques, as implemented in this study

The ROC-based evaluation used here emphasizes tradeoffs between FAR and FRR. The error trade-off approach to evaluation was championed by J. Neyman and E. S. Pearson; the Neyman-Pearson Lemma [Neyman-33] defines a criterion under which the ROC can be optimized.[2] Recently, several authors have noted the merits of Neyman-Pearson optimization for this type of evaluation (e.g., [Scott-05], [Griffin-05], [Pepe-05]). As noted by Griffin,

> *"One good way of defining optimization of biometric fusion for the purposes of verification is to find the combination of signals from each biometric such that for a given False-class acceptance rate (FAR) the True-class rejection rate (TRR) is minimized. The solution to this problem was given […] by Neyman and Pearson. The Neyman-Pearson theorem simply states that this optimization is satisfied by using decision boundaries in the total sample or probability density space such that the ratio of genuine over impostor probability densities are held constant at the boundary. Therefore, the problem at hand is to implement in a simple way this meaning of optimal fusion." [Griffin-05]*

The product of likelihood ratios technique described in this paper represents a direct approach at estimating distribution densities from sample data for the purpose of Neyman-Pearson optimization. Logistic regression is a standard statistical technique for modeling these ratios directly. Much of the fundamental theory and implementation issues relating these techniques are summarized in other papers in this series (see Parts V, VI, and VII). This paper focuses on empirical observations and practical guidance derived from our experience.

The techniques tested are a small subset of those proposed in the literature. We include techniques expected to be appropriate in only a few circumstances (simple sum of raw scores and z-normalized scores), techniques expected to do well in general (product of likelihood ratios and logistic regression),

---

[2] 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) replace each point in the scatterplot with the ratio of genuines to imposters at that point, 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.

and some "reasonable" techniques known to practitioners (Best Linear, FAR-based methods). Many of the techniques in the literature are variations on similar concepts; many of the more effective techniques are based on likelihood ratios [Sedgwick][Dass-05]. Several techniques were not evaluated due to limited resources that might be considered for future work, such as the use of support vector machines.

Our use of FAR-based techniques was influenced by the BioAPI specification [BioAPI], and the ANSI Fusion Information Format [FIF] which use estimates of FAR as a means of making scores interoperable.

Evaluation of the techniques included comparisons of the full ROCs, some of which are reproduced here. Summary results are measured at FAR=10^{-4}: this specific operating point was selected for its comparability with FpVTE results, computational feasibility, as an appropriate baseline for use in identification systems.

Note that for most of these techniques, almost all of the work (and complexity) lies in modeling the univariate (pre-fusion) score distributions. In other words, if we distinguish between normalization (univariate) and fusion (multivariate), these techniques are successful because of painstaking normalization, and fusion is merely the sum or product of the normalized scores.

4 Comparative Performance Results

Figure 1 provides a summary comparison of the eight techniques on 2-way fusion problems constructed from the following score sets: Face matchers A, B, and C; Fingerprint matchers H, I, and Q for right index (RI) and left index (LI) fingers. The charts on the following pages show greater detail.

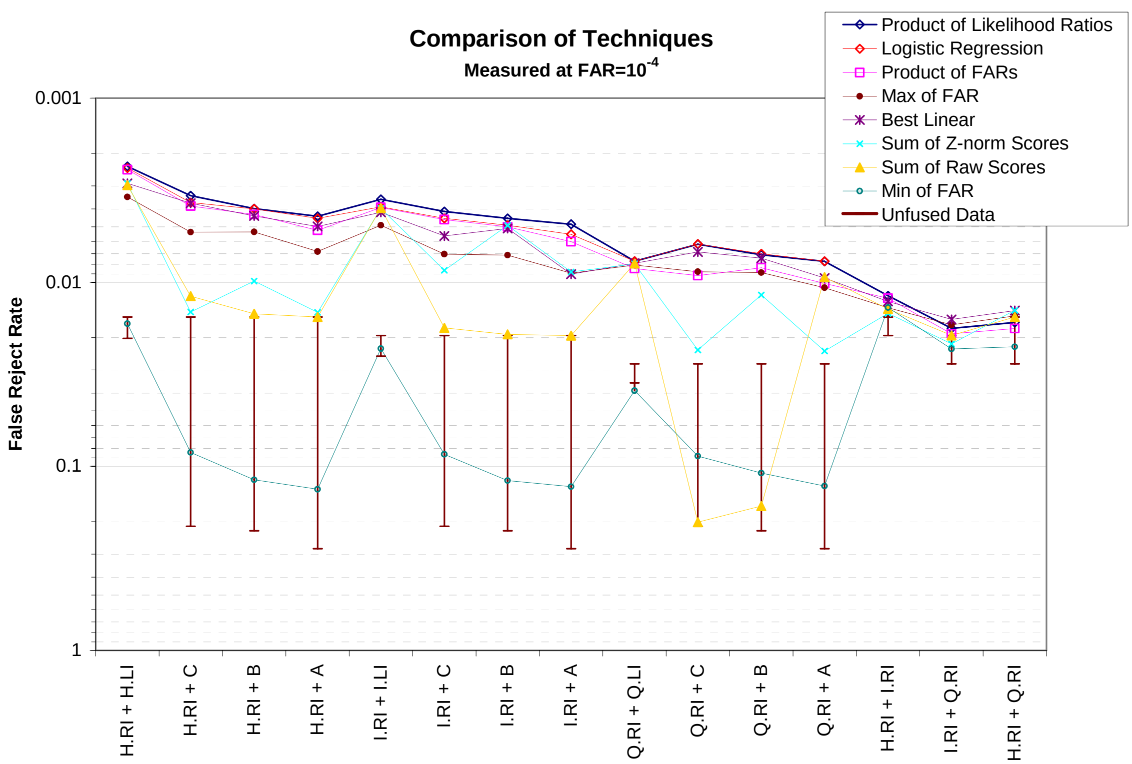

Figure 1: Eight selected techniques compared at FAR=10^{-4} on a variety of fusion tasks. Product of Likelihood Ratios performs consistently well.

This chart supports several key results:

- Product of likelihood ratios, logistic regression, product of FARs, Best Linear and max of FAR perform well consistently.
- Simple sum of raw scores performs well when the scores are of the same type (the same matcher on corresponding fingers). In other cases, z-normalization of the scores often helps, but results remain variable.
- Min of FAR is not effective.

Additionally, it should be noted that

- The probabilistic techniques are often sensitive to accurate modeling of the score distributions and implementation details (as discussed in TBD Vb and Vc). This may be a significant consideration when selecting a technique for a specific application (See Section 5).
- Max and min implement decision-level fusion, not score-level fusion (Max combines decisions with AND; Min with OR).[3] The findings of this study disprove the oft-repeated canard that decision-level fusion is ineffective: decision-level fusion was found to be highly effective, but not as effective as score-level fusion. Theoretically, Max of Likelihood Ratios (not implemented) should perform better than Max of FARs but not as well as the product of likelihood ratios. This was not verified empirically.

Figure 2 to Figure 5 provide sets of complete ROCs to show how the various techniques compare across the range of operating thresholds.

---

[3] Note that although fusion is performed at the decision level, the matchers must first be set to operate at a common decision threshold in order to maximize TAR. Calibration may utilize score-level information. The example ROCs represent the range of operating points achievable by varying this common decision threshold.

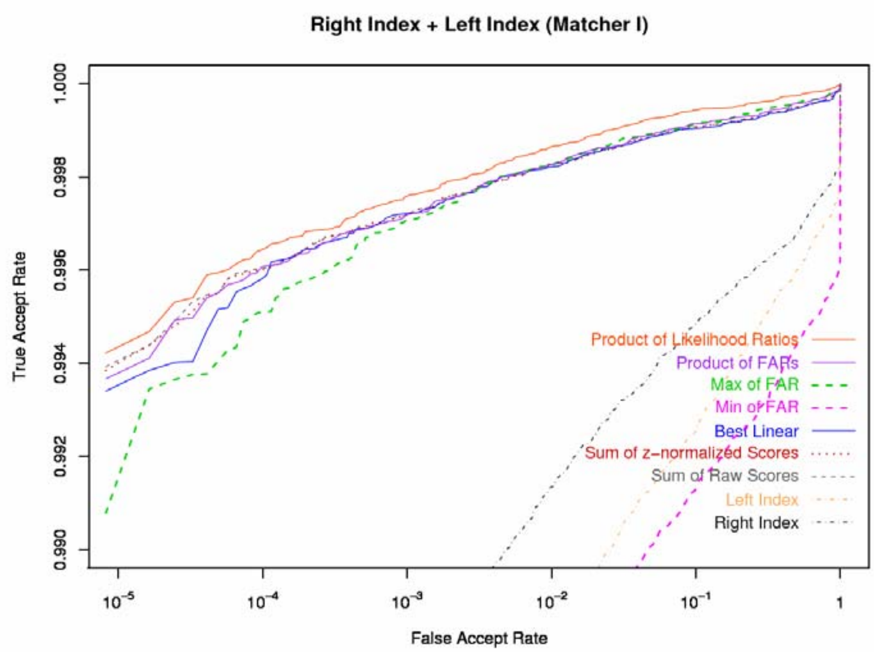

Figure 2: The result of fusing two different fingers using the same matcher. This is the least challenging type of fusion, so many techniques do well.

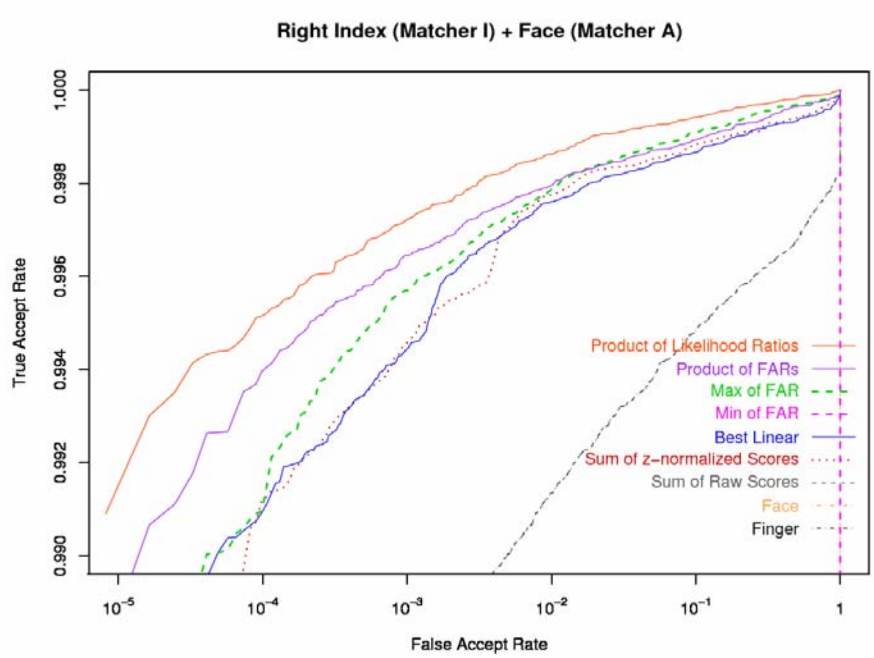

Figure 3: The result of fusing a face score and a finger score. Score distributions differ significantly, so benefits of better fusion techniques are more significant. The Face and Min of FAR lines are superimposed.

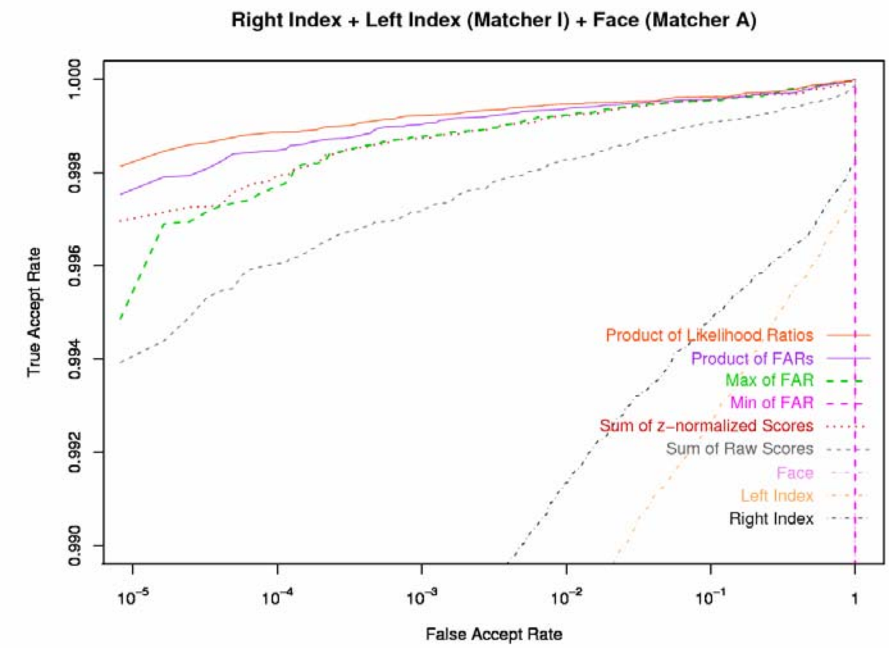

Figure 4: The result of fusing two different fingers (using the same matcher) with a face score.

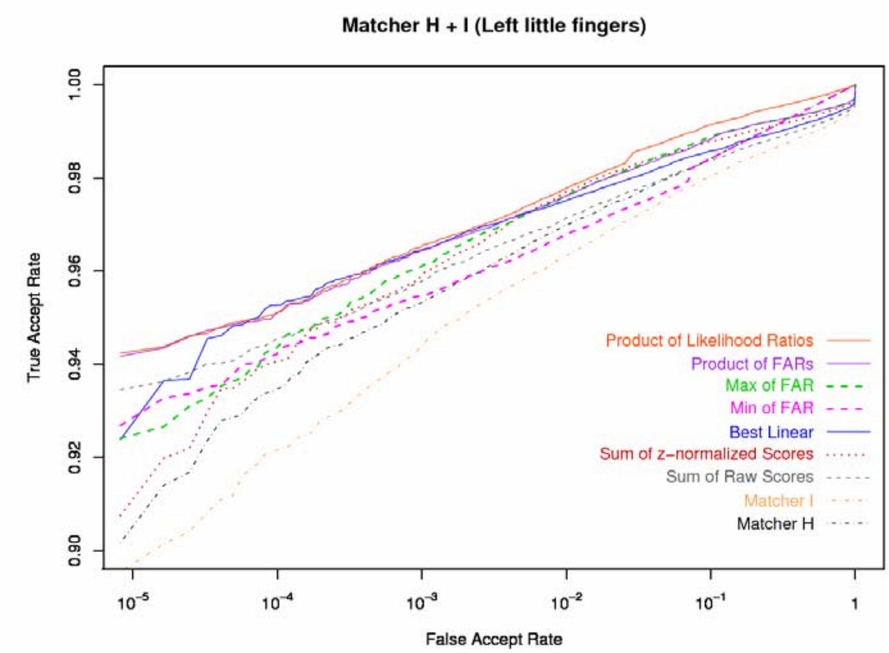

Figure 5: The result of fusing different matchers on the same source data (left little fingers). Although the scores are highly dependent, fusion still yields a benefit.

5 Guidance on Selecting Techniques

As seen in the preceding section, several fusion techniques perform well, but even among the most accurate techniques there are performance differences. One of the key considerations when selecting a technique should be what resources are available for training the algorithm. This includes many factors such as availability of training data, modeling tools and expertise, target operational settings, integrity of (possibly multiple) databases, and reliability of ground-truth information for training data. Of central concern is the training data itself and any other knowledge of the score distributions:

- Quantity of training data (separate counts of genuines and imposters)
- Quality of training data (e.g. representativeness of operational data, accuracy of ground-truth information)
- Design information (e.g. range of possible scores, whether scores represent similarity measures or probabilities, the meaning of special score values such as for FTEs, how multiple processing stages influence score distributions)
- Joint distributions (i.e., the training data includes multiple scores for each subject comparison)

Table 2 summarizes for each technique several factors that might influence the selection decision. These include simplifying theoretical assumptions (failure to satisfy these may substantially degrade performance); information required to implement the technique; observations on the difficulty of implementation and on the sensitivity of the technique to certain data characteristics; and how well the technique performed in this study. Note that this table is merely a summary: the companion papers in this series discuss the conceptual and implementation issues for these techniques.

Simple Sum of Scores	
Assumptions	Requires no training data. Assumes inputs have comparable scale and strength.
Implementation	Trivial: simple addition of scores.
Accuracy	Inconsistent and often quite bad, but good for fusion of corresponding fingers using the same matcher.
Simple Sum of Z-Scores	
Assumptions	Requires small, univariate sample distribution (imposters only). Assumes Gaussian distributions, comparable strength of inputs. Highly sensitive to assumption of comparable strength: the weaker input can pull results down. Not as sensitive to Gaussian assumption.
Implementation	Simple: Only requires normalization by standard deviation.
Accuracy	Inconsistent.
Best Linear	
Assumptions	Requires joint sample distributions. Assumes size of dataset is large enough to measure target FAR accurately, but solution generally was not highly sensitive to FAR. Does not assume fused scores are independent. Requires selection of appropriate target FAR. Insensitive to magnitude of outliers.
Implementation	Conceptually simple, but automated solution requires a special-purpose algorithm to determine the optimal weights for a given FAR.
Accuracy	Consistently near top

Product of Likelihood Ratios	
Assumptions	Assumes knowledge of univariate density distributions, for both genuine and imposter scores. Assumes fused scores are independent, but apparently not very sensitive to this assumption.
Implementation	Successful implementation requires accurate curve fitting of genuine and imposter density distributions, potentially a painstaking process requiring advanced statistics capabilities. Variable bandwidth kernel estimates yielded a good initial fit, but special care was required to achieve reasonable fits to tails and bounded distributions.
Accuracy	Consistently top.
Logistic Regression	
Assumptions	Generally requires joint sample distributions (does not assume independence). In this analysis, however, independence assumption was highly effective, so models relied exclusively on univariate distributions.
Implementation	Successful implementation requires accurate curve fitting of density distributions, which is requires statistical expertise, but is supported by standard statistical packages. Care must be taken to select appropriate modeling terms: low-order polynomial terms were found to suffice. Sensitive to outliers, but special values (e.g., spikes, FTEs) can be handled explicitly.
Accuracy	Consistently near top
Product of FARs	
Assumptions	Assumes knowledge of univariate probability distribution (impostors only). Appears to be a reasonable heuristic and does not require genuine data, but lacks a solid theoretical underpinning. Assumes fused scores are independent.
Implementation	Successful implementation requires accurate curve fitting of imposter density distributions, which is a painstaking process akin to Product of Likelihood Ratios.
Accuracy	Consistently near top
Min of FARs	
Assumptions	See Product of FARs. Note that score-level information is needed only to calibrate the decision thresholds, not during operations.
Implementation	See Product of FARs.
Accuracy	Consistently poor performer
Max of FARs	
Assumptions	See Product of FARs. Note that score-level information is needed only to calibrate the decision thresholds, not during operations.
Implementation	See Product of FARs.
Accuracy	Usually near top (inconsistent)

Table 2: Guidance on Selected Techniques

6 Conclusions

Several score-level fusion techniques were evaluated and compared in this study. There are several highly and consistently effective score-level fusion techniques to choose from. An important consideration when making a selection is what tools and information are available for modeling the score distributions.

- As implemented, the product of likelihood ratios was the most sophisticated and most accurate method. This technique requires careful modeling of the score distributions.

- Logistic regression is highly effective, and relies on standard statistical tools.
- Best linear is a conceptually simple technique. This technique requires joint sample distributions (training data). Implementation is relatively easy for 2-way fusion. It might be an appropriate choice for some evaluations, benchmarks, prototypes, etc., where highly optimal accuracy is not required.
- Product of FARs is most appropriate when no information on the genuine distribution is available; careful modeling of the imposter distribution is required.

Many proposed techniques are based on simplifying assumptions. While any of these techniques might be valid in a specific situation, they are not reliable for general-purpose use and often perform poorly.

- Simple sum of raw or z-normalized scores is not generally appropriate, but may be effective for combining multiple samples or instances from one matcher.
- Examples of other such techniques include density estimation based on a normal (Gaussian) distribution assumption (or any parametric distribution).

7 References

[Ben-Yacoub] Ben-Yacoub, Abdeljaoued, Mayoraz; "Fusion of Face and Speech Data for Person Identity Verification"; 1999.

[BioAPI] BioAPI Consortium; BioAPI Specification Version 1.1; March 16, 2001.

[Bolle-00b] R. Bolle, N. Ratha, S. Pankanti; "Evaluating authentication systems using bootstrap confidence intervals"; *Proceedings of 15th International Conference On Pattern Recognition*; pp. 835-841; 2000.

[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"; *Proc. 4th IAPR Intl. Conf. on Audio- and Video-based Biometric Person Authentication, AVBPA*, Springer LNCS-2688, pp. 830-837; 2003.

[FIF] Fusion Information Format for Data Interchange; September 12, 2005

[Griffin-04b] P. Griffin, U. Mahlmeister, Fusion for BioAPI Modules, August 2004

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

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

[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-05] Jain, Nandakumar, Ross; "Score Normalization in Multimodal Biometric Systems"; *Pattern Recognition 38* (2005); p2270-2285.

[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 1999.

[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

[Korves-05] H. Korves, L. Nadel, B. Ulery, D. Masi; "Multi-Biometric Fusion: From Research to Operations"; *Sigma*, Summer 2005

[Macskassy-04] Macskassy and Provost; "Confidence Bands for ROC Curves: Methods and an Empirical Study"; 2004.

[Micheals-03] R. Micheals; "Biometric Systems Evaluation"; 20 June 2003. (PhD Thesis)

[Neyman-33] J. Neyman, 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.

[Poh05d] N. Poh, S. Bengio; "Towards Explaining the Success (Or Failure) of Fusion in Biometric Authentication"; Institut Dalle Molle d'Intelligence Artificielle Perceptive, IDIAP-RR 05-43, 2005

[Ross-01] A. Ross, A.K. Jain; "Information Fusion in Biometrics"; *Proc. of 3rd Int'l Conference on Audio- and Video-Based Person Authentication (AVBPA)*, pp. 354-359, June 6-8, 2001.

[Scott-05] C. Scott, R. Nowak; "A Neyman-Pearson Approach to Statistical Learning"; 2005 (http://www.ece.wisc.edu/~nowak/np.pdf)

[Scott-06] C. Scott; ``Comparison and design of Neyman-Pearson classifiers'' (http://www.stat.rice.edu/~cscott/pubs/npdesign.pdf)

[SDK2] C. Watson, C. Wilson, M. Indovina, B. Cochran; "Two Finger Matching With Vendor SDK Matchers"; *NIST Interagency Report 7249*; July 2005.

[Sedgwick-03a] N. Sedgwick; "Overview of Issues on Standardisation of Multi-Modal Biometric Combiners"; Cambridge Algorithmica Limited; 16th October 2003;

[Sedgwick-03b] N. Sedgwick; "The Need for Standardisation of Multi-Modal Biometric Combination"; Cambridge Algorithmica Limited; Briefing for BSI Committee IS T/44; 6 Nov 2003;

[Sedgwick-05] N. Sedgwick; "Preliminary Report on Development and Evaluation of Multi-Biometric Fusion using the NIST BSSR1 517-Subject Dataset"; 28 May 2005

[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.

[Verlinde-00] P Verlinde, G Chollet, and M Acheroy (2000) "Multi-modal Identity Verification Using Expert Fusion", *Information Fusion 1*, 17-33.

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

[Wayman] J. L. Wayman; "Confidence Interval and Test Size Estimation for Biometric Data"; *National Biometric Test Center Collected Works 1997-2000*