

When Data Goes Missing: Methods for Missing Score Imputation in Biometric Fusion

Yaohui Ding and Arun Ross

Lane Department of Computer Science & Electrical Engineering, West Virginia University,
Morgantown, WV, USA

ABSTRACT

While fusion can be accomplished at multiple levels in a multibiometric system, score level fusion is commonly used as it offers a good trade-off between fusion complexity and data availability. However, missing scores affect the implementation of several biometric fusion rules. While there are several techniques for handling missing data, the imputation scheme - which replaces missing values with predicted values - is preferred since this scheme can be followed by a standard fusion scheme designed for complete data. This paper compares the performance of three imputation methods: Imputation via Maximum Likelihood Estimation (MLE), Multiple Imputation (MI) and Random Draw Imputation through Gaussian Mixture Model estimation (RD GMM). A novel method called Hot-deck GMM is also introduced and exhibits markedly better performance than the other methods because of its ability to preserve the local structure of the score distribution. Experiments on the MSU dataset indicate the robustness of the schemes in handling missing scores at various missing data rates.

Keywords: Biometric fusion, Missing data, Imputation, Hot-deck GMM

1. INTRODUCTION

Biometric systems that consolidate multiple sources of biometric information of the same identity are known as multi-biometric systems.¹ This consolidation of multiple traits, multiple units (of the same biometric trait), multiple classifiers or multiple samples, could be implemented at various levels, viz.: image level, feature level, rank level, score level or decision level. Fusion at the score level is the most common approach due to the trade-off between information availability and fusion complexity.²

Most techniques for score level fusion are designed for a complete score vector* where the scores to be fused are assumed to be available. These methods cannot be invoked when score vectors are incomplete. Missing information may be due to several reasons: (a) failure of a matcher to generate a score (e.g., a fingerprint matcher may be unable to generate a score when the input image is of inferior quality); (b) absence of a trait during image acquisition (e.g., a surveillance multibiometric system may be unable to obtain the iris of an individual); or (c) sensor malfunction, where the sensor pertaining to a modality may not be operational (e.g., failure of a fingerprint sensor due to wear and tear of the device). Deletion methods, which omit all incomplete vectors, will not be suitable in such cases.³ Imputation methods, on the other hand, which substitute the missing scores with predicted values are a better solution as (a) they do not delete any of the score vectors which may contain useful information for identification, and (b) their application could be followed by a standard score fusion scheme.

Many imputation methods are widely known. The shortcomings of some simple single-imputation methods like Mean Imputation, Median Imputation and Regression Substitution have been discussed in the literature.^{4,5} The more advanced imputation methods can be divided into three categories based on the model of data they assume: parametric methods, non-parametric methods and semi-parametric methods. Two popular methods, Maximum-likelihood Estimation (MLE) and Multiple Imputation (MI) via Data Augmentation⁶⁻⁸, assume a multivariate normal parametric model for the complete dataset. Zio et al.⁹ compare two semi-parametric methods - the Random Draw using GMM estimation (RD GMM) and Conditional Mean using GMM estimation (CM GMM) - and one non-parametric method called the Nearest Neighbor

Yaohui Ding: E-mail: yding@mix.wvu.edu, Telephone: 1 662 202 5569 (Contact Author)

Arun Ross: E-mail: arun.ross@mail.wvu.edu, Telephone: 1 304 293 9135

*Here, the elements of the vector are the scores generated by the individual matchers

Donor (NND). After applying the methods to both synthetic and real data, they observed that RD GMM is better at preserving both sample mean and covariance unlike CM GMM and NND.

The missing score problem in biometric fusion has received limited attention. A Bayesian approach utilizing both ranks and scores to perform fusion in an identification system was proposed by Nandakumar et al.¹⁰ which handled missing information by assigning a fixed rank value to the marginal likelihood ratio corresponding to the missing entity. Fatukasi et al.³ implemented the k-nearest neighbor (k-NN) imputation method and described three different variants for use in biometric fusion. An empirical comparison of some imputation methods such as Mean Imputation, Median Imputation and Regression Substitution was also done.

In this paper, we analyze two parametric methods (MLE and MI) and one semi-parametric method (RD GMM) for handling the missing data problem. The MSU database² containing match scores of three modalities (face, fingerprint and handgeometry) is used in the experiments. Furthermore, we introduce a novel imputation algorithm called the Hot-deck via GMM estimation, which could be viewed as an extension of the RD GMM imputation scheme.

The remainder of this paper is organized as follows. Section 2 discusses the patterns of missing data and the criteria for designing imputation methods. Algorithms employed in this work for handling the missing score problem are described in Section 3. The performance of different methods under various training and missing data parameters are reported in Section 4, and the ensuing results are discussed in Section 5. A summary is presented in Section 6.

2. PATTERNS OF MISSING DATA

Distinguishing between different patterns of missing data is important because it will determine the method used for handling the problem. Rubin⁶ defined a taxonomy for three patterns of missing data. In ‘Missing Completely At Random (MCAR)’, the probability that an entry will be missing is independent of both observed and unobserved values in the dataset. In ‘Missing At Random (MAR)’, the probability that an entry will be missing is a function of the observed values in the dataset. In ‘Missing Not At Random (MNAR)’, the missing entry depends on the observed data as well as on the value of the data which is missing. It should be noted that methods for distinguishing between MCAR and MAR are computationally expensive. Ramoni and Sebastiani¹¹ describe a novel method called the Robust Bayesian estimator (RBE), which bounds the set of all missing-value estimates consistent with the data using probability intervals, even if there is no information about the pattern of missing data. Since determining the reason behind missing information may not always be possible, it cannot be guaranteed that the occurrence of a missing observation is truly random. This makes MAR a much milder assumption (compared to MCAR) thereby allowing the synthesis of a missing-value dataset from a complete dataset.

As stated by Marker et al.,¹² two main criteria should be employed in assessing the performance of imputation methods: (a) a good imputation should preserve the natural relationship between variables in a multivariate dataset (in our case, the variables correspond to scores originating from multiple classifiers); (b) a good imputation method should embody the uncertainty caused by the imputed data by deriving variance estimates. These two criteria are applicable for imputation in a biometric score dataset. Additionally, the use of imputed data should result in comparable performance as that of raw data containing no missing observations. Some imputation methods may not result in good performance if they overstate or understate the relationship between variables, or if they omit the uncertainty in the imputed data.

3. IMPUTATION METHODS

3.1 Notation

Let $\mathbf{D} = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_N\}^t$ denote the dataset with N observations where each \mathbf{X}_i with $i = 1, \dots, N$ is a vector of scores from K different modalities (or classifiers). Different multivariate models will be assumed in the methods considered below. For example, in the MLE and MI methods, the dataset \mathbf{D} will be assumed to have a K -variate Gaussian distribution with mean $\boldsymbol{\mu} = (\mu_1, \dots, \mu_K)$ and covariance matrix $\Sigma = (\sigma_{ik})$ (with $k = 1, \dots, K$). Since both MLE and GMM methods use an iterative algorithm for estimation, let $\Theta^{(t)}$ denote all the parameters to be estimated at the t -th iteration. Note that $\Theta = (\boldsymbol{\mu}, \Sigma)$ corresponds to the multivariate Gaussian assumption.

The training set \mathbf{D}^{tr} of score vectors does not have any incomplete observations. The test set \mathbf{D}^{te} contains both complete and incomplete observations. An incomplete vector missing a score from the second classifier can be written as $\mathbf{X}_i = \{x_{i1}^{obs}, x_{i2}^{mis}, \dots, x_{iK}^{obs}\} \equiv \{X_i^{obs}, X_i^{mis}\}$. Here, X_i^{obs} denotes the scores that are available while X_i^{mis} denotes the scores that are missing.

3.2 Imputation through MLE

The theoretical benefits of Maximum-likelihood Estimation (MLE) are widely known. After incorporating the Expectation Maximization (EM) algorithm, the MLE via EM method can be used to handle the problem of parameter estimation in an incomplete dataset. The standard MLE via EM algorithm applies a two-step iterative procedure to estimate the final parameters:

- E-STEP: Replace missing scores with the conditional expectation of the missing data given the current estimate of mean and covariance matrix of the training data; the sweep operator⁴ is applied to get regression equations with X^{obs} as predictors.
- M-STEP: Obtain the new ML estimates of the mean vector and covariance matrix using both the observed data and the imputed data from the E-STEP.

When this method is used in the biometric scenario involving match scores, some additional constraints are required: the different observations (vectors) should be considered independent, and this assumption should be maintained as much as possible during the estimation and imputation procedure. In order to accommodate this assumption, while computing the distribution parameters and performing imputation for the incomplete vector $\mathbf{X}_i = (\mathbf{X}_i^{obs}, \mathbf{X}_i^{mis})$, we only use the training set \mathbf{D}^{tr} and the observed part of this vector X_i^{obs} . This strategy is relevant in a practical situation where the training set is relatively fixed.

With the assumption of K-variate normal distribution, the hypothetical complete dataset \mathbf{D} belongs to the regular exponential family. So $\sum_{i=1}^n x_{ik}$ and $\sum_{i=1}^n x_{ik}x_{ij}$ are sufficient statistics of samples from this distribution ($j, k = 1, \dots, K$). The modified t -th iteration of E-STEP can then be written as:

$$E\left(\sum_{i=1}^n x_{ik} | \mathbf{D}^{tr}, X_i^{obs}, \Theta^{(t)}\right) = \sum_{i=1}^n x_{ik}^{(t)}, \quad k = 1, \dots, K \quad (1)$$

$$E\left(\sum_{i=1}^n x_{ik}x_{ij} | \mathbf{D}^{tr}, X_i^{obs}, \Theta^{(t)}\right) = \sum_{i=1}^n \left(x_{ik}^{(t)}x_{ij}^{(t)} + c_{ijk}^{(t)}\right) \quad (2)$$

where,

$$x_{ik}^{(t)} = \begin{cases} x_{ik}, & \text{if } x_{ik} \text{ is observed} \\ E\left(x_{ik} | \mathbf{D}^{tr}, X_i^{obs}, \Theta^{(t)}\right), & \text{if } x_{ik} \text{ is missing.} \end{cases} \quad (3)$$

and

$$c_{ijk}^{(t)} = \begin{cases} 0, & \text{if } x_{ik} \text{ or } x_{ij} \text{ is observed} \\ Cov\left(x_{ik}, x_{ij} | \mathbf{D}^{tr}, X_i^{obs}, \theta^{(t)}\right), & \text{if } x_{ik} \text{ and } x_{ij} \text{ are missing.} \end{cases} \quad (4)$$

The M-STEP of the EM algorithm is straightforward and is a standard MLE process, i.e.,

$$\mu_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^n x_{ik}^{(t)}, \quad k = 1, \dots, K; \quad (5)$$

$$\sigma_{jk}^{(t+1)} = \frac{1}{n} E\left(\sum_{i=1}^n x_{ik}x_{ij} | \mathbf{D}^{tr}, X_i^{obs}\right) - \mu_k^{(t+1)}\mu_j^{(t+1)}. \quad (6)$$

The algorithm will iterate repeatedly between the two steps until the difference between covariance matrices in subsequent M-STEPs falls below some specified convergence criterion. Then the imputed values are calculated by performing the E-STEP one more time.

A notable drawback of EM algorithm should be pointed out. The imputed scores from the EM algorithm lack the residual variability which are present in the training set with complete data, because they fall exactly on the regression line when using the parameters estimated by the iterations. The MI method described below will impart variability and uncertainty to these imputation values.

3.3 Multiple Imputation Method

The primary shortcoming associated with the MLE method - its inability to accommodate variability/uncertainty - can be attenuated by the Multiple Imputation (MI) method. Proposed by Rubin,⁶ the MI method accounts for missing data by restoring not only the natural variability in the missing-data, but by also incorporating the uncertainty caused by the estimation process.

Although several variants of Multiple Imputation are available, the NORM package proposed by Schafer⁸ which employs the Data Augmentation algorithm, exhibits good performance. Data Augmentation (DA) is an iterative simulation technique that bears a strong resemblance to the EM algorithm. In the DA process, missing data are first imputed by drawing them from their conditional distribution given the observed data and assumed values for the parameters (I-Step). Next, a set of *multiple* values for the parameters is drawn from the Bayesian posterior distribution given the observed data and the most recently imputed values for the missing data (P-Step). Alternating between these two steps continues until convergence produces multiple imputations (say, m) of the missing data.

The converged parameter estimates corresponding to the m imputations is stored in this case. Then the following analysis is conducted in order to obtain an estimate of the imputed data : a) compute the point estimate of each parameter; b) compute the estimate of the standard error for each parameter; and c) obtain the critical ratio (point estimate divided by the estimated standard error). After that, Rubin's Rules⁶ are used to combine the results of this analysis to derive an overall estimate. A reasonable point estimate of the parameter is the simple average of the m estimates. The estimate of the standard error is a combination of the within-imputation variability, $\bar{\mu}$, and the between-imputation variability, B :

$$T = \bar{\mu} + [(1 + 1/m) * B].$$

3.4 Imputation via GMM Estimation

Imputation method via Gaussian Mixture Model (GMM) estimation contains two parts: density estimation using the assumption of a GMM and imputation using the estimated density.

The density estimation problem can be stated as follows. Given our dataset \mathbf{D} , which can be interpreted as N points in K dimensions, and a family \mathbf{F} of probability density functions, find the probability density $f(\mathbf{x}) \in \mathbf{F}$ that is most likely to have generated the given points.

Gaussian Mixture Model defines the family \mathbf{F} by assuming that each member has the same mathematical form - a multivariate Gaussian distribution - and using different $\Theta = (p_1, \dots, p_C; \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_C; \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_C)$ which denotes all sets of parameters in the model to distinguish different members, i.e.,

$$f(\mathbf{x}; \Theta) = \sum_{c=1}^C p_c N(\mathbf{x}; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c), \quad (7)$$

where $\sum p_c = 1, p_c \geq 0$ for $c = 1, \dots, C$. Let us introduce the vector of indicator variables $\mathbf{z}_i = (z_{i1}, \dots, z_{iC})$ where $z_{ic} = 1$ if observation i belongs to c -th multivariate Gaussian distribution, and 0 otherwise.

Hunt and Jorgensen¹³ designed a GMM Estimation algorithm based on the maximum-likelihood estimates via EM algorithm.

3.4.1 Density Estimation of GMM

Although a mixture model has great flexibility in modeling, a restriction of the number of components C is still required, because along with an increase in the number of parameters, the estimation of those parameters from the training data might imply a greater variance for each of the parameters. In our work, the Bayesian Information Criterion (BIC) is used as the statistical criterion to obtain a reasonable value of C .

The Expectation (E-STEP) and the Maximization (M-STEP) are described as follows:

- E-STEP:

$$\begin{aligned}\hat{\tau}_{ic}^{(t)} &= E\left(z_{ic} | \mathbf{D}^{\text{tr}}; \Theta^{(t)}\right) \\ &= \frac{p_c^{(t)} N(\mathbf{D}^{\text{tr}}; \boldsymbol{\mu}_c^{(t)}, \boldsymbol{\Sigma}_c^{(t)})}{\sum_{c=1}^C p_c^{(t)} N(\mathbf{D}^{\text{tr}}; \boldsymbol{\mu}_c^{(t)}, \boldsymbol{\Sigma}_c^{(t)})}.\end{aligned}\quad (8)$$

Here, the first expected value $\hat{\tau}_{ic}^{(t)}$ is the estimated posterior probability (at iteration t) that the i -th observation belongs to the c -th group.

- M-STEP: The mixing proportions are given by:

$$p_c^{(t+1)} = \frac{1}{n} \sum_{i=1}^n \hat{\tau}_{ic}^{(t)}, \quad c = 1, \dots, C \quad (9)$$

The multivariate Gaussian parameters are obtained as:

$$\boldsymbol{\mu}_{ck}^{(t+1)} = \frac{1}{np_c^{(t+1)}} E\left(\sum_{i=1}^n \hat{\tau}_{ic}^{(t)} x_{ik} | \mathbf{D}^{\text{tr}}; \boldsymbol{\mu}_c^{(t)}, \boldsymbol{\Sigma}_c^{(t)}\right), \quad k = 1, \dots, K \quad (10)$$

$$\begin{aligned}\boldsymbol{\Sigma}_{cjk}^{(t+1)} &= \frac{1}{np_c^{(t+1)}} E\left(\sum_{i=1}^n \hat{\tau}_{ic}^{(t)} x_{ik} x_{ij} | \mathbf{D}^{\text{tr}}; \boldsymbol{\mu}_c^{(t)}, \boldsymbol{\Sigma}_c^{(t)}\right) \\ &\quad - \boldsymbol{\mu}_{ck}^{(t+1)} \boldsymbol{\mu}_{cj}^{(t+1)}, \quad j, k = 1, \dots, K\end{aligned}\quad (11)$$

3.4.2 Two Imputation Methods via GMM Estimation

- **Random Draw Imputation:** The estimates of the Gaussian mixture model parameters are obtained as:

$$f(X_i; \Theta) = \sum_{c=1}^C p_c N(X_i; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c), \quad (12)$$

In practice, the random drawing of a value as x_i^{mis} from the distribution of

$$f(X_i^{\text{mis}} | X_i^{\text{obs}}; \Theta) = \sum_{c=1}^C \tau_{ic} N(X_i^{\text{mis}} | X_i^{\text{obs}}; \Theta), \quad (13)$$

could be accomplished by two simple steps: First, draw a value c from the distribution $\text{Multinomial}(1; \tau_{1c}, \dots, \tau_{ic})$; Then, given c , generate a random value from the multivariate conditional Gaussian distribution as the imputation of the missing value $N(X_i^{\text{mis}} | X_i^{\text{obs}}; \Theta)$.

- **Hot-deck Imputation:** The main principle of the hot deck method is to use the current set of existing scores or simulated scores (donors) to provide imputation values for the incomplete vectors. In our experiment, a simulated dataset based on the estimation of the mixture model parameters was used as the donor dataset.

The procedure can be described in two steps: Firstly, use the estimation of mixture model parameters Θ to simulate a dataset $\mathbf{D}^{\text{sim}} = \{\mathbf{Y}_1, \dots, \mathbf{Y}_M\}^T, M \gg n$. Then, in order to impute values to an incomplete vector \mathbf{X}_i , use the observed part X_i^{obs} to find the simulated vector \mathbf{Y}_m in \mathbf{D}^{sim} , which has the smallest Euclidean distance with X_i^{obs} , and then replace the missing part X_i^{mis} with the exact value of the corresponding vector \mathbf{Y}_m . For instance, if $\mathbf{X}_i = (X_{i1}^{\text{obs}}, X_{i2}^{\text{mis}}, \dots, X_{iK}^{\text{obs}})$ (implying that the score of the second classifier is missing), the missing score will be imputed by the score of the second feature in \mathbf{Y}_m , $X_{i,2}^{\text{imp}} = Y_{m,2}$.

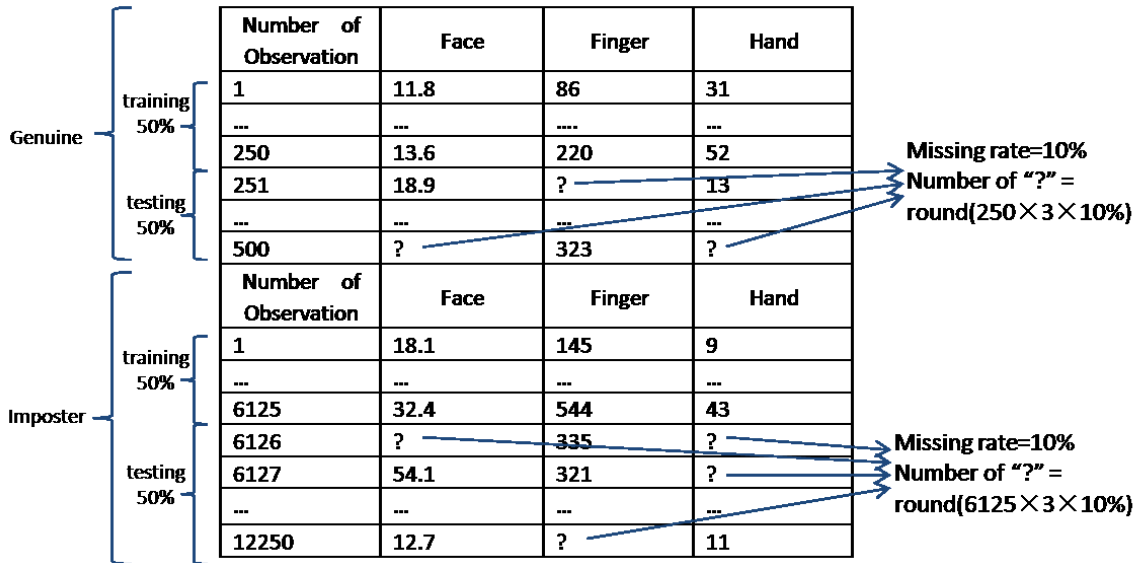


Figure 1. Example of the structure of the database used in our experiments. Here, 50% of the dataset is used as training data, and a missing rate of 10% is specified for the test set.

4. EXPERIMENTS

Experiments were conducted using the Michigan State University (MSU) dataset consisting of face, fingerprint and hand-geometry information pertaining to 50 users with each user providing 5 samples of each biometric. Each modality includes 500 genuine scores and 12,250 imposter scores.² The column vector $\{X_1, X_2, X_3\}$ corresponds to the (similarity) scores of face, fingerprint and hand-geometry, respectively.

The size of the training set is varied as three different proportions of the complete score database: 10%, 25% and 50%, respectively, and each training set is used by the four imputation methods. It should be noted that the training set is always completely observed without any missing data, i.e., it consists of complete score vectors. The remaining part of the database (90%, 75% and 50%, respectively) is used for testing and evaluating the models.

For each modality in the test set, we randomly remove the matching scores at three different rates: 5%, 10% and 50%, in order to artificially generate missing score vectors. At the same time, we ensure that there is at least one observed score available for every observation (vector). Since missing data is simulated at different rates on a modality-by-modality basis, the total number of observations (score vectors) containing missing data can be greater than the product of the size of the test set and the specified missing rate. Figure 1 gives an example of the structure of the database used in the experiment. It should be noted that the training subsets corresponding to both genuine and imposter score vectors are combined into a single training set when estimating the parameters.

As a result, nine combinations encompassing different training and missing data rates are generated. Although the above procedure to generate missing data is completely random and the artificial database appears to conform to the MCAR scenario, the MAR assumption may still be tenable since in real-world operational datasets there could be potential reasons for the missing part to depend on the observed part.

Several auxiliary software were used to obtain the imputed values in this work. NORM is a free program distributed by Schafer(1997), and it has been employed for ML estimation and Multiple Imputation, both of which assume the multivariate normal model. Gaussian Mixture Model estimations are obtained using the R software with the “mclust” package, Version 3.1-10 (2008), for Windows.

5. RESULTS

The min-max normalization scheme followed by the simple Sum Rule fusion has been observed to result in reasonable improvement in matching accuracy of a multibiometric system.² This technique was used in this work for reporting

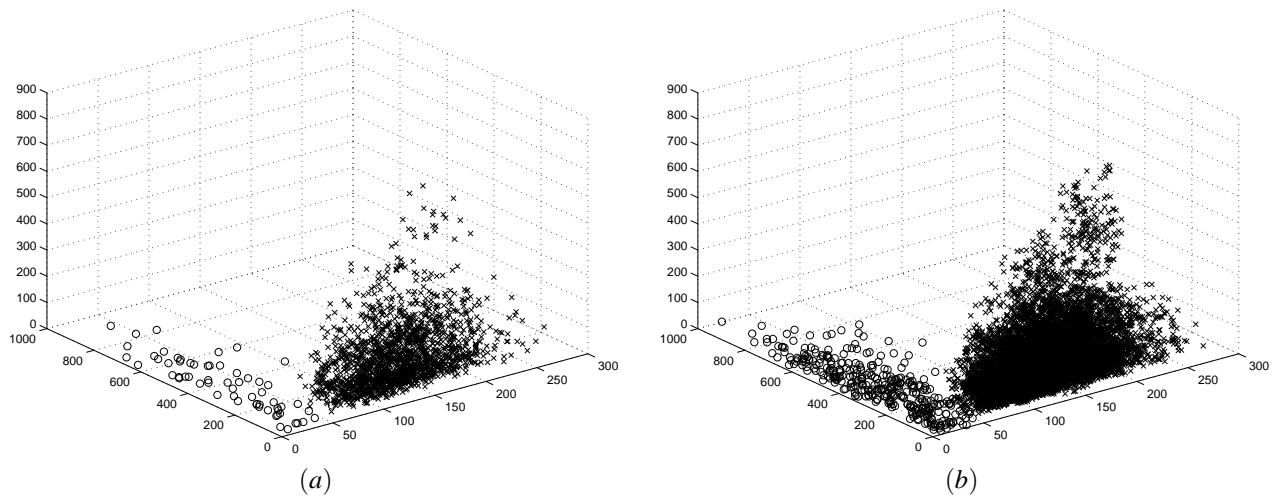


Figure 2. Scatter plots of the test set ('o': Genuine and 'x': Imposter) before artificially removing some points. (a) When 50% of the MSU dataset is employed for testing; (b) When 75% of the MSU dataset is employed for testing

matching performance. Also, in the various graphs, the legend “Raw Fingerprint/Face/HandGeometry” is used to indicate the performance when there is no missing data and the legend “Imputed Fingerprint/Face/HandGeometry” is used to denote performance after application of one of the imputation methods.

The following are some comments on the recognition performance of different imputation methods:

- From Fig.5 and Fig.6, we observe that imputation through MLE performs the worst among the four methods, and the MI method which uses the same model (multivariate normal distribution) and the same estimated parameters performs much better. From these figures we even notice that the MLE imputation method results in poor fusion performance that is worse than a single modality system (fingerprint) at a low FAR ($\sim 0.01\%$). This can also be seen in Table 2. When the rate of missing data reaches 50% (Mi50), under the MLE method, the GAR (Genuine Accept Rate) of Sum Fusion decreases to that of fingerprints at a FAR of 0.1%. This does not occur with the other methods used in our experiments.
- From Fig.5 and Table 1, we can observe that the MI method provides a comparable performance with the GMM assumption when the missing rate is low ($\leq 10\%$). When the missing rate increases to 50%, although the fusion performance due to the MI method is rather high (the EER of sum fusion is below 0.1% in this case), from Fig.6b we notice that the performances of the single modalities depart significantly from that of the raw dataset, which means the imputed data perturbs the natural characteristics of the original raw dataset. On the other hand, the Hot-deck GMM method delivers a low EER for fusion whilst simultaneously retaining the original shape of the ROC curves of the individual modalities.
- From both the EER and the GAR tables, we note that the fingerprint modality is sufficiently robust to various missing rates and different imputation methods. Therefore, assigning a comparatively high weight parameter to the similarity scores of the fingerprint modality during fusion will increase the robustness of handling missing data for this dataset.
- As a complement to the Random Draw via GMM method, the Hot-deck via GMM method performs much better under a higher missing rate. From Table 2 it can be found that the GARs of the face and handgeometry modalities decrease steeply when the size of the training set is small. This might be due to the nuances of the imputation process. The value c which is drawn from $Mult_c(1; \hat{\tau}_1, \dots, \hat{\tau}_K)$ plays a critical role in the imputed data, since it decides which component of the mixture model should be used to generate the imputation. A slightly biased value for c will cause an enormous deviation from the true component of the mixture model to which the raw score belongs. Therefore, if the size of training set is not large enough, the RD GMM method is more likely to generate a large bias from the raw data. In contrast, the Hot-deck via GMM method does not only rely on the estimated parameters, but also uses the information corresponding to the observed part of the incomplete vector to choose the nearest simulated data. It also requires the sample size of the simulated data to be large enough, and this requirement can be easily satisfied.

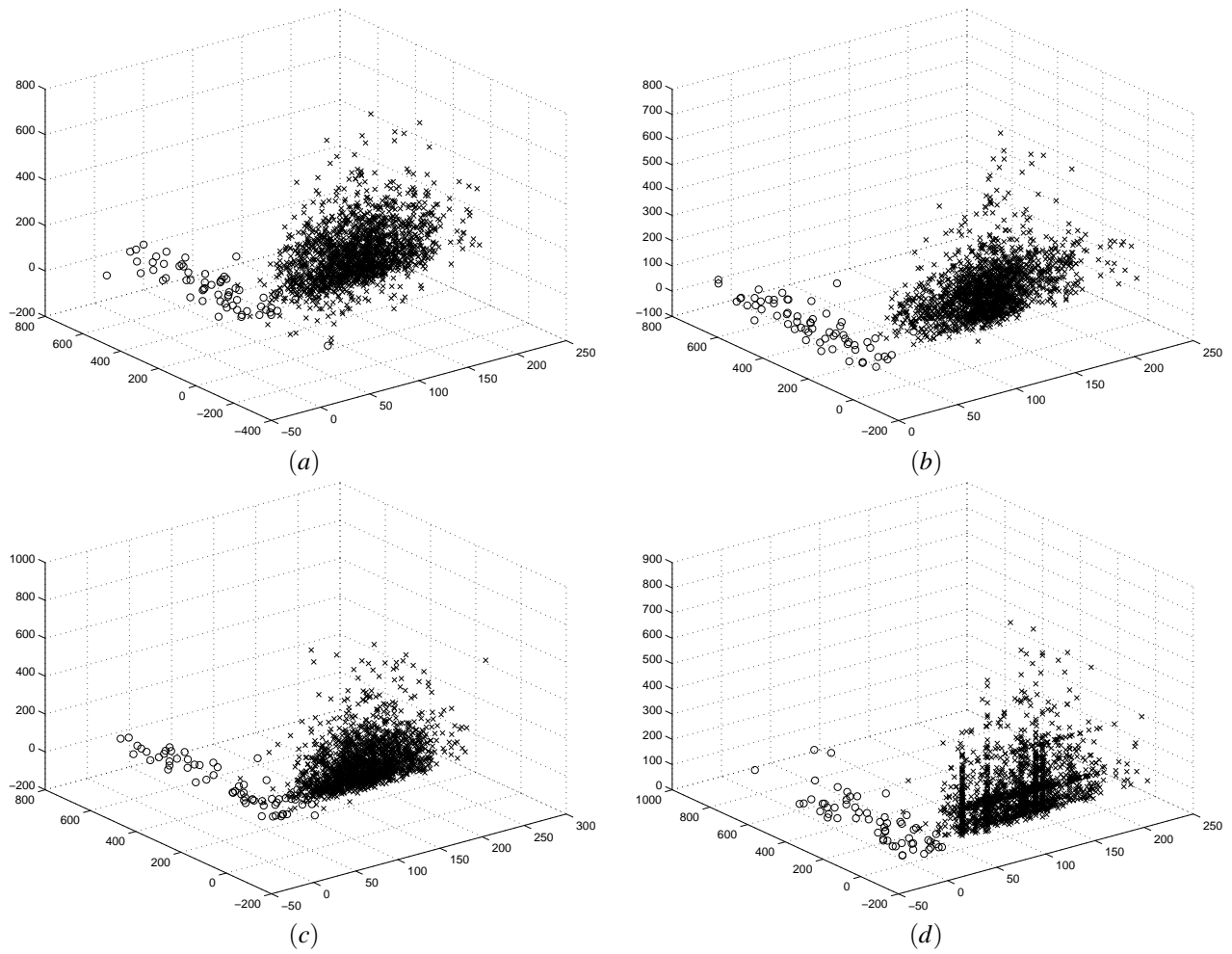


Figure 3. Scatter plots of the matching scores ('o': Genuine and 'x': Imposter) after imputation using different methods. Here, 50% of the dataset is employed as training data, and 10% of the test set is artificially missing: (a) MLE; (b) MI; (c) RD GMM; (d) Hot-deck GMM.

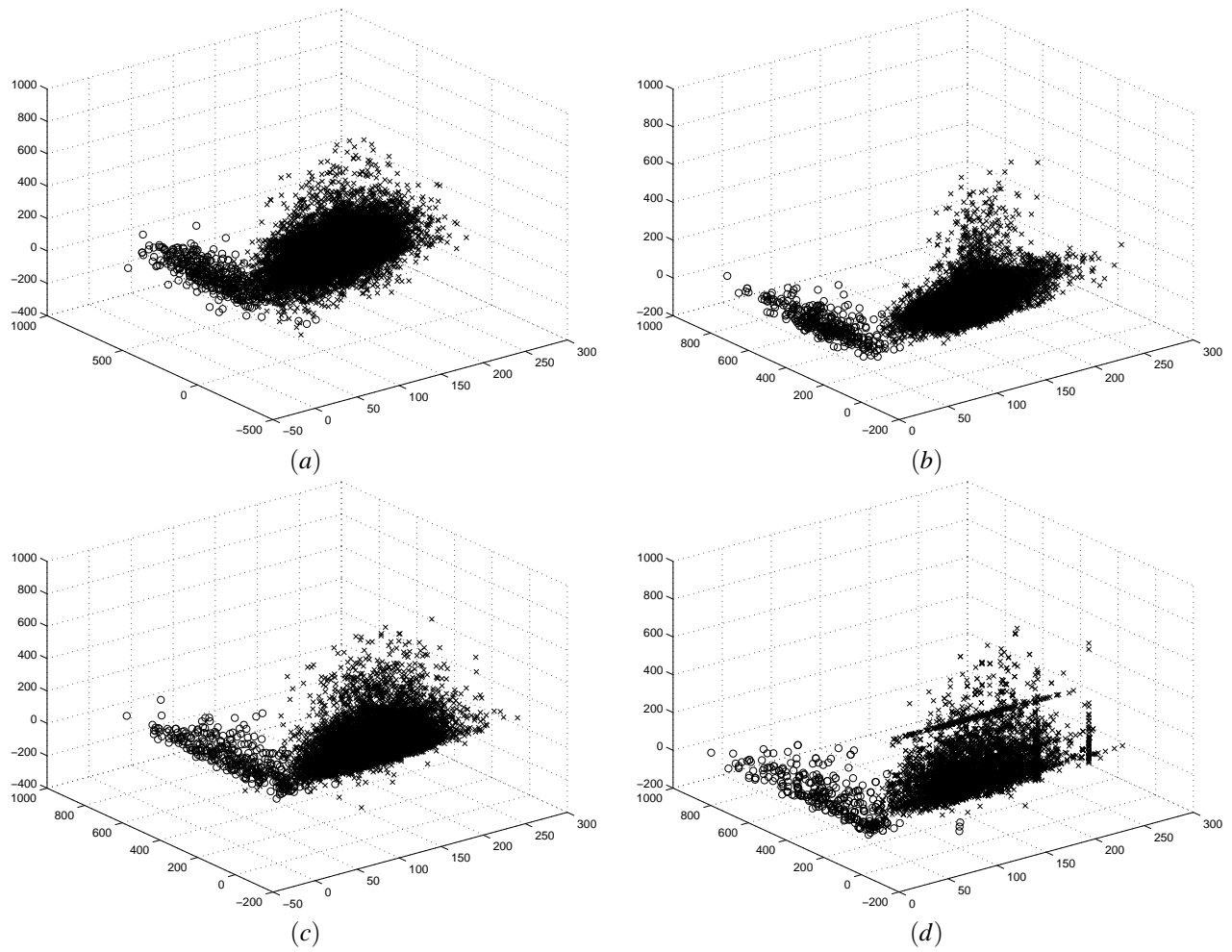


Figure 4. Scatter plots of the matching scores ('o': Genuine and 'x': Imposter) after imputation using different methods. Here, 25% of the dataset is employed as training data, and 50% of the test set is artificially missing: (a) MLE; (b) MI; (c) RD GMM; (d) Hot-deck GMM.

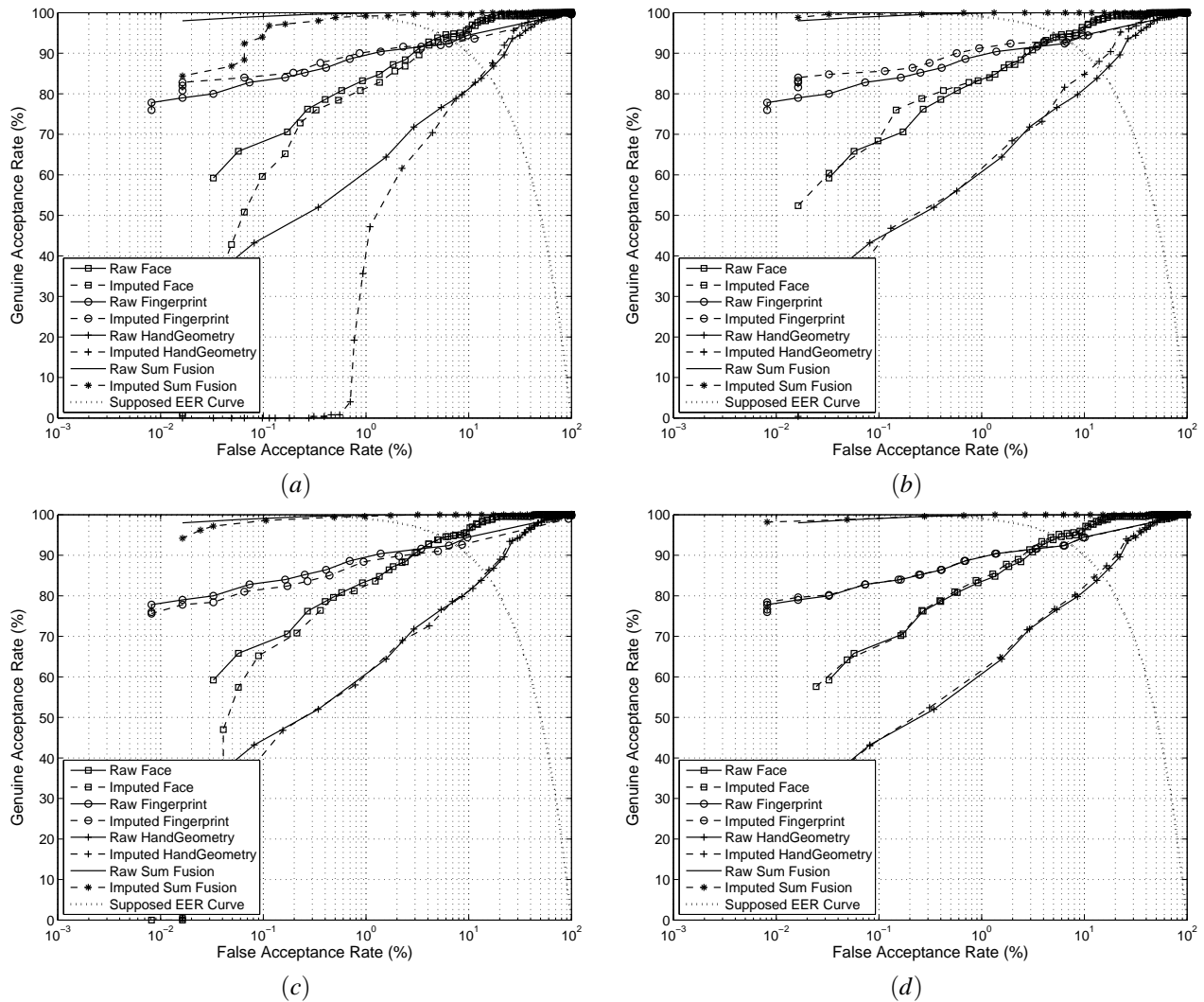


Figure 5. Recognition performance of different imputation methods where 50% of the dataset is employed as training data, and 10% of the test set is artificially missing: (a) MLE; (b) MI; (c) RD GMM; (d) Hot-deck GMM.

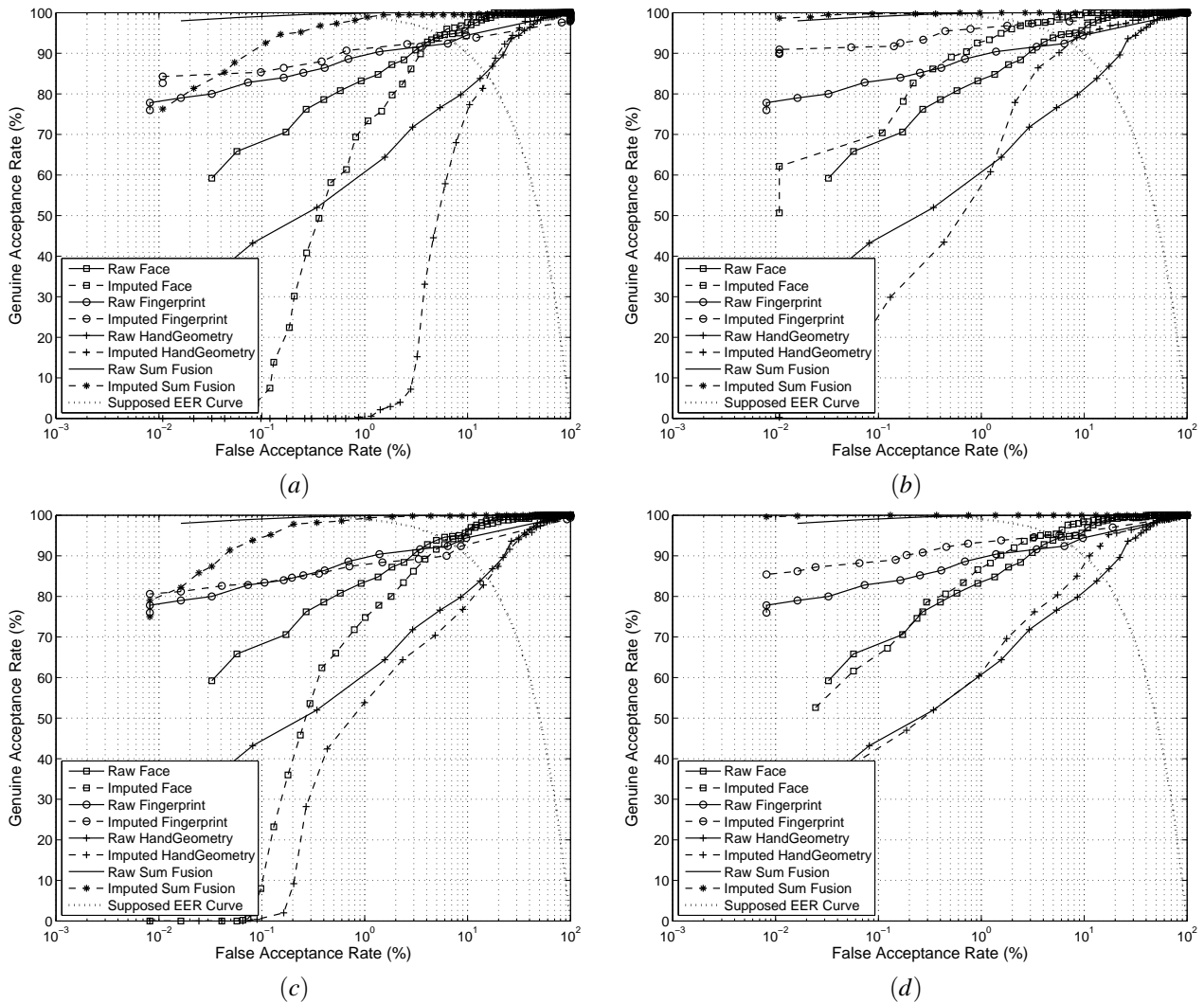


Figure 6. Recognition performance of different imputation methods where 25% of the dataset is employed as training data, and 50% of the test set is artificially missing: (a) MLE; (b) MI; (c) RD GMM; (d) Hot-deck GMM.

Table 1.

Equal Error Rate (EER) of different imputation methods. ‘Tr50’ means 50% of the dataset is used as training set, and ‘Mi5’ means 5% of the test set is artificially missing.

		Raw Data	Tr50			Tr25			Tr10		
			Mi5	Mi10	Mi50	Mi5	Mi10	Mi50	Mi5	Mi10	Mi50
MLE	Face	5.65	6.80	6.32	4.40	5.06	5.40	5.50	6.02	5.83	5.61
	Finger	7.11	6.12	7.34	10.12	5.78	5.57	6.76	7.65	7.10	6.69
	Hand	14.92	14.20	14.56	15.94	14.88	14.51	15.81	15.28	14.71	15.53
	Sum Fusion	0.36	0.31	0.90	1.99	0.58	0.48	1.07	0.60	0.57	1.40
MI	Face	5.65	6.85	5.99	3.18	5.02	4.73	2.77	5.43	5.31	2.99
	Finger	7.11	6.80	7.02	3.63	5.35	5.62	2.77	6.98	6.43	3.64
	Hand	14.92	13.53	12.84	8.32	13.28	13.01	7.41	14.59	13.35	7.46
	Sum Fusion	0.36	0.33	0.32	0.06	0.30	0.40	0.25	0.39	0.26	0.09
RD GMM	Face	5.65	5.64	5.75	4.80	5.68	5.48	6.48	5.73	6.65	6.99
	Finger	7.11	7.04	7.73	7.80	7.10	7.66	8.04	7.22	7.10	5.76
	Hand	14.92	14.68	14.56	15.05	14.82	15.25	15.81	14.56	13.97	13.81
	Sum Fusion	0.36	0.39	0.58	0.76	0.69	0.73	0.90	0.80	0.58	0.92
Hot-Deck GMM	Face	5.65	5.48	5.18	3.85	5.65	5.34	4.73	5.56	5.41	3.22
	Finger	7.11	6.72	7.11	5.66	6.97	7.11	4.96	6.94	6.50	3.98
	Hand	14.92	14.44	14.20	11.65	14.66	14.63	10.90	14.32	13.40	7.28
	Sum Fusion	0.36	0.37	0.36	0.29	0.36	0.34	0.08	0.28	0.25	0.06

6. SUMMARY

The results in the previous sections indicate that the imputation of missing data through the Gaussian Mixture Model is a powerful scheme for multimodal biometric fusion. Particularly, imputation via Hot-deck from the simulated dataset generated using the estimated Gaussian mixture model, results in a better recognition performance than the others. Imputation by randomly drawing from the estimated Gaussian mixture model is also a suitable option when it is difficult to generate a large simulated dataset. The imputation methods described in this paper focus only on recovering the missing similarity scores, but do not concern themselves with the subsequent processes in biometric fusion like score normalization and fusion rule. We are currently designing schemes in which the imputation and fusion can be performed simultaneously.

ACKNOWLEDGMENTS

This work was supported by the NSF Center for Identification Technology Research (CITeR).

REFERENCES

- [1] A. A. Ross, K. Nandakumar, and A. K. Jain, *Handbook of Multibiometrics*, Springer, Secaucus, NJ, USA, 2006.
- [2] A. Jain, K. Nandakumar, and A. Ross, “Score normalization in multimodal biometric systems,” *Pattern Recognition* **38**(12), pp. 2270–2285, 2005.
- [3] O. Fatukasi, J. Kittler, and N. Poh, “Estimation of missing values in multimodal biometric fusion,” in *IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS)*, 2008.
- [4] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, Wiley, New York, 1st ed., 1987.
- [5] J. Schafer and J. Graham, “Missing data: Our view of the state of the art,” *Psychological Methods*, 2002.
- [6] D. B. Rubin, *Multiple imputation for nonresponse in surveys*, Wiley, 1987.
- [7] R. J. A. Little, “Regression with missing x’s: A review,” *Journal of the American Statistical Association* **87**(420), pp. 1227–1237, 1992.
- [8] J. Schafer, *Analysis of incomplete multivariate data*, London, 1997.

Table 2.

Genuine acceptance rate (GAR)(%) of different imputation methods at a 0.1% false acceptance rate (FAR)

		Raw Data	Tr50			Tr25			Tr10		
			Mi5	Mi10	Mi50	Mi5	Mi10	Mi50	Mi5	Mi10	Mi50
MLE	Face	68.16	63.99	59.94	3.99	65.95	47.06	5.21	65.46	62.76	12.45
	Finger	83.25	86.07	84.36	85.21	85.34	84.97	85.46	82.64	83.74	85.09
	Hand	44.60	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06
	Sum Fusion	99.08	95.64	94.17	86.56	97.57	95.80	91.83	97.73	96.99	86.93
MI	Face	68.16	68.53	68.40	72.70	68.40	69.63	70.00	68.77	66.69	71.96
	Finger	83.25	86.56	85.46	93.68	86.07	85.71	91.47	83.37	85.09	91.10
	Hand	44.60	42.39	42.88	38.96	42.76	40.67	25.21	43.62	44.60	30.12
	Sum Fusion	99.08	99.45	99.57	99.94	99.20	99.08	99.69	99.08	99.08	99.82
RD GMM	Face	68.16	68.04	70.25	35.03	65.71	63.13	8.77	61.78	58.22	1.53
	Finger	83.25	83.74	82.76	82.52	83.25	83.13	83.13	83.25	85.09	88.40
	Hand	44.60	43.87	48.53	14.54	29.88	29.75	0.80	36.50	1.53	0.06
	Sum Fusion	99.08	99.08	98.96	97.85	98.59	98.10	94.42	97.85	97.48	95.80
Hot-Deck GMM	Face	68.16	68.16	67.79	69.39	67.30	67.79	65.46	68.28	66.56	65.71
	Finger	83.25	83.99	83.25	87.67	83.62	83.99	88.53	83.62	85.34	91.35
	Hand	44.60	43.87	44.36	42.39	42.64	42.02	42.52	41.66	43.62	35.40
	Sum Fusion	99.08	98.83	99.08	99.33	98.96	99.20	99.82	99.08	99.33	99.82

- [9] M. Di Zio, U. Guarnera, and O. Luzi, "Imputation through finite gaussian mixture models," *Computational Statistics & Data Analysis* **51**(11), pp. 5305–5316, 2007.
- [10] K. Nandakumar, A. K. Jain, and A. Ross, "Fusion in multibiometric identification systems: What about the missing data?," in *IEEE/IAPR International Conference on Biometrics*, pp. 743–752, 2009.
- [11] M. Ramoni and P. Sebastiani, "Robust learning with missing data," *Machine Learning* **45**(2), pp. 147–170, 2001.
- [12] D. A. Marker, D. R. Judkins, and M. Winglee, "Large-scale imputation for complex surveys," in *Survey Nonresponse*, John Wiley and Sons, 1999.
- [13] L. Hunt and M. Jorgensen, "Mixture model clustering for mixed data with missing information," *Computational Statistics & Data Analysis* **41**(3-4), pp. 429–440, 2003.