Handling Multimodal Information Fusion with Missing Observations Using the Neutral Point Substitution Method

David Windridge[1], Norman Poh[1], Vadim Mottl[2], Alexander Tatarchuk[2], and Andrey Eliseyev[2]

[1] CVSSP, University of Surrey, The Stag Hill, Guildford, GU2 7XH, UK
[2] Computing Center of the Russian Academy of Sciences, Vavilov St. 40, Moscow, 119991, Russia

Abstract. We have previously introduced, in purely theoretical terms, the notion of neutral point substitution for missing kernel data in multimodal problems. In particular, it was demonstrated that when modalities are maximally disjoint, the method is precisely equivalent to the Sum rule decision scheme. As well as forging an intriguing analogy between multikernel and decision-combination methods, this finding means that the neutral-point method should exhibit a degree of resilience to class misattribution within the individual classifiers through the relative cancelling of combined estimation errors (if sufficiently decorrelated).

However, the case of completely disjoint modalities is unrepresentative of the general missing data problem. We here set out to experimentally test the notion of neutral point substitution in a realistic experimental scenario with partially-disjoint data to establish the practical application of the method. The tested data consists in multimodal Biometric measurements of individuals in which the missing-modality problem is endemic. We hence test a SVM classifier under both the modal decision fusion and neutral point-substitution paradigms, and find that, while error cancellation is indeed apparent, the genuinely multimodal approach enabled by the neutral-point method is superior by a significant factor.

1 Introduction

In a paper given at the last MCS meeting [9], we set out a strategy for addressing the problem of missing modalities in multimodal kernel data. The problem of missing features is well-known in general pattern recognition, but can be addressed (aside from simply omitting the missing-data samples) using methods such as mean substitution [1], at the simplest level, or else via more complex methods (eg [3]) that take into account specifics of the distribution statistics and morphology.

However, in multimodal kernel decision problems the issue of missing features becomes acute (multimodal kernel decision problems are those in which feature maps $\hat{\phi}$ giving $\mathcal{R}^N$ outputs for detected objects ω are associated either with particular sensor spaces; $\hat{\phi^m}(S_m(\omega)) \rightarrow \mathcal{R}^{N_m}$, or else with particular kernel

J.A. Benediktsson, J. Kittler, and F. Roli (Eds.): MCS 2009, LNCS 5519, pp. 161–170, 2009.

measures $K_m(\hat{\phi}^m(\cdot), \hat{\phi}^m(\cdot)) \rightarrow \mathcal{R}$ on some, possibly even *common*, sensor-output space S)[1]. The difficulties arise because we cannot, in general, assume that the Kernel matrix $\mathbf{K_n} = K_n(\hat{\phi}^n(S(\omega_i)), \hat{\phi}^n(S(\omega_j)))$ defined on a per-mode basis (ie applicable only to mode n) will give rise to the *same* Mercer embedding space, $\hat{\psi}^n(S) = (\psi_1^n(S), \psi_2^n(S), \psi_3^n(S), \ldots)'$ when the set from which i and j are drawn has differing cardinalities, r, due to the missing data[2]. (The functions $\psi_i^n(S)$ being Eigenfunctions of the Kernel matrix $\mathbf{K_n}$; ie 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 $u_i = \psi_i(S(\omega_i)))$.

We cannot therefore simply assume that modes can be combined into a composite (Mercer) pattern space in which to perform classification (as in standard pattern recognition)[3]. Furthermore, even if this composition of spaces can be achieved, there are no kernels defined *a priori* within it since there are no *inter*-modal kernels defined at the outset, as would be required for pattern recognition based on Euclidean (l^2-norm), or quasi-Euclidean (l^n-norm) assumptions. Consequently, there is an ambiguity as to how the problem should be approached.

We therefore, rather, approach the problem from the opposite direction, firstly defining a composite kernel capable of accommodating missing Kernel values, and only secondarily considering the nature of the space in which this composite kernel is embedded, consistent with the ideals of kernel-based approaches generally.

The paper is therefore structured as follows: in the following section we recap the neutral-point approach to missing value substitution, and indicate its relation to the Sum-Rule decision scheme. In section 3 we detail the application of the method to multi-modal Biometric data (data in which the missing value problem arises naturally). Section 4 discusses experimental outcomes and makes concluding remarks.

[1] In which case the problem is similar to that of multikernel learning [6]. However, we here assume that that the association of kernels with output sensors with is in some way intrinsic to the experimental setup (e.g. the association of genetic-distance with gene-data), rather than being determined after the fact via an optimisation process as is sometimes the case in Multikernel learning).

[2] We could, of course, simply overlook this issue, and combine outputs in the sensors' tensor space; however this is to make the notion of using *mode-specific* Kernels redundant, in particular, losing the inherent advantages of the problem-relevant, minimalist and linearised embedding spaces so constructed. Furthermore, in certain domains (eg step-wise gene-distance measurements [8]), where there is no absolute underlying sensor space, it is only possible to work with kernel embedding spaces.

[3] This is not necessary problematic for other Kernel applications. For instance, a method for approaching the missing data issue at the objective function level in Kernel PCA is given in [7]; it utilises cross entropy with respect to a Gaussian distribution as the cost function to be minimised with respect to the missing values. However, this makes implicit assumptions about the data (namely that it can be modelled as a Gaussian with the given kernel), and takes a significant amount of time to compute.

2 The Neutral Point Method

We here recap the essentials of the neutral point method; for more detailed information refer to [9]. For clarity in this section, we assume an underlying unidimensional sensor space within each mode, 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 (ie where x maps objects ω into a common real valued space):

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

Any kernel $K_i(x_i', x_i'')$ embeds the scale of the respective sensor $\mathcal{X}_i$ (equipped with with inner product) into a hypothetical linear space (the embedding space), $\hat{\mathcal{X}}_i \supseteq \mathcal{X}_i$, in which the null element and linear operations are defined.

For a single modality, the training set:

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

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

$$\varOmega_i^\star => \{\mathbf{K}_i = \lfloor K_i(x_i(\omega_j), x_i(\omega_l)), \omega_j, \omega_l \in \varOmega_i^\star \rfloor, y(\omega_j), \omega_j \in \varOmega_i^\star\} \tag{3}$$

Support Vector Machines (SVMs) are the most common Kernel-based approach approach to 2-class pattern recognition, the problem being to find maximal margin discriminant hyperplane in space $\boldsymbol{\mathcal{X}}_i$:

$$\boldsymbol{y}_i(x_i(\omega)) = K_i(\theta_i, x_i(\omega)) + b_i \overset{>}{<} 0 \tag{4}$$

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

This leads to the standard SVM Training Criterion:

$$K_i(\theta_i, \theta_i) + C \sum_{\omega_j \in \varOmega_i^\star} \delta_j \to \min(\theta_i \in \boldsymbol{\mathcal{X}}_i, b \in \mathcal{R}, \delta_j \in \mathcal{R}) \tag{5}$$

Subject to:

$$y_i \lfloor K_i(\theta_i, x_i(\omega_l)) + b \rfloor \geq 1 - \delta_j, \delta_j \geq 0 \tag{6}$$

The (Wolfe) dual form of the criterion is a quadratic programming problem with respect to the Lagrangian multipliers, λ:

$$\sum_{\omega_j \in \varOmega_i^\star} \lambda_{i,j} - (1/2) \sum_{\omega_j \in \varOmega_i^\star} \sum_{\omega_l \in \varOmega_i^\star} \lfloor y_j y_l K_i(x_i(\omega_j), x_i(\omega_l)) \rfloor \lambda_{i,j} \lambda_{i,l} \to \max \tag{7}$$

Subject to:

$$\sum_{\omega_j \in \varOmega_i^\star} y_j \lambda_{i,j} = 0, 0 \leq \lambda_{i,j} \leq C/2, \omega_j \in \varOmega_i^\star \tag{8}$$

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

$$\hat{f}(x_i(\omega)) = \sum_{j:\omega_j \in \Omega_i^\star} y_j \hat{\lambda}_{i,j} K_i(x_i(\omega_j), x_i(\omega_l)) + \hat{b}_i \gtrless 0 \tag{9}$$

with:

$$\hat{b}_i = -\left(\sum_{j:\omega_j \in \Omega_i^\star} \hat{\lambda}_{i,j} \sum_{l:\omega_l \in \Omega_i^\star} y(\omega_l) \hat{\lambda}_{i,l} K_i(x_i(\omega_j), x_i(\omega_l)) / \sum_{j:\omega_j \in \Omega_i^\star} \hat{\lambda}_{i,j} \right) \tag{10}$$

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

$$\hat{x}_{\phi,i} \in \boldsymbol{\mathcal{X}}_{\phi,i}, \boldsymbol{\mathcal{X}}_{\phi,i} = \{x_i \in \boldsymbol{\mathcal{X}}_i : K_i(\hat{\theta}_i, x_i) + \hat{b}_i = 0\}, \hat{b}_i = -K_i(\hat{\theta}_i, x_{\phi,i}) \tag{11}$$

These are the neutral points. In the following, we do not, at any stage, need to explicitly calculate them. In particular, where an individual neutral point is used in calculation, we shall find that it is only required that the neutral point be one drawn from the total set of of neutral points, without having the requirement of specifying *which* neutral point it is. In other words the designator of an individual neutral point behaves like a 'particularity' operator and not an indexical operator.

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:

$$K_i(\theta_i, \theta_i) + C \sum_{\omega_j \in \Omega_i^\star} \delta_j \to \min(\theta_i \in \boldsymbol{\mathcal{X}}_i, b \in \mathcal{R}, \delta_j \in \mathcal{R}) \tag{12}$$

Subject to:

$$\lfloor y_j(K_i(\theta_i, x_i(\omega_j)) + \sum_{l=1, l \neq i}^{n} K_l(\theta_l, x_l(\omega_j)) + b) \geq 1 - \delta_j, \delta_j \geq 0, \omega_j \in \Omega_i^\star \rfloor, i = 1, \ldots, n \tag{13}$$

However, the question arises as to the existence of the terms $K_l(\theta_l, x_l(\omega_j))$, when $l \neq i$; that is, where an object designated within one mode's kernel embedding space also exists within another mode's kernel space. If, for instance, multi-modal training sets are partially disjoint (e.g. when training sets have missing feature values) then the multi-mode kernel problem is not soluble in itself. If multi-modal training sets are completely disjoint (for instance, when the training sets within each mode are proprietary) then the multi-modal kernel problem is maximally intractable.

However, because of the presence of the individual modes' decision problems in the above constraint optimisation problem, we can apply the neutral point substitution as constituting the least biasing value substitution *given the decision problem in question.* Thus, rather than proposing a missing data approach that makes strong assumptions about the form of the data (perhaps 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 the replace 'missing' sensor values $x_l(\omega_j), l \neq i$, by unbiased neutral points: $\hat{x}_{\phi,i} \in \hat{\mathcal{X}}_{\phi,i}$.

It was shown [9] that, in the case of completely disjoint modalities, the solution to the above equation with appropriate neutral point substitutions (such that $K_l(\theta_l, x_l(\omega_j)) + b = 0$) becomes linearly separable in b, and defaults to the sum rule decision scheme for the individual modes' SVMs:

$$\hat{f}(x_i(\omega), i =, \ldots, n) = \sum_{1=1}^{n} \lfloor K_i(\hat{\theta}_i, x_i(\omega_l)) + \hat{b}_i \rfloor \gtrless 0 \qquad (14)$$

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 (ie 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 Mercer space is not so straightforwardly decomposable into its marginal components. In this case, we can regard the the completed data (ie the data without missing components) as 'weighting' the summed marginal decisions on the basis of the intra-modal correlations to an extent that is governed by their proportion of the total data. The exact degree to which this occurs will be data and kernel dependant. 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 (eg, of the sort that occur in the field of census data returns).

3 Experimental Findings

3.1 Database, Reference Systems and Experimental Protocols

The data used in our evaluation scheme is taken from the Biosecure database. *Biosecure*[4] 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,

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

Table 1. 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

(b) 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

(c) Protocols

Data sets		No. of match scores per person	
		dev (51 persons)	eva (156 persons)
S1	Gen	1	1
	Imp	103×4	51×4
S2	Gen	2	2
	Imp	103×4	126×4

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). A report on the complete Biosecure database is being drafted.

For the purpose of our experiments, we used the subset of desktop scenario, 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 system, and the experimental protocols are summarized in Table 1.

Note that for the purpose of performance assessment, the main objective of this paper, 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: development and the evaluation sets (see Table 1(c)). In this table, S1 means the session 1 data whereas S2 means the session 2 data. The data in S1 consists of two samples collected within the same session. They are collected to facilitate the development of a baseline system. It is known that intra-session performance is biased [4]. For this reason, we shall use the S2 data for our evaluation. A plot of EER for the 8 channels of data is shown in Figure 1. The iris baseline system used here is far from the performance claimed by Daugman's implementation [2]. 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).

Two factors can result in missing modalities. First, during the data collection process, some volunteers did not complete a whole session. Second, some acquired

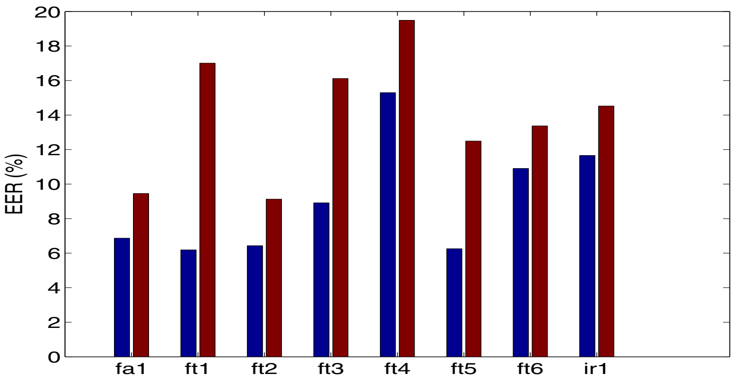

Fig. 1. The error of the development set (blue) versus that of evaluation set (red) of the 8 systems used in the cost-sensitive evaluation of the Biosecure data set

Table 2. Correlation matrix of genuine scores. fa1=face, ft1–6: fingerprints, ir1=iris match scores.

fa1	ft1	ft2	ft3	ft4	ft5	ft6	ir1
1.00	0.15	-0.03	-0.07	-0.13	0.12	-0.08	-0.07
0.15	1.00	0.33	0.40	0.39	0.56	0.44	0.08
-0.03	0.33	1.00	0.60	0.41	0.58	0.64	0.02
-0.07	0.40	0.60	1.00	0.51	0.62	0.66	0.02
-0.13	0.39	0.41	0.51	1.00	0.51	0.56	-0.04
0.12	0.56	0.58	0.62	0.51	1.00	0.73	-0.12
-0.08	0.44	0.64	0.66	0.56	0.73	1.00	-0.11
-0.07	0.08	0.02	0.02	-0.04	-0.12	-0.11	1.00

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.

3.2 Correlation Analysis of the Match Scores

We may summarise the data as paired biometric systems delivering impostor match scores[5] (the corresponding genuine user match scores are similar and, hence not considered here).

In particular, 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.

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

Table 3. Impostor scores

1.00	-0.02	-0.02	-0.05	-0.01	-0.03	-0.05	0.10
-0.02	1.00	0.06	0.08	0.09	0.04	0.07	-0.03
-0.02	0.06	1.00	0.12	0.05	0.14	0.10	-0.03
-0.05	0.08	0.12	1.00	0.08	0.10	0.15	-0.05
-0.01	0.09	0.05	0.08	1.00	0.05	0.07	-0.02
-0.03	0.04	0.14	0.10	0.05	1.00	0.14	-0.03
-0.05	0.07	0.10	0.15	0.07	0.14	1.00	0.02
0.10	-0.03	-0.03	-0.05	-0.02	-0.03	0.02	1.00

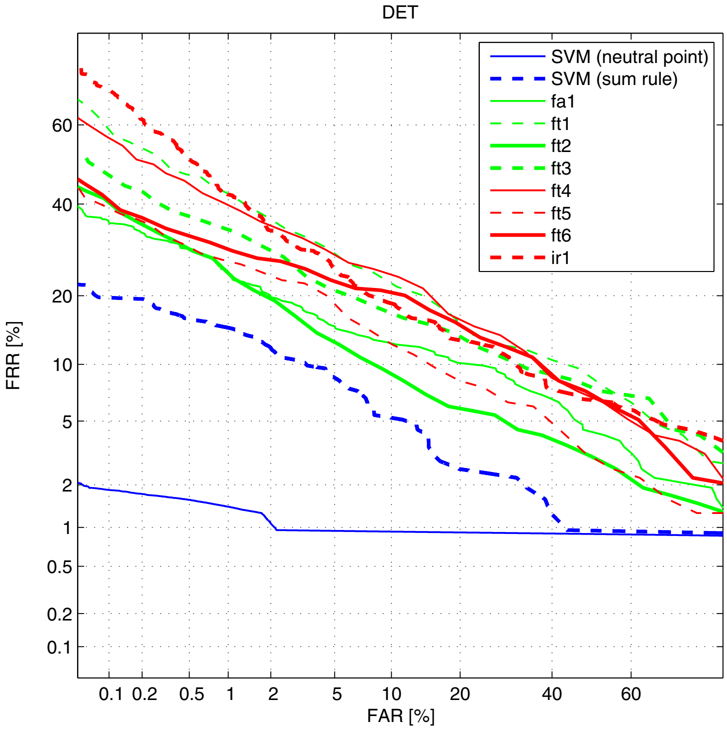

Fig. 2. Performance of the baseline expert systems and that of fusion with SVM using the sum rule and the neutral point substitution method

negative) correlation. The correlation matrix of the impostor and client match scores calculated on the development set are shown in Table 2 and Table 3.

There are three 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. rows 2 to 7) are *all* positive, albeit having small values. Third, the correlation among the genuine match scores of all the six fingers (columns 2 to 7, resp. rows 2 to 7) have relatively high values (from 0.3 to 0.6). According to [5], 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 kernelised in terms of SVM recognition within the individual modes.

3.3 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 kernelisation 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.

The results of these tests are given as superimposed ROC curves in Figure 2.

4 Discussion and Conclusions

It was demonstrated theoretically 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-kernelisation and the missing data problem arose naturally, in order to complement the theoretical analysis derived for the asymptotic scenario of complete data-disjunction.

Results (Fig. 2) demonstrate that the sum rule decision scheme is indeed superior to any individual modal decision rule on the tested data, but that very significantly greater advantage arose from the composition of Kernels (which would, in itself, be impossible without missing value substitution). We hypothesise that this result will be typical for naturally-arising multi-kernel, missing-data problem (data in which missing values are relatively rare). The neutral point method is thus an appropriate 'first-resort' strategy to consider in these cases, as opposed to modal fusion; particularly as the latter is implicit in the former.

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 linearised 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 conclude that the neutral point method can be characterised as an empirically safe and theoretically-unbiased approach to missing data substitution.

Acknowledgements

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.

This work was supported partially by the advanced researcher fellowship PA0022_121477 of the Swiss National Science Foundation, and by the EU-funded Mobio project grant IST-214324 (www.mobioproject.org).

References

1. Cuesta, J.M.L., de Cordoba Herralde, R., D'Haro Enriquez, L.F.: Applying feature reduction analysis to a pprlm-multiple gaussian language identification system. In: Articulo en actas de las V Jornadas de Tecnologia, pp. 29–32 (2008)
2. Daugman, J.: How Iris Recognition Works, ch. 6. Kluwer Publishers, Dordrecht (1999)
3. Lin, T.I., Lee, J.C., Ho, H.J.: On fast supervised learning for normal mixture models with missing information. Pattern Recognition 39(6), 1177–1187 (2006)
4. Martin, A., Przybocki, M., Campbell, J.P.: The NIST Speaker Recognition Evaluation Program, ch. 8. Springer, Heidelberg (2005)
5. Poh, N., Bengio, S.: How Do Correlation and Variance of Base Classifiers Affect Fusion in Biometric Authentication Tasks? IEEE Trans. Signal Processing 53(11), 4384–4396 (2005)
6. Rakotomamonjy, A., Bach, F., Canu, S., Grandvalet, Y.: Simplemkl. Journal of Machine Learning Research (2008)
7. Sanguinetti, G., Lawrence, N.D.: Missing data in kernel pca (2006), `http://www.dcs.shef.ac.uk/intranet/research/resmes/CS0608.pdf`
8. Walsh, B.: Estimating the time to the mrca for the y chromosome or mtdna for a pair of individuals. Genetics 158, 897–912 (2001)
9. Windridge, D., Mottl, V., Tatarchuk, A., Eliseyev, A.: The neutral point method for kernel-based combination of disjoint training data in multi-modal pattern recognition. In: Haindl, M., Kittler, J., Roli, F. (eds.) MCS 2007. LNCS, vol. 4472, pp. 13–21. Springer, Heidelberg (2007)