Addressing Missing Values in Kernel-based Multimodal Biometric Fusion using Neutral Point Substitution

Norman Poh, David Windridge, Vadim Mottl, Alexander Tatarchuk and Andrey Eliseyev

***Abstract*— In multimodal biometric information fusion, it is common to encounter missing modalities in which matching cannot be performed. As a result, at the match score level, this implies that scores will be missing. We address the multimodal fusion problem involving missing modalities (scores) using support vector machines with the Neutral Point Substitution (NPS) method. The approach starts by processing each modality using a kernel. When a modality is missing, at the kernel level, the missing modality is substituted by one that is unbiased with regards to the classification, called a neutral point. Critically, unlike conventional missing-data substitution methods, explicit calculation of neutral points may be omitted by virtue of their implicit incorporation within the SVM training framework. Experiments based on the publicly available Biosecure DS2 multimodal (scores) data set shows that the SVM-NPS approach achieves very good generalization performance compared to the sum rule fusion, especially with severe missing modalities.**

***Index Terms*— Multimodal biometrics, multiple classifiers system, biometric authentication, information fusion, missing features**

I. Introduction

A. Motivations

In order to improve confidence in verifying the identity of individuals seeking access to physical or virtual locations, both government and commercial organizations are implementing more secure personal identification systems. The challenge of creating a well-designed, highly secure and accurate personal identification system has always been a central goal in security businesses. This challenge can be responded to by the use of multimodal biometric systems [1], [2], [3] where both the security and performance levels can be further increased.

Although multimodal biometric fusion is well studied, as evidenced by [1] (and references herein), little attention has been focused on how to handle the case of missing biometric modalities, which results in missing features in the joint-score (output) space. A recent work reported in [4] as well as the first known multimodal benchmark evaluation [5] shows that the problem of missing features can indeed be handled. In fact, according to [5], given a carefully designed fusion mechanism, a multimodal system can degrade gracefully in performance with increasingly many missing features

There are several causes for the missing modalities, as listed below:

- **Temporary/permanent alteration of biometric traits**: Underlying biometric traits are living tissues that may change both over short (within days) and long period of time (years). For instance, a cough may temporarily change the vocal tracts of a person and this is likely to result in false rejection. Some drugs are known to permanently change the fingerprint minutiae. In either circumstances, the user of a biometric device, or the operator involved, may decide not to use the device.
- **Malfunctioning of a subset of biometric devices**: Biometric device may be worn over time.
- **Desire to increase the authentication throughput**: Last but not least, for some applications, e.g., entrance to a theme park, where access request is larger than expected, it may be sensible to increase the throughput of biometric authentication by reducing the number of biometric traits needed.

One can distinguish two types of incomplete data samples: those in training (i.e., during the classifier design stage) and those in testing (i.e, when the classifier is operational). In training, incomplete data samples can be discarded if the proportion of incomplete samples versus the complete ones is very small. During testing, however, one still has to classify incomplete data samples. In the problem of biometric authentication/identification, it can be assumed that the training data is complete but the testing data may be incomplete due to the above mentioned reasons.

B. Existing Solutions to Missing Features

In the pattern recognition literature, the problem of missing features can be handled in three ways:

- **Imputation**: replacing the missing features via mean substitution [6], [7], at the simplest level, or else via more complex methods (e.g. [8]) that take into account specifics of the distribution statistics and morphology.
- **Exhaustive fusion design**: designing a fusion classifier for all possible combinations of observable features (also reported in [6])
- **Naive Bayes fusion**: assuming independence in the feature (joint-score) space, e.g., [4]

None of the above solution is thoroughly satisfactory. For instance, by replacing the missing features with their corresponding expected values, the imputation approach could po-

Norman Poh and David Windridge are with CVSSP, University of Surrey, Guildford, GU2 7XH, Surrey, UK. E-mails: normanpoh@ieee.org, d.windridge@surrey.ac.uk, Vadim Mottl, Alexander Tatarchuk and Andrey Eliseyev are with the Computing Center of the Russian Academy of Sciences, Vavilov St. 40, Moscow, 119991, Russia

tentially bias towards one class or another. The dynamic fusion solution, on the other hand, requires an exhaustive design of all possible fusion problems defined by the observable joint-score subspace. If there are N biometric systems to combine, $2^N - N - 1$ fusion classifiers will be needed (subtracting the cases involving the empty set as well as a single feature). Hence, this is not a scalable approach. Finally, by working only with the marginal distributions, the Naive Bayes approach cannot handle possible correlation among the expert outputs. As will be evidenced in our experiments, in the fusion problem involving multiple fingerprints (see Figures 3 and 4, for instance), the match (genuine) scores among different fingers of the same subject, for the same as well as different (left versus right) hands, are actually correlated. This implies that the need to work with the joint-score space directly.

There are two categories of solutions to the missing feature problem, depending on the type of classifier framework: generative versus discriminative classifiers. For the generative classifier, an obvious approach is to model the joint-score distribution. Then, during testing, one can simply *marginalize* the joint-score distribution with respect to the missing features, in order to obtain the distribution marginals with only the observed features subspace. Inference via the Bayes rule (estimating posterior probabilities) or the log-likelihood ratio test then becomes straightforward. This solution was reported in [5].

[BEGIN ADDED TEXT] For discriminative classifiers, *marginalization* will make the learning the parameter (search) problem NP-complete [9]. An alternative solution that often leads to acceptable performance is to ignore or to skip parameters corresponding to missing information, during both inference and prediction. This approach was pursued in [9] for discriminative classifiers based on Bayesian network classifiers and logistic regression. However, this is not always possible and, in general, parameter-omission must be treated on a classifier-by-classifier basis. The purpose of this paper is to propose a possible implementation of this strategy using support vector machines (SVM), in a manner that is naturally congruent with the underlying Kernel-based discriminative strategy. **[END ADDED TEXT]**

C. Our Proposal

In this paper, we thus propose a discriminative classifier capable of dealing with missing features using a kernel-based SVM approach. The multimodal biometric fusion problem is formulated as one of combining multiple kernels, in which each kernel is designed for a particular biometric modality (such that normalization of each output may be required in order to handle the different range of each kernel prior to combining). Our particular problem is to combine multiple kernels when the experimental data is not fully represented in each kernel.

In order to handle the missing features during testing/inference, we will adopt the neutral point substitution (NPS) method [10]. An SVM works by projecting features into a linear kernel embedding space defined with respect to those same features. Missing modalities (i.e., features in the joint-score space) thus, in general, cannot be located within this space in order to be substituted. This difficulty arises because the symmetric, positive-definite Kernel matrix that specifies pairwise relations among all the training samples (and which can be regarded as a Gram matrix in the embedding space [11]) is undefined for the missing features/modalities. The NPS method adopts a decision-agnostic approach with regard to the substitution of these values, i.e. it assumes that missing modalities do not contribute to any bias in the discrimination of one class from another. Critically, unlike conventional missing-data substitution methods, the determination of these points can be implicitly incorporation within the SVM training framework.

In [10], it has been shown that the NPS method is theoretically equivalent to the sum rule fusion scheme when the modalities are maximally disjoint (i.e., there are no common samples). As justified by Kittler *et al.* [12], this finding means that the neutral-point method should exhibit a degree of resilience to class misattribution within the individual classifiers through the relative canceling of combined estimation errors (if sufficiently decorrelated). We would therefore like to quantify this result for a typical (i.e., non-maximally disjoint) data set.

D. Contributions and Paper Organization

The contributions of this paper are two-fold:

- to apply the NPS method, in an SVM framework, to solve the multimodal biometric fusion problem with missing modalities (features).
- to validate this approach using a publicly available database (for repeatability of experiments), i.e., the Biosecure DS2 score-and-quality database [13].

The experimental results show that while the sum fusion rule attains performance that is better than any single biometric modality (confirming to findings in the fusion literature [12], [14]), our proposal using the NPS method achieves even better results. We hypothesize that this will be typical for naturally-arising multi-kernel, missing-data problem such as multimodal biometrics.

This paper is organized as follows: Section II presents the theory of NPS. It will first introduce the notation for multimodal kernel design problem, and then present the neutral point method. Section III supports the presented theory by an experimental validation. This is followed by discussions and conclusions in Section IV.

II. Multimodal Kernel Decision and Neutral Point Substitution

A. The Multimodal Kernel Fusion Problem

Specifying a generalized feature map $\hat{\phi}$ to be that which generates an output in $\mathcal{R}^N$ for a detected object ω, we consider a multimodal kernel decision problem to be one in which feature maps are associated either with a set $\{S_m\}$ of m distinct sensor spaces; $\hat{\phi^m}(S_m(\omega)) \rightarrow \mathcal{R}^{N_m}$, or else associated with m distinct kernel measures $K_m(\hat{\phi}^m(\cdot), \hat{\phi}^m(\cdot)) \rightarrow \mathcal{R}$ defined on arbitrary (possibly even *common*) sensor-output spaces $\{S\}$. The former case, where distinct kernels are associated with distinct modalities, may be considered as representative of an *experimentally driven* scenario; the latter case is more typical of a ensemble-learning-driven scenario, in which multiple kernels are employed to capture different aspects of the learning problem. Since the latter case subsumes the former, for maximum generality we consider only this approach and omit explicit subscripting of the sensor space S.

When considered on a per-modality basis (i.e., applicable only to modality n), one cannot, in general, assume that a Kernel matrix $\mathbf{K_n} = K_n(\hat{\phi}^n(S(\omega_i)), \hat{\phi}^n(S(\omega_j)))$ will give rise to *identical* Mercer embedding spaces, $\hat{\psi}^n(S) = (\psi_1^n(S), \psi_2^n(S), \psi_3^n(S), \ldots)'$ when the set of objects from which i and j are drawn undergoes variation due to missing features/modalities. This is even more acute when the selected subsets have differing cardinalities, r, given the relation of this quantity to the embedding space dimensionality (the dimensionality of the space will always be$\leq r$ for inner-product kernels). This makes the substitution of, for instance, mean valued vectors for the missing values non-trivial, unlike the standard parametric missing value problem.

(Here, $\psi_i^n(S)$ are Eigenfunctions of the integral linear operator associated with Kernel K_n; i.e., such that $\hat{\phi}(S(\omega_i)) = \lambda^{\frac{1}{2}} u_i$, where $\mathbf{K_n} = \mathbf{U \Lambda U'}$ and $\mathbf{U} = (u_1, u_2, u_3, \ldots u_r)$, with $\mathbf{\Lambda} = diag(\lambda_1, \lambda_2, \ldots \lambda_n)$ the eigenvalue matrix, and $u_i = \psi_i(S(\omega_i))$ [i.e., the u_i are Mercer features]).

In the following section, we shall address the missing modality problem though SVM theory.

B. The Neutral Point Method

Let us assume an underlying unidimensional sensor space within each modality, and omit explicit consideration of the sensor-space/feature-map relation $\phi(S(\omega))$ as it does not effect findings:

We thus consider a set of Kernel measures, K_i in relation to which sensor outputs can be defined for each entity ω (i.e., where x maps objects ω into a common real-valued space):

$$\mathcal{X}_i = \{x(\omega), \omega \in \Omega\} \tag{1}$$

Any kernel $K_i(x_i', x_i'')$ embeds (via the inner product equivalence) the scale of the respective sensor $\mathcal{X}_i$ into a hypothetical linear space (the embedding space) in which the null element and linear operations are defined. If the Kernel is itself an inner product on the sensor outputs then $\hat{\mathcal{X}}_i \supseteq \mathcal{X}_i$: however, this relation does not hold for general kernels.

For a single modality, the training set:

$$\Omega_i^\star = \{\omega_j, j = 1, \ldots, N_i\} \tag{2}$$

is *completely* defined by kernel matrix and class indices y ($y = \pm 1$):

$$\Omega_i^\star => \{\mathbf{K}_i = \lfloor K_i(x_i(\omega_j), x_i(\omega_l)), \omega_j, \omega_l \in \Omega_i^\star \rfloor, y(\omega_j), \\ \omega_j \in \Omega_i^\star\}$$

Support Vector Machines (SVMs) are the most common Kernel-based approach to two-class pattern recognition, the problem being to find maximal margin discriminant hyperplane in space $\hat{\mathcal{X}}_i$:

$$\vec{y}_i(x_i(\omega)) = K_i(\theta_i, x_i(\omega)) + b_i \gtrless 0$$

(which generally has a much more complex (i.e non-linear) decision boundary in $\mathcal{X}_i$).

This leads to the standard SVM Training Criterion:

$$K_i(\theta_i, \theta_i) + C \sum_{\omega_j \in \Omega_i^\star} \delta_j \rightarrow \min(\theta_i \in \vec{\mathcal{X}}_i, b \in \mathcal{R}, \delta_j \in \mathcal{R})$$

Subject to:

$$y_i \lfloor K_i(\theta_i, x_i(\omega_l)) + b \rfloor \geq 1 - \delta_j, \delta_j \geq 0$$

The presence of the slack variables, δ_j, gives rise to a "softer" margin, allowing solutions to classification problems that are not linearly separable (δ_j measures the degree of misclassification of each object). These variables disappear in the dual formulation of the problem, leaving only C as a free configuration parameter for specifying the trade-off between margin maximization and error minimization. The (Wolfe) dual form of the criterion is hence a quadratic programming problem with respect to the Lagrangian multipliers, λ:

$$\sum_{\omega_j \in \Omega_i^\star} \lambda_{i,j} - \frac{1}{2} \sum_{\omega_j \in \Omega_i^\star} \sum_{\omega_l \in \Omega_i^\star} \lfloor y_j y_l K_i(x_i(\omega_j), x_i(\omega_l)) \rfloor \lambda_{i,j} \lambda_{i,l} \\ \rightarrow \max$$

Subject to:

$$\sum_{\omega_j \in \Omega_i^\star} y_j \lambda_{i,j} = 0, 0 \leq \lambda_{i,j} \leq C/2, \omega_j \in \Omega_i^\star$$

This gives rise to a decision rule defined by the support objects $\hat{\Omega}_i \in \Omega_i^\star$ as the remaining Lagrange multipliers tend to zero $\lambda_{i,j} \rightarrow 0$ (leaving $\hat{\lambda}_{i,j} > 0$):

$$\hat{f}(x_i(\omega)) = \sum_{j:\omega_j \in \Omega_i^\star} y_j \hat{\lambda}_{i,j} K_i(x_i(\omega_j), x_i(\omega_l)) + \hat{b}_i \gtrless 0$$

with:

$$\hat{b}_i = -\left(\frac{\sum_{j:\omega_j \in \Omega_i^\star} \hat{\lambda}_{i,j} \sum_{l:\omega_l \in \Omega_i^\star} y(\omega_l) \hat{\lambda}_{i,l} K_i(x_i(\omega_j), x_i(\omega_l))}{\sum_{j:\omega_j \in \Omega_i^\star} \hat{\lambda}_{i,j}} \right)$$

However, there exists a continuum of points for each i for which no decision is given:

$$\hat{x}_{\phi,i} \in \vec{\mathcal{X}}_{\phi,i} \quad , \quad \vec{\mathcal{X}}_{\phi,i} = \{x_i \in \vec{\mathcal{X}}_i : K_i(\hat{\theta}_i, x_i) + \hat{b}_i = 0\} \\ \hat{b}_i = -K_i(\hat{\theta}_i, x_{\phi,i})$$

These are the neutral points (see Figure 1). In the following, we do not, at any stage, need to explicitly calculate them. In particular, where a neutral point is indicated within a calculation, we shall find that it is only required that it be an individual drawn from the total set of neutral points, without any corresponding requirement of specifying *which* specific neutral point it is. In other words, the designator of an

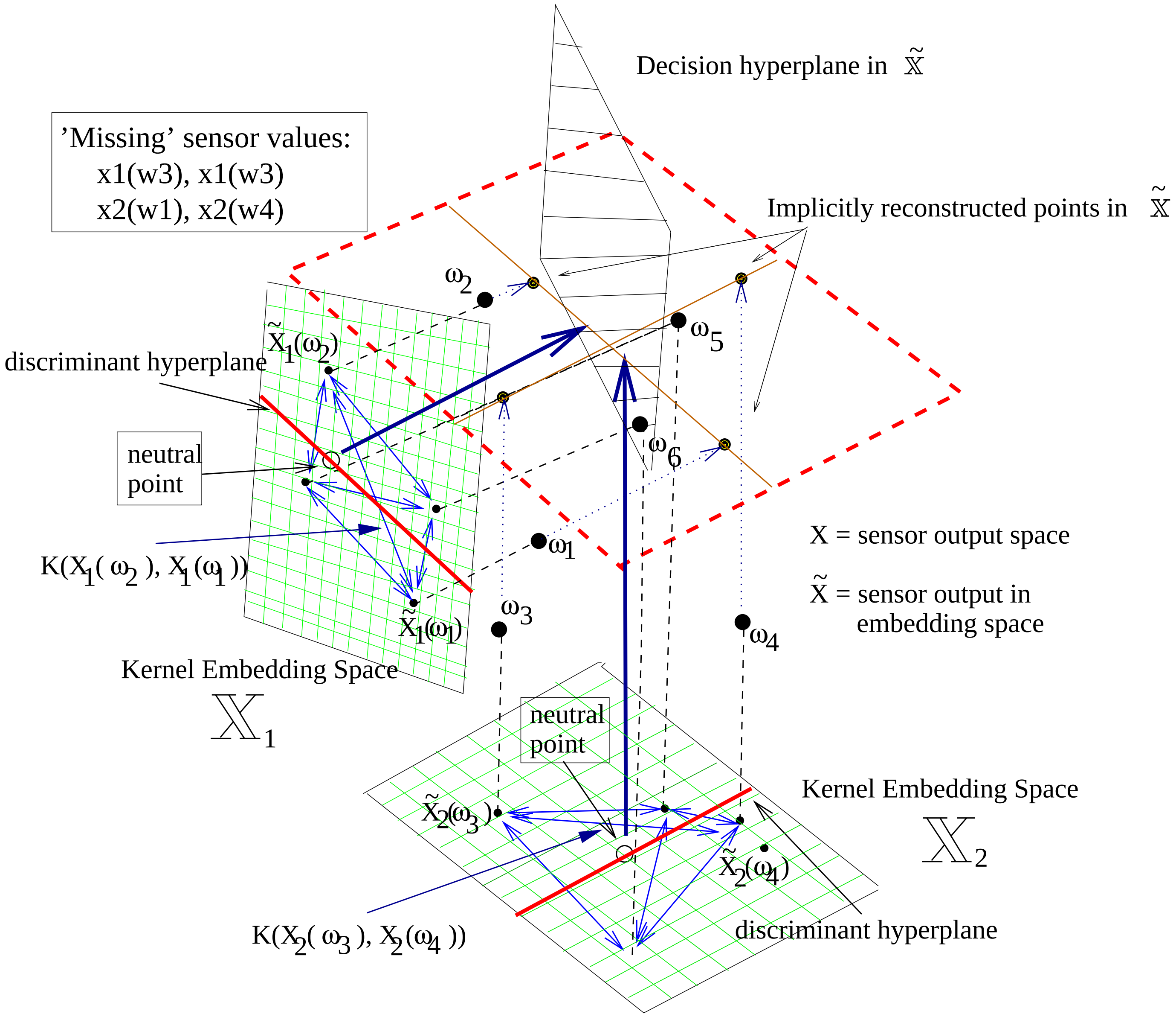


Fig. 1. An illustration of the decision space embedding implicitly constructed by NPS kernel fusion under conditions of partially disjoint modalities - note ill-posedness of embedding without NPS. (For visual clarity, only 3 dimensions of embedding space are rendered, and Kernel similarities are displayed in terms of Euclidean distances)

individual neutral point behaves as a "particularity" operator and not as an indexical within equations.

To proceed further, we now need to explicitly consider the multikernel decision problem. Substituting the most straightforward multi-modal Kernel, the linear kernel where $K(x', x'') = \sum_{i=1}^{n} K_i(x'_i, x''_i)$ into the (non-dual) SVM decision problem, we find that the training criterion becomes:

$$\sum_{i=1}^{n} K_i(\theta_i, \theta_i) + C \sum_{\omega_j \in \Omega_i^\star} \delta_j \rightarrow \min(\theta_i \in \vec{\mathcal{X}}_i, b \in \mathcal{R}, \delta_j \in \mathcal{R})$$

Subject to:

$$\lfloor y_j(K_i(\theta_i, x_i(\omega_j)) + \sum_{l=1, l\neq i}^{n} K_l(\theta_l, x_l(\omega_j)) + b) \geq 1 - \delta_j, \delta_j \geq 0, \omega_j \in \Omega_i^\star \rfloor, i = 1, \ldots, n$$

The question arises immediately as to the existence of the summation terms $K_l(\theta_l, x_l(\omega_j))$ when $l \neq i$; i.e., whether an object designated within one modality-specific kernel embedding space also exists within another modality-specific embedding space. If, for instance, multi-modal training sets are partially disjoint (e.g. when training sets have missing feature values) then the multi-modal kernel problem as specified is not soluble in itself. If multi-modal training sets are completely disjoint (for instance, when the training sets within each modality are proprietary) then the multi-modal kernel problem is maximally ill-posed.

However, because of the presence of the individual modality decision problems in the above constraint optimization problem, we can apply the neutral point substitution as constituting the least biasing value substitution. Thus, rather than proposing a missing data approach that makes strong assumptions about the form of the data (e.g. that it is Gaussian in nature), or else takes only very partial consideration of the nature of the data (as in mean-substitution), we propose to adopt a missing-data approach that is *relevant to the classification problem in hand*. Hence, we replace "missing" sensor values, $x_l(\omega_j), l \neq i$, in (3) by unbiased neutral points: $\hat{x}_{\phi,i} \in \hat{\mathcal{X}}_{\phi,i}$.

As was shown in [10] for the case of completely disjoint modalities, if we make the appropriate neutral point substitutions (i.e., $x_l(\omega_j), l \neq i \rightarrow \hat{x}_{\phi,l}$ within the summation), then the solution to the above equation exhibits linear separability. In fact, it defaults to the sum rule decision scheme for the individual modality-specific SVMs:

$$\hat{f}(x_i(\omega), i = 1, \ldots, n) = \sum_{1=1}^{n} \lfloor K_i(\hat{\theta}_i, x_i(\omega_l)) + \hat{b}_i \rfloor \gtrless 0 \quad (3)$$

This is a very reassuring result, in that it shows that our choice of unbiased substitution for missing data naturally corresponds to the only alternative way of dealing with the completely disjoint data problem (i.e., treating it as a case of decision fusion). Further, it indicates that neutral point substitution readily permits room for the error decorrelation effect to take place (which can be important if the composite Kernel increases the dimensionality of the embedding space to the point at which the "curse of dimensionality" becomes apparent). What is not immediately clear, however, is the extent to which this effect is advantageous for partially disjoint data, where the composite decision space is not so straightforwardly decomposable.

In such partial cases, it is still possible to apply the same neutral point substitution as in the disjoint case, provided that we decompose b into its components at the outset: $b = \sum_{1=1}^{n} \hat{b}_i$ (as in the separable solution). The constraint equation thus becomes:

$$\lfloor y_j(K_i(\theta_i, x_i(\omega_j)) + \hat{b}_i \sum_{l=1, l\neq i}^{n} \Big(K_l(\theta_l, x_l(\omega_j)) + \hat{b}_l\Big))$$
$$\geq 1 - \delta_j, \delta_j \geq 0, \omega_j \in \Omega_i^{\star} \rfloor, i = 1, \ldots, n$$

By applying the neutral point substitution such that $K_l(\theta_l, \hat{x}_{\phi,i}) + b_l = 0$ for missing values within the summation, we have that the summation need only be performed over the known quantities, and the solution is found as for the standard SVM. We hence do not need to explicitly calculate the $\hat{x}_{\phi,i}$.

[BEGIN ADDED TEXT]

We know that the feature vector with missing modality values exists within the composite space when augmented by the neutral points because it is fully defined within the Kernel matrix (i.e. because the neutral points are linearly dependent the support objects). Thus, even if the decision in the composite space can be shown to be decomposable into the individual modalities (i.e. for fully independent data sets), we do not, even in these circumstances, commence classification from the individual modalities; we always work with the composite feature space which potentially has far more classification information than fused output from the individual modalities.

[END ADDED TEXT]

From the neutral-point perspective, it is thus possible to regard SVM classification on partially-disjoint multi-modal data (i.e., data with missing feature components) as being "weighted" towards the sum of the marginal decisions according to the proportion of incomplete data. The exact degree to which this affects overall classification will be data and kernel dependent. We would therefore like to quantify this result for a typical data set.

We hence now turn to an empirical exploration of the neutral point method in a realistic scenario, in which the modal data is only very partially disjoint; that is, where the multimodal data is largely complete, apart from a few missing values (for instance, of the sort that occur in the field of census data returns, for which the method was first developed).

III. Experimental Findings

A. Database, Reference Systems and Experimental Protocols

The data used in our evaluation scheme is taken from the Biosecure database. *Biosecure*[1] is a European project whose aim is to integrate multi-disciplinary research efforts in biometric-based identity authentication. Application examples are a building access system using a desktop-based or a mobile-based platform, as well as applications over the Internet such as tele-working and web or remote-banking services. As far as the data collection is concerned, three scenarios have been identified, each simulating the use of biometrics in remote-access authentication via the Internet (termed the "Internet" scenario), physical access control (the "desktop" scenario), and authentication via mobile devices (the "mobile" scenario) [15].

For the purpose of our experiments, we used the subset of the desktop scenario[2] which further contains a subset of still face, 6 fingers and iris modalities, denoted by fa1, ft1–6 and ir1, respectively. These 8 channels of data, as well as the reference systems, and the experimental protocols are summarized in Table I.

Note that for the purpose of performance assessment, the data set and experimental protocols are not the primary concern; any database could have been used. The only requirement is that a wide variety of biometric modalities are used in order to illustrate the generality of our approach.

It is important to note that there are two score data sets: the development and the evaluation sets (see Table I(c)). In this table, S1 means the session 1 data whereas S2 means the session 2 data. For *each client*, the data in S1 consists of two samples collected within the same session. They are collected to facilitate the development of a baseline system (i.e., for enrollment). It is known that intra-session performance is biased [16].

To illustrate this systematic bias, we compare the performance of the same session (S1) versus that of different-session (i.e., S2), for each of the 8 channels of data, in terms of Equal Error Rate (EER), in Figure 2. As can be observed, the same-session performance is systematically better than the different-session performance.

[1] http://www.biosecure.info/

[2] The matching scores used in the experiments are available for download at: http://personal.ee.surrey.ac.uk/Personal/Norman.Poh/web/fusionq.

TABLE I

A LIST OF CHANNELS OF DATA FOR EACH BIOMETRIC MODALITY CAPTURED USING A GIVEN DEVICE.

(a) Channels of data

Label	template ID {n}	Modality	Sensor	Remarks
fa	1	Still Face	web cam	Frontal face images (low resolution)
ft	1–6	Fingerprint	Thermal	1/4 is right/left thumb; 2/5 is right/left index; 3/6 is right/left middle finger
ir	1–2	Iris image	LG	1 is left eye; 2 is right eye

TABLE II

A LIST OF CHANNELS OF DATA FOR EACH BIOMETRIC MODALITY CAPTURED USING A GIVEN DEVICE.

(a) Reference systems

Modality	Reference systems
Still Face	Omniperception's Affinity SDK face detector; LDA-based face verifier
Fingerprint	NIST Fingerprint system
Iris	A variant of Libor Masek's iris system

(b) Protocols

Data sets		No. of matching scores	
		dev (51 persons)	eva (156 persons)
S1	Genuine	1 × 51	1 × 156
	Impostor	103 × 4 × 51	51 × 4 × 156
S2	Genuine	2 × 51	2 × 156
	Impostor	103 × 4 × 51	126 × 4 × 156

[BEGIN ADDED TEXT] Acronyms: S1 = session one; S2 = session two; dev = development (training) set; eva = evaluation (test) set
Example: The entry "103 × 4 × 51" in column dev and row S2:Impostor indicates the number of scores due to comparing 51 client references against the queries of 103 impostors, each having 4 attempts. The entry "2 × 156" in column eva and row S2:Genuine indicates the number of genuine matching scores due to comparing 156 client references each having two genuine samples. **[END ADDED TEXT]**

Due to the above systematic bias, we shall use the S2 development (dev) set for training and the S2 evaluation set (eva) set for testing all fusion algorithms.

The iris baseline system used here is far from the performance claimed by Daugman's implementation [17]. We verified that this is due to bad iris segmentation and a suboptimal threshold for distinguishing eyelashes from iris (being baselines, no effort was made to optimize performance; the only requirement is that all systems output match scores. In case of failure to match or to extract features, the system will output a dummy value ("-999") to denote missing a missing score.

Two factors can result in missing modalities. First, during the data collection process, some volunteers did not complete a whole session. Second, some acquired biometric samples are so low in quality that they cannot be processed by our feature extraction algorithm, or the resultant extracted features could not be used for matching. Being well controlled, the development set contains almost complete observations; however a fraction of samples in the evaluation set (8348 out of 76920) contain some missing modalities.

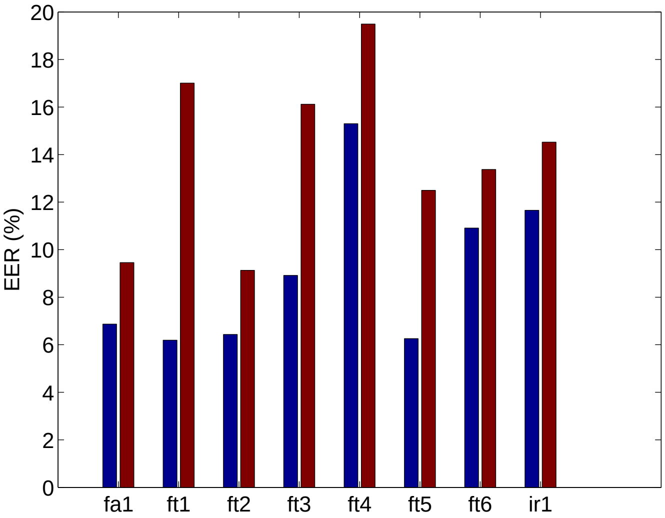


Fig. 2. The error of the development set (blue) versus that of evaluation set (red) of the 8 systems used in the cost-sensitive evaluation of the original Biosecure data set.

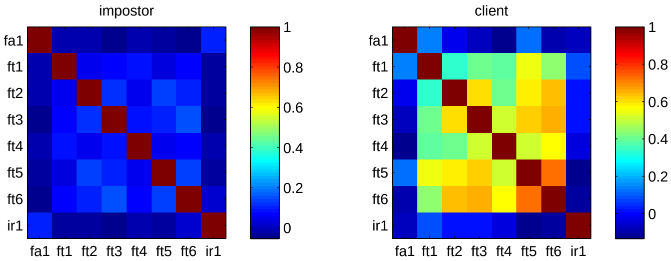


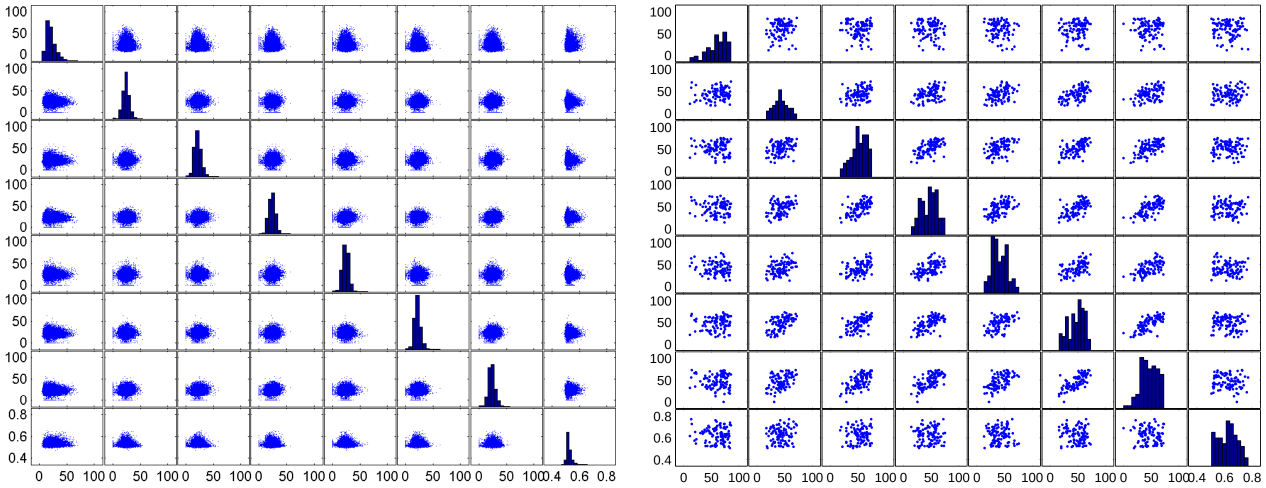
Fig. 4. Correlation matrix of impostor and genuine (client) scores on the left and right panels, respectively

B. Correlation Analysis of the Match Scores

A matrix plot consisting of a pairs of biometric systems delivering impostor match scores is shown in Figure 3. The corresponding genuine user match scores are similar and, hence not shown here.

It is useful to summarize the two class-conditional covariance matrices by their correlation matrices since correlation is invariant to variable scaling and is bounded in $[-1, 1]$, with 1 (resp. -1) being perfect positive (resp. negative) correlation. The correlation matrix of the impostor and client match scores calculated on the development set are shown in Figure 4.

There are two points to note. First, the impostor match scores have generally correlation entries close to zero. Second, the correlation among all the six fingers (columns 2 to 7, resp.



(a) impostor

(b) client

Fig. 3. Scatter plot of (a) impostor and (b) client match scores

rows 2 to 7) are *all* positive, between 0.3 to 0.6. According to [18], this indicates that combining two fingerprint systems may not be as effective as combining two different biometric traits, e.g., a fingerprint and a face biometric. The problem is therefore implicitly *multi*-modal, and can be Kernelized in terms of SVM recognition within the individual modalities.

C. Results

Using the neutral point substitution method outlined in Section 2, we therefore specified an experimental scenario in which the SVM classifier acts both individually upon the modalities of the Biosecure database, and collectively via sum rule decision fusion and composite Kernelization. Composite kernelization is carried-out via the linear kernel $K(x', x'') = \sum_{i=1}^{n} K_i(x_i', x_i'')$ with neutral point substitution undertaken for the missing values. An inner product kernel is chosen for transparency within the individual modalities.

[BEGIN ADDED TEXT] Although the original data contains some missing modalities, it is instructive to examine the effect of missing modalities to various degrees. In achieve this, we use the following procedure: Let M be a matrix of scores of N samples by d dimensions (corresponding to all the d columns of match scores from $d = 8$ devices: face, 6 fingers and 1 iris). The total number of elements in M is $d \times N$. Missing values were gradually introduced by replacing T observed values with "-999" (the dummy value denoting missing value) in such a way that all the elements in the matrix M have equal probability of being deleted. We varied T such that the ratio of $T/(dN)$ is 10%, 20% and 30% and that the subsequent subset always contained the missing values of its precedent subset. **[END ADDED TEXT]**

[BEGIN ADDED TEXT] The results of these tests are given as superimposed DET curves in Figure 5. The methods are explained below:

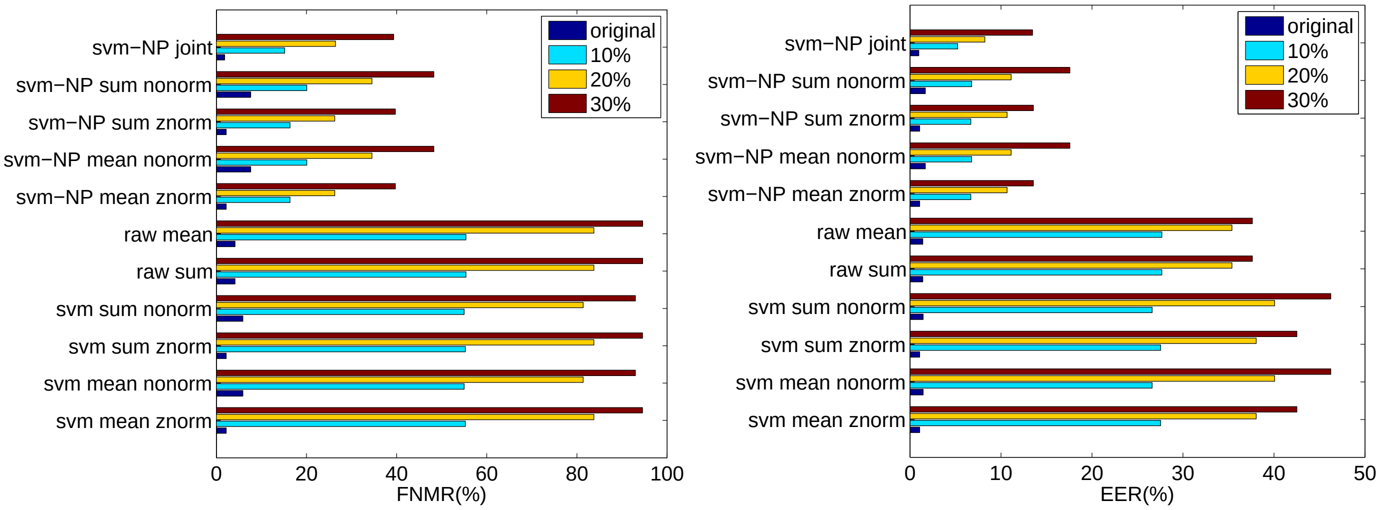
1) `SVM-NP joint` denotes the proposed SVM approach with neutral point substitution. In this case, a single SVM was trained with the joint-score space of 8 features (corresponding to the 8 expert outputs described in Section III) using full observations. During inference, missing values are substituted with neutral points.
2) `SVM-NP sum` $norm$, where $norm \in \{\mathtt{znorm}, \mathtt{nonorm}\}$, is another implementation of SVM-NP which assumes the expert outputs to be maximally disjoint. An SVM is therefore trained for each expert and the outputs of the SVM ensemble (for observable modalities) are then summed. The $norm$ here indicates the type of normalization procedure, which can either be with the zero-mean unit variance normalization (`znorm`), or without any normalization (`nonorm`). The parameters of Z-norm, i.e., mean and standard deviation, are calculated on the output of each SVM on the entire training set.
3) `SVM-NP mean` $norm$ is the same as configuration 2 above, except that the mean fusion rule is used instead of the sum rule.
4) `SVM` $fuse$ $norm$ for $fuse \in \{\mathtt{sum}, \mathtt{mean}\}$ and $norm \in \{\mathtt{znorm}, \mathtt{nonorm}\}$ is the same as the configurations 1–3, as explained above, except that SVM are used without NPS.
5) `raw` $fuse$, where $fuse \in \{\mathtt{sum}, \mathtt{mean}\}$, indicates direct fusion of the raw scores using either the sum or the mean rule.

[END ADDED TEXT]

[BEGIN ADDED TEXT] In the experiments, two performance indicators are used: Equal Error Rate (EER) and False Match Rate (FNMR) at False Non-Match Rate (FMR) of 0.1%. EER is the operating point where FMR and FNMR are equal. FMR, also known as false acceptance rate, is the estimated probability of accepting an impostor. FNMR, on the other hand, is the probability of rejecting a client. Both EER and FNMR@FMR=0.1% are commonly used in the biometric literature. **[END ADDED TEXT]**

The following observations can be made:

- `SVM-NP` $fuse$ $norm$ is better than any `SVM` $fuse$ $norm$, for any fusion strategy $fuse \in \{\mathtt{sum}, \mathtt{mean}\}$ and any normalization strategy $norm \in \mathtt{nonorm}, \mathtt{znorm}\}$.



(a) Verification rate

(b) Equal Error Rate

Fig. 5. Performance of the baseline expert systems and that of fusion with various SVM method as well as that of the sum rule.

- `SVM-NP` $fuse$ `nonorm`, for any fusion strategy, $fuse \in \{\texttt{sum}, \texttt{mean}\}$, is slightly inferior to its counterpart `SVM-NP` $fuse$ `znorm`. This shows the important of normalizing the outputs of SVM in the maximally disjoint case.
- `SVM-NP joint` achieves the maximal generalization performance in all data sets.

IV. Discussion and Conclusions

At the outset of this investigation it was conjectured on theoretical grounds that the neutral point method is an appropriate strategy for treating missing values in multi-kernel problems with the potential to retain the error-decorrelation advantages of the sum-rule decision scheme in typical test scenarios with partial missing data. Experiments were consequently conducted on multimodal biometric data from the Biosecure database, in which both multi-kernelization and the missing data problem arose naturally, in order to complement the earlier theoretical analysis derived for the asymptotic scenario of complete data-disjunction.

Results (Fig. 5) demonstrate that the sum rule decision scheme is indeed superior to any individual modal decision rule on the tested data, but that significantly greater advantage arises from using a composite kernel (which would, in itself, be impossible without missing value substitution). **[BEGIN ADDED TEXT]** The experiments thus demonstrate that the advantage of the NP method is two-fold; firstly, we are able to exploit all of the available measurement data (i.e without discarding any), and secondly, we are able to exploit all of the available correlation information.

The observed performance improvement in moving from SVM-NP-joint to SVM-NP-sum-znorm is thus a measure of the discriminative information available within the modalities when properly-correlated. Even if such correlation were not evident, the argument for using the NP method is that it can perform *no worse* than decision fusion, and that NP substitution is therefore an intrinsically safe default option. (Thus, in practice, we need not consider the issue of correlation and simply use SVMs via NPS in *any* circumstances with missing data; when there is correlation, this approach produces significantly better results than any given decision fusion strategy). **[END ADDED TEXT]**

We hypothesize that the latter will be typical for naturally-arising multi-kernel, missing-data problems (i.e. data in which missing values are relatively rare). The neutral point method is thus the appropriate "first-resort" strategy to consider in these cases, as opposed to the alternative of multimodal decision fusion; particularly as decision fusion is asymptotically implicit in the neutral point approach.

Because of the nature of the derivation of the neutral point method, there is no explicit requirement for actual value substitution, and the method gives rise to minimal changes to the cost function of linearized kernel composition. Furthermore, the method differs from previous approaches in that the missing values are related to the decision problem rather than to the data distribution. In this way it is consistent with the broad philosophy of maxim margin SVM-based approaches. We thus characterize the neutral point method as an empirically safe, well-posed and discriminitively-unbiased approach to missing data substitution. Moreover, its straightforward methodological application in terms of complete training cases makes it naturally congruent with the problem of multimodal biometrics.

Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n^o 215078 and the grant IST-214324 (www.mobioproject.org). We would also like to thank EPSRC for funding received through grant EP/F069626/1 and the Swiss NSF grant through the advanced research fellowship PA0022_121477.

REFERENCES

[1] A. Ross, K. Nandakumar, and A.K. Jain, *Handbook of Multibiometrics*, Springer Verlag, 2006.

[2] C. Sanderson, *Automatic Person Verification Using Speech and Face Information*, Ph.D. thesis, Griffith University, Queensland, Australia, 2002.

[3] K. Nandakumar, "Integration of M]ultiple Cues in Biometric Systems," M.S. thesis, Michigan State University, 2005.

[4] Karthik Nandakumar, Anil K. Jain, and Arun Ross, "Fusion in multibiometric identification systems: What about the missing data?," in *ICB*, 2009, pp. 743–752.

[5] N. Poh, T. Bourlai, J. Kittler, L. Allano, F. Alonso-Fernandez, O. Ambekar, J. Baker, B. Dorizzi, O. Fatukasi, J. Fierrez, H. Ganster, J. Ortega-Garcia, D. Maurer, A. A Salah, T. Scheidat, and C. Vielhauer, "Benchmarking quality-dependent and cost-sensitive score-level multimodal biometric fusion algorithms," *IEEE Trans. on Information Forensics and Security*, vol. 4, no. 4, pp. 849–866, 10 2009.

[6] O. Fatukasi, J. Kittler, and N. Poh, "Estimation of missing values in multimodal biometric fusion," in *IEEE Conference on Biometrics: Theory, Applications and Systems*, Washington, D.C., 2009, pp. 1–6.

[7] Juan Manuel, Lucas Cuesta, Ricardo De Crdoba Herralde, Luis Fern, O Dharo Enrquez, and Grupo De Tecnologa Del Habla, "Applying feature reduction analysis to a pprlm-multiple gaussian language identification system," 2009.

[8] T I. Lin, J C. Lee, and H J. Ho, "On fast supervised learning for normal mixture models with missing information," *Pattern Recognition*, , no. 6, pp. 1177–1187, 2006.

[9] T. Roos, H. Wettig, P. Grünwald, P. Myllymäki, and H. Tirri, "On discriminative bayesian network classifiers and logistic regression," *Mach. Learn.*, vol. 59, no. 3, pp. 267–296, 2005.

[10] D. Windridge, V. Mottl, A. Tatarchuk, and A. Eliseyev, "The neutral point method for kernel-based combination of disjoint training data in multi-modal pattern recognition problem," in *LNCS 4472/2007, Proc. Multiple Classifier Systems 2007 (MCS 2007)*, 2008, pp. 13–21.

[11] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2007.

[12] J. Kittler, M. Hatef, R. P.W. Duin, and J. Matas, "On Combining Classifiers," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226–239, 1998.

[13] N. Poh, T. Bourlai, and J. Kittler, "A multimodal biometric test bed for quality-dependent, cost-sensitive and client-specific score-level fusion algorithms," in *Pattern Recognition Journal*, March 2010, pp. 1094–1105, doi:10.1016/j.patcog.2009.09.011.

[14] A. Jain, K. Nandakumar, and A. Ross, "Score Normalisation in Multimodal Biometric Systems," *Pattern Recognition*, vol. 38, no. 12, pp. 2270–2285, 2005.

[15] J. Ortega-Garcia, J. Fierrez, F. Alonso-Fernandez, J. Galbally, M. R. Freire, J. Gonzalez-Rodriguez, C. Garcia-Mateo, J-L. Alba-Castro, E. Gonzalez-Agulla, E. Otero-Muras, S. Garcia-Salicetti, L. Allano, B. Ly-Van, B. Dorizzi, J. Kittler, T. Bourlai, N. Poh, F. Deravi, R. Ng, M. Fairhust, J. Hennebert, A. Humm, M. Tistarelli, L. Brodo, J. Richiardi, A. Drygajlo, H. Ganster, F. Sukno, S-K. Pavani, A. Frangi, L. Akarun, and A. Savran, "The multi-scenario multi-environment biosecure multimodal database (bmdb)," *IEEE Trans. on Pattern Analysis and Machine*, 2009, accepted for publication.

[16] A. Martin, M. Przybocki, and J. P. Campbell, *The NIST Speaker Recognition Evaluation Program*, chapter 8, Springer, 2005.

[17] J. Daugman, *How Iris Recognition Works*, chapter 6, Kluwer Publishers, 1999.

[18] N. Poh and S. Bengio, "How Do Correlation and Variance of Base Classifiers Affect Fusion in Biometric Authentication Tasks?," *IEEE Trans. Signal Processing*, vol. 53, no. 11, pp. 4384–4396, 2005.