Kernel Combination via Debiased Object Correspondence Analysis

David Windridge[a,], *Fei Yan*[b]

[a] *Computer Science Dept., Middlesex University, The Burroughs, Hendon, London NW4 4BT*
[b] *Faculty of Engineering and Physical Sciences, University of Surrey, Guildford, Surrey GU2 7XH, UK*

Abstract

This paper addresses the problem of combining multi-modal kernels in situations in which object correspondence information is unavailable between modalities, for instance, where missing feature values exist, or when using proprietary databases in multi-modal biometrics. The method thus seeks to recover inter-modality kernel information so as to enable classifiers to be built within a composite embedding space. This is achieved through a principled group-wise identification of objects within differing modal kernel matrices in order to form a composite kernel matrix that retains the full freedom of linear kernel combination existing in multiple kernel learning. The underlying principle is derived from the notion of tomographic reconstruction, which has been applied successfully in conventional pattern recognition.

In setting out this method, we aim to improve upon object-correspondence insensitive methods, such as kernel matrix combination via the Cartesian product of object sets to which the method defaults in the case of no discovered pairwise object identifications. We benchmark the method against the augmented kernel method, an order-insensitive approach derived from the direct sum of constituent kernel matrices, and also against straightforward additive kernel combination where the correspondence information is given *a priori*. We find that the proposed method gives rise to substantial performance improvements.

Keywords: classifier combination, support vector machines, kernel

Email address: d.windridge@mdx.ac.uk, Tel: +44 (0)1483 686048, Fax: +44 (0)1483 686031 (David Windridge)

methods, tomography

1. Introduction

The problem of multiple kernel learning (MKL) was identified by Lanckriet *et al.* [1] and is now well established within the literature [2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. It builds on the widespread adoption of kernel-based methods within machine learning for a variety of tasks, in particular regression and classification [12, 13]. The latter category includes state-of-the-art methods such as support vector machines (SVMs) [14, 12] and kernel Fisher discriminant analysis (kernel FDA) [15, 16].

Kernel methods have in common that they map observations into an inner product space, provided that they fulfil the Mercer conditions. A wide choice of kernels is typically available for any given learning problem; each of these kernels can be seen as capturing a different aspect of the data. In classification problems, arbitrarily morphologically-complex (*i.e.*, nonlinear) decision boundaries may be obtained within a linear input space via the choice of kernel. Early work on learning the kernel includes [17], where kernel parameters are optimized by minimizing estimates of the generalization error of SVMs, and [18], where the complexity of learning the kernel matrix for SVM classification is analyzed.

Multiple kernel learning seeks to learn an appropriate linear combination of such base kernels, linear combination being chosen because this crucially retains the Mercer properties. Lanckriet *et al.*'s formulation [1] utilizes linear combination of M $m \times m$ training kernel matrices $K_k, k = 1, \ldots, M$ and m class labels $y_i \in \{1, -1\}, i = 1, \ldots, m$, with m the number of training samples, this being equivalent to forming the Cartesian product of the associated feature spaces. The goal of MKL is then to optimize the 'scaling factors' of the feature spaces with respect to the classification. Other MKL formulations address tractability issues when m is large. These include, e.g., the semi-infinite linear programming (SILP) formulation of [3], and the reduced gradient descent algorithm of [6]. The ℓ_1 regularization in [1] can also be generalized to an ℓ_p $(p > 1)$ norm [19] to avoid solution sparsity if required. Other variants of MKL approaches include, to name a few, hyperkernels [20], information theoretic MKL [21], multiple kernel FDA [22, 23], multiclass MKL [4], multilabel MKL [24] and nonlinear MKL [25].

A key distinction that may be made between multiple kernel methods is whether they implicitly require object correspondence information; additive kernel combination such as the method of Lanckriet *et al.* assumes that this

information is present. Thus, the ordering of the objects defining the K_k is assumed to be the same across all modalities. However, methods do exist that are not dependent on this correspondence, the principle such method being *augmented kernel combination* [26]. In augmented kernel combination, the direct sum of kernel matrices is formed, resulting in a block-diagonal kernel matrix (i.e. so that all of the constituent kernel matrices are embedded along the diagonal of the resultant matrix, with all inter-kernel values set to a value of zero); [26] compares the geometric interpretation of linear combination and augmented kernel combination. It is shown in [27] that augmented kernel combination is closely related to classifier fusion.

In general, the problem domain will determine whether object correspondence information is available. For instance, it is not uncommon in multimodal biometrics to obtain distinct sets of exemplar subjects for each individual biometric measurement (*e.g.*, iris scans, finger prints, photographic images), particularly when employing separate commercial sources [28]. In this case, we would wish to utilize the information collectively contained within each data set for a given test subject, but would lack object correspondences in the collective set of multi-modal data sets. In other words, we have object correspondence in the test set but not the training set. The augmented kernel approach to classification of individual test subjects in this case would be to build a composite kernel matrix via the direct sum of kernel matrices associated with each modality and then utilize this, in combination with a corresponding vector of class labels, for classifier training. (The direct sum kernel matrix is order-insensitive with regard to the training objects within individual modalities provided that the class label vector is correspondingly permuted).

However, the argument of this paper is that such methods, by omitting the possibility of re-deriving correspondence information, potentially overlook important classification information. To address this, we propose a kernel-based adaptation of a method developed for standard non-kernelized pattern recognition that is capable of bringing about this correspondence[1]. The resulting method for multiple kernel learning gives rise to a kernel matrix that defines an appropriate composite embedding space that, as nearly as possible, approximates the Kernel matrix that would exist if all object correspondence information were available. It does so by removing the bi-

[1]Thus, our method is an MKL method to the extent that it proposes a linear sum of kernels to be optimized. However, the method of generating these kernels is by no means linear.

asing factors associated with linear methods of kernel combination.

1.1. Linear combination bias in non-kernelized pattern recognition

In conventional (*i.e.*, non-kernelized) pattern recognition, it may be demonstrated [29, 30] that linear classifier combination methods impose a bias on the composite decision space formed by decision combination[2] (by "decision space" we here mean the space in which the decision boundary is formed). This bias comes about via the limitations of linear combination in dealing with correlated information in the marginal classifiers (i.e the feature-selected classifiers constituting the combination), and prevents the optimal decision boundary being constructed, leading to suboptimal overall performance. We thus consider the classifiers within a combination as representing, to some degree of approximation, the marginal distributions of the composite pattern space in which the decision boundary is formed (see Section 2.1 for a pictorial example of this process; in the remainder of the Introduction we give a qualitative account).

This biasing behavior occurs, for instance, when feature selection is applied to an input space of arbitrary dimensionality, S, such that a set of classifiers (indexed by $i \in I$) become associated with non-coincident (*i.e.*, non-overlapping) feature sets that collectively span S (or a subset of it). In such cases, classifier combination effectively acts to combine, in the original input space, the set of orthogonal marginals distributions that are implicitly modelled within the individual classifiers, i (modelling need not be exact, e.g. in the case of discriminative classifiers; see Section 2.1 for an example with artificial neural networks).

A similar situation exists in multi-modal fusion problems, where modalities may equally be regarded as the features of some composite decision space, allocated to specific classifiers associated with the modalities. It was the effort of [29] to demonstrate that this bias is specifically a form of *sampling bias*. The bias attributable to linear combination methods within the composite space is thus due to the mismatch of the very low number of angular samples of the composite decision space (equivalent in magnitude to $|I|$) created by the orthogonal 'marginal' distributions of the feature-selection process in comparison to their linear sampling rate. (The linear sampling rate equates to the total number of distinguishable input vectors)[3]. However,

[2]This applies in situations in which it can be reasonably assumed that there exists no *a priori* restriction on density distributions in the decision space, for instance, prior knowledge of feature independence.

[3]Note this only represents *combination* bias; classifier bias also contributes. *cf.* [30]

to fully represent arbitrary distributions in the composite space, angular and linear sampling would have to be of the same order (the orthogonal nature of this angular sampling is depicted in Section 2.1).

This mismatch between angular and linear sampling of the composite decision space suggests an analogy with tomography theory, for which the component classifiers of the combination essentially represent Radon-projections (linear integrals) of the composite decision space. Linear combination then acts as the inverse operation to Radon-projection, *i.e.*, *back projection* (essentially a summation over the Radon Projections that intersect at the point of reconstruction). However, in tomography theory back-projection only recovers a *biased* simulacra of the original unprojected composite space (the outcome of back-projection being the original distribution in the space *convolved* with an artefact defined by the angular frequency of the Radon sampling). The process of tomography is thus concerned with the pre- or post-combination filtration of this artefact in order to remove the sampling bias.

Similarly, this bias is represented within tomographic classifier combination theory as a convolution of the true underlying distribution of pattern vectors (denoted F_{true}) in the decision space with an artefact (denoted F_{samp}) deriving from the sampling (F_{true} and F_{samp} are thus density distributions defined over the entirety of S). The 'recovered' density distribution induced by classifier combination is thus $F_{comb} = F_{true} \star F_{samp}$, with $\star$ the convolution operation. F_{samp} is thus defined in the composite space by the response of an origin-centered Dirac delta function, firstly to representation as a series of individual Dirac delta functions in the marginal spaces associated with each classifier, and secondly to the action of the combination rule that reconstructs an 'image', F_{comb}, of the original Dirac delta function within the composite space. That is, F_{samp} is what is obtained if one were to take a single pattern vector from the true underlying distribution of pattern vectors in the decision space, represent it within the individual classifiers via feature-selection, and then 're-project' it back onto in the decision space by applying the combination rule. The resulting entity formalizes the systematic convolutional 'bias' introduced by the combination rule: convolutional biasing is thus a manifestation of the inevitable failure of the combination rule to interpolate the missing distribution information lost by feature se-

for a fuller discussion of the bias/variance breakdown under this paradigm. See also [31, 32, 33] for a general discussion of bias-variance-covariance decomposition in classifier ensembles.

lection.

As part of the theoretical development of this paper, we will, in the next section, formalize this process for the sum rule decision scheme before demonstrating how the tomographic fusion principle may be explicitly kernelized and applied to the problem of kernel fusion. In doing so, we note the similarity of this approach to the mechanism (though not the application) of the pyramid match kernel (PMK), in that it involves a bottom-up pairwise identification of entities in a kernel context. However, note, that while the PMK is used to generate a positive-semidefinite (PSD) kernel matrix from a set of features contained within different objects, the current algorithm combines kernel *matrices* associated with differing modalities into a single, inter-modal kernel matrix.

The remainder of this paper is thus organized as follows. In Section 2, we set out the tomographic fusion methodology as applied to the sum rule decision scheme, and demonstrate how it relates to morphological correspondence. In Section 3 we demonstrate how this method can be applied to kernel spaces and, in doing so, set out a novel algorithm for multiple kernel combination. Section 4 then applies this method to a series of experimental scenarios, firstly on an illustrative simulated domain, and secondly on a range of data sets obtained from the UCI machine learning repository [34], and from the CAL500 semantic retrieval problem. In Section 5 we discuss the relative performance of the method, and conclude in Section 6 with a summary of achievements.

2. Morphologically-Unbiased Classifier Fusion

2.1. Pictorial example

We illustrate the process of morphologically-unbiased classifier fusion with a simple example. Suppose a multi-modal dataset consists of just two modalities (with each modality consisting of one real-valued feature), such that the totality of the data can be represented within a composite two-dimensional pattern space. Suppose that the underlying distribution of pattern vectors for single classes in this space are well-separated unimodal two-dimensional multivariate distributions (e.g. Gaussians). Further suppose that these individual modalities (labelled x and y) are represented by individual perceptron classifiers (or equivalently that the composite 2D feature space has been allocated to two distinct perceptron classifiers with non-overlapping feature-sets following feature-selection). The 2D data distribution is thus 'integrated' along each of two axes in order to be represented within the two 1-D classifiers.

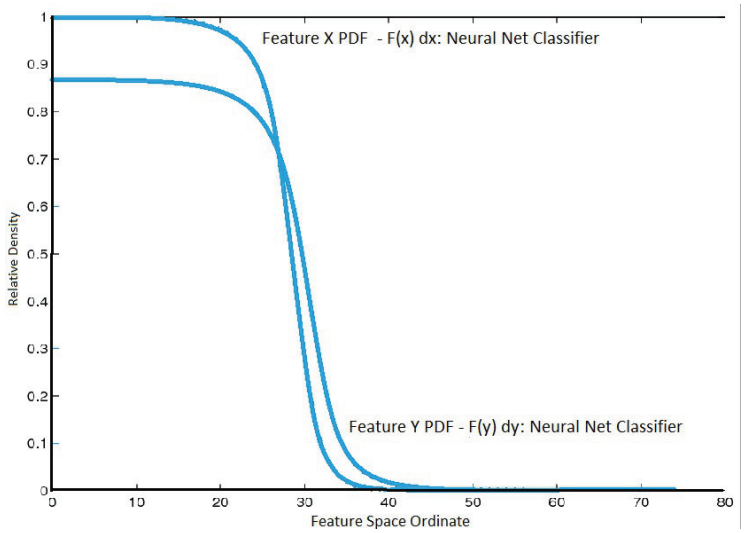

Figure 1: Sigmoid pseudo-density functions for each individual perceptron classifiers representing the two feature ordinates labelled x and y.

We illustrate the resulting sigmoid activation functions, $F_x^l(\mathbf{x})$ and $F_y^l(\mathbf{y})$, in Figure 1 for the single class, l. Thus, although perceptrons are primarily discriminative, $F_x^l(\mathbf{x})$ and $F_y^l(\mathbf{y})$ are treated as density approximating functions throughout the following: clearly explicitly generative classifiers will exhibit a closer correspondence to the underlying density distribution. It is then apparent that a straightforward linear summation combination rule:

$$class\_label(x, y) = \arg\max_l \left(F_x^l(\mathbf{x}) + F_y^l(\mathbf{y})\right)$$

will generate an implicit class density function, $F_{comb}^l(x, y) = \left(F_x^l(\mathbf{x}) + F_y^l(\mathbf{y})\right)$ within the decision space S as depicted in Figure 2 (left). (The decision space is equivalent to the original composite pattern space in this example). Thus, the final class allocation in the decision scheme will be $l_{chosen} = \arg\max_l \ F_{comb}^l(x_1, y_1)$ for a test pattern vector with a value $x = x_1$ in the first modality and a value $y = y_1$ in the second modality).

However, it is evident that the distribution $F_{comb}^l(x, y)$ by no means resembles a unimodal two-dimension multivariate Gaussian; the distribution $F_{comb}^l(x, y)$ exhibits long extensions of excess density along the axes. The reason for this can be understood in terms of tomography theory [35]; $F_x^l(\mathbf{x})$ and $F_y^l(\mathbf{y})$ approximate *Radon transformations* of the underlying distribution $F_{true}^l(x, y)$ (*i.e.*, they are projections along the axis). In standard to-

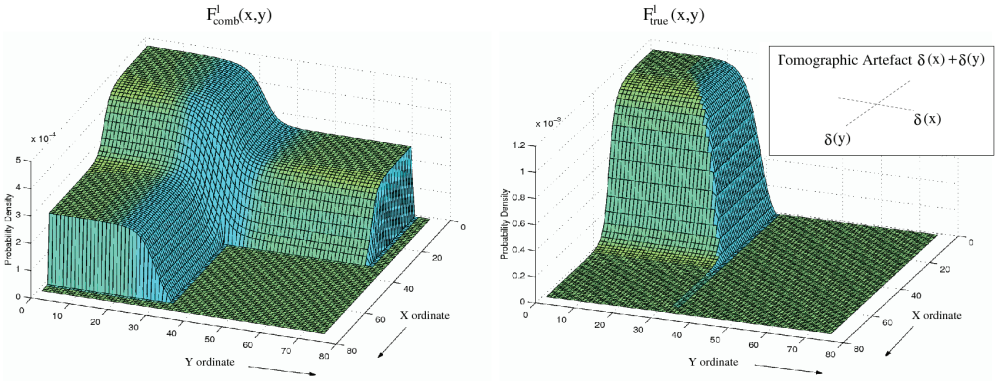

Figure 2: left: composite class density function, $F^l_{comb}(x, y)$, right: the modified composite density function $F^l_{comb}(x, y)$ after deconvolution.

mographic reconstruction, the inverse of the Radon transformation process is the back projection operation, which is equivalent in this case to generating the 2D function $F^l_x(\mathbf{x}) + F^l_y(\mathbf{y})$. However, back projection introduces an artefact into the reconstruction that is defined by the angular sampling rate of the original Radon transformation (*i.e.*, the number of Radon 'slices'). In our case this angular sampling rate is extremely low; we are integrating (*i.e.*, Radon transforming) only along the two orthogonal axes defined by the ordinates x and y.

There is thus introduced an artefact, $F_{samp}(x, y) = \delta(\mathbf{x})dx + \delta(\mathbf{y})dy$, into the reconstruction (see inset of Figure 2, right). This artefact acts to 'blur out' the attempted reconstruction of $F^l_{true}(x, y)$ via convolution. In particular, it acts to blur the attempted reconstruction along the axes of integration.

Any truly unbiased reconstruction of the 2D space thus requires that we deconvolve out this artefact. Figure 2 (right) illustrates the modification of the composite density function $F^l_{comb}(x, y)$ when this deconvolution takes place. It is immediately evident that the resulting distribution more closely resembles the underlying unimodal two-dimension multivariate Gaussian; what we have achieved by the deconvolution is to eliminate all traces of axial bias from the decision space (*i.e.*, the projection along the axes). Note that the deconvolved space still gives an identical marginal projection to the undeconvolved space, so that $F^l_x(\mathbf{x})$ and $F^l_y(\mathbf{y})$ are unaffected by the deconvolution. Typically, when carried out for all classes, this results in significantly improved classification performance (though obviously the method will fail if the underlying distribution *does* exhibit axial bias; however this

tends not to occur when modalities are of intrinsically different kind). In the following, we derive this more formally. (We will continue to derive the methodology in terms of the Sum Rule decision scheme, both because this exemplifies linear combination methods, but also because it serves to simplify the mathematics; note, however, that *any* linear combination rule will give rise to axial artefacts).

3. Kernelization of Morphologically Unbiased Fusion

The technique for morphological debiasing of classifier combination discussed in the previous section is explicitly density-centric, and might on first inspection appear inapplicable to multi-kernel approaches, for which the embedding space is collectively defined by the component kernel matrices, $K_1, K_2 \ldots$, rather than given *a priori*. However, morphologically unbiased combination, at its most basic, relies upon the principle that there exists some composite decision space composed of component ('marginal') spaces relating to the individual input classifiers, each of which perhaps associated with a distinct modality. It further assumes that the typically orthogonal angular sampling rate implicit in this marginalization is significantly smaller than the linear sampling of the input space (essentially, the potentially infinite number of distinct pattern vectors for which the classifier provides an output).

Both of the above are true of multi-kernel learning problems for which the composite Reproducing Kernel Hilbert Space (RKHS) is generally the (weighted) Cartesian Product of the component RKHSs in order to comply with the Mercer constraints. Note that the notion of sampling in this context assumes that the RKHS are equipped with a Kernel norm. Throughout the following we shall, for clarity, consider only the unweighted kernel fusion problem; the presence of weight coefficients does not affect the methodology of combination, other than by introducing an additional 'scaling factor' to the marginal components.

Morphologically unbiased combination also assumes that there is no *a priori* reason to suppose that the angular sampling implied by the marginalization process has any *intrinsic* relation to the underlying morphology of the distribution of vectors within the decision space. While multiple kernel learning does not employ the notion of distributions of pattern vectors existing within a pre-existing space, as such, the implicit construction of a decision space via the Cartesian product of the component embedding spaces does not itself depend on the morphology of those embedding spaces. Consequently, the orthogonal angular sampling implied by the formation of

the Cartesian product of the component kernels' embedding spaces is independent of the values within the component kernel matrices $K_1, K_2, \ldots$ provided that these are consistent with the formation of the same embedding space (*i.e.*, they have the same Mercer features $(\lambda_1^{\frac{1}{2}} u_1, \lambda_2^{\frac{1}{2}} u_2, \ldots)$ (or some permutation of these) under the eigenvalue decomposition, $K_n = U\Lambda U'$: where $U = (u_1, u_2, \ldots, u_r)$ and $\Lambda = diag(\lambda_1, \lambda_2, \ldots))$.

We thus assume that there is no *a priori* restriction on object distributions within the composite embedding space arising from kernel composition. This is generally true in *multi-modal* kernel fusion problems where there is no reason to suppose that embedding spaces associated with one modality are *intrinsically* related to those of another modality (*specific* kernel measurements are still free to exhibit intra-modal correlations, however)[4].

The underlying motivations behind morphologically unbiased combination classifier combination thus remain valid within a kernel-based context. Moreover, by removing the explicit requirement for reconstruction of a density distribution in the composite space, we shall show that it is possible to maximize the efficiency of the method.

The goal of tomographic kernel fusion is thus to generate a novel kernel matrix, K_{comp}, from the input matrix for each modality $K_1, K_2, \ldots, K_M$ (along with their associated class labelings), such that a classification algorithm (typically an SVM) can act in the most effective manner on the composite data, optimally taking advantage of correlations within the data. We further wish to do this in a way that does not generate an excessive number of objects within K_{comp}, in contrast to the tensor product approach.

In the following subsection, we give a breakdown of the key algorithmic stages of morphologically-unbiased kernel combination, with full pseudocode set out in Algorithm 1.

3.1. Algorithmic approach to morphologically unbiased combination kernel combination

3.1.1. Initialization requirements

First, the Kernel matrices is split for each modality $K_1, K_2, \ldots, K_M$ into class l specific components: $K_1^l, K_2^l, \ldots, K_M^l$, $1 \leq l \leq L$, after which a map

[4]These assumptions are less certain, though not necessarily invalid, in the case of multiple kernels measurements applied to the *same* data sets. Thus, while the morphologically unbiased combination fusion may be applied without restriction to any kernel combination problem, we expect its advantages to hold predominantly in the multi-modal kernel fusion domain.

$f_i(l,j)$ is defined that takes an object index j in kernel matrix l to its original index in K_i. Thus, for a set of objects, $x^{m,l} \in X$, and class labels, L, such that $c : X \to \{l \in L\}$ we obtain the kernel matrices K_n^l defined such that:

$$K_m^l(x_i^{m,l}, x_j^{m,l}) = K_m(x_i^{m,l}, x_j^{m,l}) \quad iff \quad c(x_i^{m,l}) \to l \wedge c(x_j^{m,l}) \to l$$
$$\forall l, \forall i, \forall j, 1 < m < M, m \in \mathcal{I}$$

Note that, defining the magnitude of a kernel matrix, $|K|$, to be the number of objects that it indexes, we shall assume for simplicity throughout the following that $\forall l, l', m, m', |K_m^l| = |K_{m'}^{l'}|$. The method, however, does not depend on this assumption. This class separation is necessary in the following since we do not expect, *a priori*, observations to exhibit inter-class kernel dependencies. Thus $K_m(x_i^{m,l}, x_j^{m,l})$ tells us nothing about $K_m(x_k^{m,l'}, x_l^{m,l'})$ intrinsically. Note however that we *will* need these relations in order to build the final output matrix K_{comp} of the procedure, since inter-class kernel dependencies define, in part, the Mercer embedding space of K_{comp}. Note that all of the following there is no explicit dependency on object ordering within the K_m; the method assumes no object correspondence at the outset. Obviously, if this information exists, it can be directly utilized as $K_{\text{output}} = \Sigma_m K_m$.

A squared distance matrix is then defined via the standard kernel norm independently for each modality and class:

$$D_m^l(x_i^{m,l}, x_j^{m,l}) = K_m^l(x_i^{m,l}, x_i^{m,l}) + K_m^l(x_j^{m,l}, x_j^{m,l}) - K_m^l(x_i^{m,l}, x_j^{m,l}) - K_m^l(x_j^{m,l}, x_i^{m,l}) \quad (1)$$

where symmetry implies that this only need be explicitly calculated for $i \geq j$. This matrix is treated throughout the following as being indicative, within the bounds of stochastic variability, of the (class-wise and modality-wise) inverse-squared pairwise density distribution within the embedding space.

For each $D_m^l(x_i^{m,l}, x_j^{m,l})$, an index set of ordered[5] object pairs

$$\{S_m^l(t_m^l) = (x_i^{m,l}, x_j^{m,l})_t : 1 \leq t \leq |K_m^l|^2\}$$

is then defined such that

$$D_m^l(S_m^l(t_m^l)) \geq D_m^l(S_m^l(t_m^l - 1)), \forall t_m^l, t_m^l > 1.$$

[5]The accumulated sort procedures introduce a time penalty of $O(\sum_l \sum_m |K_m^l|^2 log|K_m^l|^2) \approx O(\sum_m |K_m|^2 log|K_m|^2)$ for typical sort algorithms, reducible to $\approx O(max(|K_m^l|^2) log(max(|K_m^l|^2))) \approx O((|K_m|^2/L) log(|K_m|^2/L))$ if executed in parallel.

t_m^l thus varies over the indices of object pairs ordered in terms of their density in the embedding space. Consecutive indices may relate to the same density value (*i.e.*, if they are degenerate with respect to D_m^l), in which case the ordering of the index may be arbitrarily permuted over these values without consequence.

To initialize the iterative procedure, the set of object pairs that are degenerate with the density maxima is obtained for each modality and class in the same manner as the marginalized morphological correspondence algorithm of [35]. That is, for all l and m, we obtain the sets: $\{S_m^l(t_m^l) : S_m^l(t_m^l) = S_m^l(0)\}$. Note that zeros on the leading diagonals of the D_m^l are ignored in the above and throughout the following.

The morphological debiasing algorithm treats density maxima that occur within the distinct modalities as correspondent, *i.e.*, such that they collectively *co*-ordinate maxima in the composite decision space. The maximal set of such ordinal correspondences obtained from the density maxima of the different modalities is given via the Cartesian product over all of the modalities:

$$S^l(0) = \bigotimes_m \{S_m^l(t_m^l) : S_m^l(t_m^l) = S_m^l(0)\}$$

$S^l(0)$ thus consists of ordered sets of the form $\left(x^{1,l}, x^{2,l}, \ldots, x^{M,l}\right)$, where $x^{m,l}$ is either component of the pair $(x_i^{m,l}, x_j^{m,l})_t$ when t is in the considered density band – here the lowest density band – for each of the modalities. Note that the Cartesian product here is over the *constituents* of the $S_m^l(t_m^l)$, such that $S_m^l(t_m^l)$ is treated as a set rather than an ordered pair.

This implicitly establishes a set of object correspondences, $x^{m',l} \rightarrow x^{m'',l}$, across the differing modalities m' and m'' when $x^{m',l}$ and $x^{m'',l}$ co-occur within the ordered sets S^l. When there are several ordered sets for which this occurs (*i.e.*, if for any m $|\{S_m^l(t) : S_m^l(t) = S_m^l(0)\}| > 1$), then there is a ambiguity of association, and object pairs within each modality have a *set* of object pairs with which they are associated[6]. In the ideal case, *i.e.*, $|\{S_m^l(t) : S_m^l(t) = S_m^l(0)\}| = 1 \ \forall m$, object pairs in the different modalities exist in a one-to-one correspondence. (Of course, this must apply at every single iteration of the association procedure for this there to be perfect pairwise correspondence across each modality). Ambiguity of association thus serves to increase the intrinsic size of the kernel matrix derived from

[6]Note that using pairwise object correspondences also introduces an intrinsic ambiguity of association, such that m distinct pairs in each modality gives rise to 2^m points within the composite space.

the procedure (although we shall discuss methods for reducing this). In the asymptotic case of one-to-one object correspondence, the output kernel matrix of the procedure would have the same magnitude as the input matrix, *i.e.*, $|K_{comp}| = |K_m|$. (Note that, whatever the final size of $|K_{comp}|$, we always utilize all of the available object correspondences; i.e. we discard none of the available data).

The above procedure thus equates to the compilation of the Cartesian product of the highest density *level sets* of the marginal density distributions. We have not yet, however, constructed the kernel relations implicit in this notion; at present we are considering *only* the establishment of object correspondences.

3.1.2. Iterative Object Correspondence Determination

In order to utilize the above initialization within an iterative context, the set $S^l(0)$, with 0 indicating that this is the first iteration, is added to what will become the cumulative set of correspondences, $\mathcal{S}^l$ (note caligraphization):

$$\mathcal{S}^l(n) = \mathcal{S}^l(n-1) \bigcup S^l(n), \quad n \in \mathcal{I}$$

The 'marginalized' per-modality set of object pairs for which correspondences have been established, $\mathcal{S}^l_m$ is also cumulatively aggregated:

$$\mathcal{S}^l_m(n) = \mathcal{S}^l_m(n-1) \bigcup \{S^l_m(n) \in S^l(n)\}$$

Individual ordinal counters k^l_m are then instigated for each separate modality and class, which are able to increment independently with each iteration number n. $k^l_m(n)$ is thus the value of the ordinal index at iteration n, such that $S^l_m(k^l_m(n))$ is the $k^l_m(n)$'th ordinal pair, *i.e.*, $(x_i^{m,l}, x_j^{m,l})_{k^l_m(n)}$. This will allow the algorithm to accommodate differingly-sized density degeneracy sets at each iteration, and corresponds to the $\Delta z_x^{(n)}$, $\Delta z_y^{(n)}$ variation that we employed in the marginal version of morphological correspondence algorithm.

A reference counter, nominated as that of modality $m = 1$, *i.e.*, $k^l_1(n)$, is increased to the index value immediately above that of the highest value in the degeneracy set obtained by the previous iteration of the kernelized Högbom procedure[7]. It has an initial value of $k^l_1(0) = 0$, and is set after

[7]It is also possible to subdivide the reference counting if too many degenerate values are found within the Kernel matrices (see experimental section).

each iteration, to a value

$$k_1^l(n) = |\{S_1^l(t) : S_1^l(t) = S_1^l(1)\}| + k_1^l(n-1)$$

The counters of the remaining modalities, *i.e.*, $k_m^l(1),\ m \neq 1$, are then incremented until the density value closest to the reference density value is obtained[8]. Thus:

$$k_m^l(n) = \arg\min_t \left\{ |D_m^l(S_m^l(t)) - D_1^l(S_1^l(k_1^l(n)))| \right\},\ \forall m \neq 1$$

Note that, where the minimum values are degenerate, the lowest $k_m^l(n)$ value is selected.

We denote the set of ordered object pairs within a given modality that falls within this density band as $T_m^l(n)$. Thus:

$$T_m^l(n) = \{S_m^l(t) : k_m^l(n-1) < t \leq k_m^l(n)\}$$

The $T_m^l(n)$ are then the kernelized equivalent of the level sets of the marginalized Högbom deconvolution algorithm in [35]. Thus, for the iterative update we have that: $S^l(n) = \bigotimes_m T_m^l(n)$.

The Cartesian product inherent in $S^l(n)$ establishes object correspondences via the density correlations between modalities. However, once density maxima correlations have been established (*i.e.*, peaks have already been established in the class distribution in the composite embedding space), any sequence of object pairs, $s' \in S^l(n+1)$, at an incrementally higher density level, $n+1$, can in principle be associated with sequences of object pairs contained within $S^l(n)$ simply via the topological connectivity implicit in the definition of s and $S^l(n)$. In this case there is no ambiguity in inter-modal object association, and the Cartesian product need not be carried out[9]. We model this as follows:

For all modalities m, let the set of first and second objects of all the ordinal pairs in $T_m^l(n)$ be denoted $T_m^{l,1}(n)$ and $T_m^{l,2}(n)$, respectively. $T_m^{l,1}(n)$ is thus the set of lowest-valued object indices in the set of object pairs that make up $T_m^l(n)$. (Recall that the associated object pairs $(x_i^{m,l}, x_j^{m,l})_t$ are ordered with the lowest index value first).

[8] This is in principle a logarithmic $O(log|K_m^l|)$ time-complexity process (binary search), though incremental search is generally sufficient given the relatively small reference index increment at each iteration.

[9] Such points are topologically attached to an existing density maxima, and are therefore exempt from the Cartesian multiplication with other modalities that gives rise to novel peaks in the first place, *i.e.*, there is no ambiguity as to their inter-modal correspondence.

Let $g_m^l(n)$ be an single element of $T_m^{l,1}(n)$ for each modality, such that the composite $(g_1^l(n), g_2^l(n), \ldots, g_M^l(n)) \in G^l$ represents a coordinate of the composite embedding space for all modalities. A corresponding point can be derived from the second objects of the ordinal pairs:

$$(h_1^l(n), h_2^l(n), \ldots, h_M^l(n)) \in H^l \quad \text{such that} \quad h_m^l(n) \in T_m^{l,2}(n)$$

Provided that the first point is an element of $\mathcal{S}^l(n)$:

$$(g_1^l(n), g_2^l(n), \ldots, g_M^l(n)) \in \mathcal{S}^l(n)$$

then the second point can be added to the set $\mathcal{S}^l(n)$, *i.e.*,

$$\mathcal{S}^l(n) \bigcup (h_1^l(n), h_2^l(n), \ldots, h_M^l(n))$$

and the pairs $(g_m^l(n), h_m^l(n))$ can be removed from $T_m^l(n)$ for each modality:

$$T_m^l(n) = T_m^l(n) / (g_m^l(n), h_m^l(n))$$

The Cartesian product, $S^l(n) = \bigotimes_m T_m^l(n)$, thus involves fewer elements and consequently less ambiguity of object association.

Note that the newly added point is a modality-independent maximum distance of $(\sum_m D_1^l(k_1^l(n)))^{\frac{1}{2}}$ from the topologically connected point in the embedding space, and the separate index counters, $k_1^l(n)$, are designed to ensure equal density sampling in the embedding space. Thus we implicitly employ the l^1-norm for determining topologically-connected points; however a further l^2-norm constraint can be applied such that peaks in the embedding-space class density distribution are established and aggregated in a radially symmetric manner; this would be the equivalent of employing a fully density-based morphologically unbiased combination model.

The reference counter $k_1^l(n)$ is incremented according to the above criteria, and the two processes of Cartesian product multiplication and topological association are continuously iterated until the indices all achieve their maximum wherein all pairs are sampled. The output of the iterative process is a complete set of object correspondences across the various modalities for each class, *i.e.*, $\mathcal{S}^l(n_{\max})$.

3.1.3. Reconstruction of Composite Kernel Matrix

These object correspondences are transformed back into a Kernel matrix in the following way. First, a class-index vector $\mathcal{C}$ is compiled for each

meta-object, $O = \left(x^{1,l}, x^{2,l}, \ldots x^{M,l}\right)$ in $\mathcal{S}^l(n_{\max})$ generated by the twin processes of Cartesian multiplication and object/peak association for each of the classes in sequence. Thus $\mathcal{C}(t') = l$ if

$$t' = \Sigma_{k=1}^{k=l-1} |\mathcal{S}^k(n_{\max})| + t$$

where

$$\left(x^{1,l}, x^{2,l}, \ldots x^{M,l}\right)_t \in \mathcal{S}^l(n_{\max})$$

Next, the morphologically unbiased kernel matrix is compiled for each meta-object across all of the classes. Recall that, although separately constituted, the different class distributions exist within the same embedding space. Thus, exploiting the fact that Mercer properties are preserved under additivity, and assuming coefficients of unity:

$$K_{comp}(O_I, O_J) = \Sigma_m K_m(x_I^m, x_J^m)$$

where

$$1 < \{I, J\} < |\mathcal{C}|, \quad O_I = \left(x_I^1, x_I^2, \ldots, x_I^m\right)_I$$

The x_I^m are thus the objects within modality m re-indexed by the class-concatenated variables I, J (implicitly utilizing the mapping functions $f_i(l, j)$ involved in the original class separation). This typically gives rise to a kernel matrix K_{comp} of magnitude:

$$sup(|K_m|) \leq |K_{comp}| << e^{\Sigma_m ln|K_m|}$$

with actual size dependent on the data.

This matrix may then be used alongside the class labelings contained within $\mathcal{C}$ for, *e.g.*, SVM training in the normal manner. If a test matrix is required (such that test points are contained within the same embedding space as the support vectors of the SVM), then this may be compiled in the same manner by utilizing the fact that the test objects can be embedded in each of the individual modality's embedding spaces, *i.e.*,

$$K_{test}(O_i^{\text{test}}, O_J) = \Sigma_m K_m^{\text{test}}(x_i^{m,\text{test}}, x_J^m)$$

where the x_{test}^m are test objects indexed via i, and K_m^{test} is the test kernel matrix generated by forming the kernel product of the test and training data. Note that the each index I value in x_I^m does not necessarily correspond to the same object in each modality m.

A further refinement of the method permits sub-sampling of the Cartesian products generated during the procedure. Thus, if during iteration

n calculation of the product $S^l(n) = \bigotimes_m T_m^l(n)$ is indicated, a randomly-selected subsample of the product can instead be obtained with $> \Sigma_m |T_m^l(n)|$ samples (this sampling minimum is chosen so as to preserve marginal sample rates). The subsampling procedure is thus designed so as to give rise to an equivalent areal coverage of generated points to that of the full Cartesian product method, and thus no significant performance degradation is expected (identical support vectors should be selected). Such sub-sampling helps to ensure that K_{comp} does not become excessively large in that case of minimal morphological correlation.

Algorithm 1 sets out pseudocode for the above method. Note we have for simplicity assumed that the K_m are either intrinsically of similar value ranges, or else have been normalized so as to have similar or identical value ranges. However the sort-based nature of the method means that this is not a critical consideration.

Algorithm 1 will thus solve any l-class, m-modality problem for which the inputs are the inter-modality kernel matrices, K_m, the inter-modality test matrices, K_m^{test} and the class labels $c(x_j^{m,l}) \rightarrow l$. Outputs are the composite kernel matrix K_{comp}, the class labels $\mathcal{C}$ for the constructed objects in K_{comp}, and the test kernel matrix K_{test} defined within the embedding space of K_{comp}.

3.2. Complexity Considerations

The order of complexity of the proposed method can be broken down in terms of the three main components of the procedure as follows.

3.2.1. Initialization

Initialization consists of the following series of processes, with respective complexities given in square brackets: *class-based matrix splitting* $[O(|K|)]$, *a squared kernel distance calculation* $[O(m|K|^2)]$, *distance sorting* $[O(m|K|^2 log(|K|^2))]$ and, finally, a *linear search for degenerate densities* $[O(m\ log(|K|^2))]$. (Note that we do not need an explicit co-ordinatization step of order $O(|K|^{2m})$ since it is only required that we *identify* the degenerate sets within each modality constituting the initial coordinate m-tuples; we do not calculate each point individually at this stage).

Initialization is thus dominated by the $O(m|K|^2 log(|K|^2))$ sorting process.

3.2.2. Main Loop

The main loop essentially carries out a series of n inter-modally co-dependent searches within each modality's sorted list of kernel distances

Algorithm 1 PseudoCode for Morphologically unbiased kernel fusion

initialization

For each modality, split the kernel matrix K_m into l classes with independent mapping indices;

Compile squared kernel norm matrix for each modality m and class l with zero-valued lead diagonal, using Equation (1) and $D_m^l \rightarrow D_m^l - D_m^l I$;

Find the set of density-sorted object pairs $S_m^l(t_m^l)$ degenerate with the density maxima: *i.e.*, $\forall l, m$ obtain: $\{S_m^l(t_m^l) : S_m^l(t_m^l) = S_m^l(0)\}$;

Specify an initial peak set via the Cartesian product $S^l(0) = \bigotimes_m \{S_m^l(t_m^l) : S_m^l(t_m^l) = S_m^l(0)\}$: set counters $n = 1$, $k_1^l(-1) = 0$;

main loop

while $k_1^l(n) \leq |K_m^1|$ **do**

  Set $k_1^l(n) = |\{S_1^l(t) : S_1^l(t) = S_1^l(1)\}| + k_1^l(n-1)$;

  $\forall m,\ m \neq 1$ find $k_m^l(n)$ for which $|D_1^l(S_1^l(k_1^l(n))) - D_m^l(S_1^l(k_m^l(n)))|$ is a minimum;

  Associate object correspondences within density level set to existing peaks if topologically connected, *i.e.*, if $(g_1^l(n), g_2^l(n), \ldots, g_M^l(n)) \in \mathcal{S}^l(n)$, then $\mathcal{S}^l(n) = \mathcal{S}^l(n) \bigcup (h_1^l(n), h_2^l(n), \ldots, h_M^l(n))$;

  For all m, remove connected pairs $(g_m^l(n), h_m^l(n))$ from $T_m^l(n)$, *i.e.*, $T_m^l(n) = T_m^l(n)/(g_m^l(n), h_m^l(n))$ where $T_m^l(n) = \{S_m^l(t) : k_m^l(n-1) < t \leq k_m^l(n)\}$;

  Form the Cartesian product: $S^l(n) = \bigotimes_m T_m^l(n)$ (Sub-sampling if required);

  Update correspondence set: $\mathcal{S}^l(n) = \mathcal{S}^l(n-1) \bigcup S^l(n)$, $n \in \mathcal{I}$;

  $n = n + 1$;

end while

reconstruction of kernel matrix

Compile class-index vector $\mathcal{C}$ for each meta-object created by object associations, $O = \left(x^{1,l}, x^{2,l}, \ldots, x^{M,l}\right)$, in $\mathcal{S}^l(n_{\max})$;

Form composite meta-objects kernel matrix from corresponding reindexed modality kernel matrices: $K_{comp}(O_I, O_J) = \Sigma_m K_m(x_I^m, x_J^m)$;

Form test matrix via: $K_{test}(O_i^{\text{test}}, O_J) = \Sigma_m K_m^{\text{test}}(x_i^{m,\text{test}}, x_J^m)$;

Train classifier using K_{comp} and $\mathcal{C}$;

in order to identify density-degenerate sets for the co-ordinatization process, where n is essentially a density resolution parameter. This process is inherently of order $O(mlog(|K|^2))]$ within each of the n iterations. If we assume the worst-case scenario in which the number of iterations n scales linearly with the number of objects $|K|$ (i.e. such that we assume a consistent density-bin size, irrespective of the number of objects), then the total loop complexity is $O(m|K|log(|K|^2))$. The main loop is thus asymptotically less complex than the initialisation phase.

3.2.3. Final Kernel Matrix Construction

Final kernel matrix construction cannot be derived from the coordinate set identifications derived during the main loop iteration without additional cost, since composite kernel distance must now be explicitly calculated via the kernel addition process (an inherently linear complexity process with respect to kernel matrix entries). For the purpose of calculating the complexity, it may be assumed that this process is carried out explicitly following each identification of degenerate density sets during the main loop iteration, thus extending the original $O(m|K|log(|K|^2))$ process into a $O(m|K|(log(|K|^2) + |K|^2) = O(m|K|(|K|^2) = O(m|K|^3)$ process (using subsampling). The reconstruction phase will thus tend to dominate asymptotically, having a polynomial complexity similar to that of the support vector machine that typically constitutes the next step of the process.

The process as a whole is thus not prohibitive in terms of its order of complexity, though execution time will vary greatly with the degree of morphological correlation between data sets; for data sets exhibiting a typical degree of morphological correlation, such as those of the UCI data set, the SVM process will generally constitute the bottleneck.

4. Experiments

4.1. Illustrative example

As a simple illustrative example of the above process, we consider a 2-class problem in which individual class distributions are defined by a 2-D Gaussian density function with randomly-generated covariance matrix. Differing modalities are simulated by marginalization: two kernel matrices K_1 and K_2 are generated via the dot products of, respectively, the x and y ordinates of a set of 10 pattern vectors generated using the two 2-D class distributions (see Figure 3 left). (Each modality's kernel matrix thus generates a 1D embedding space; the original 2D space represents a hypothetical

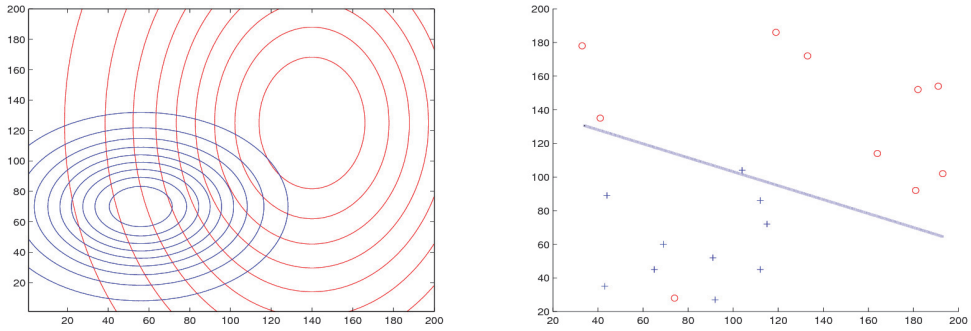

Figure 3: Left: Contour plots of the two class probability density distributions. Right: Decision boundary in the original space for $K_1 + K_2$ (with object correspondence)

composite multi-modal pattern space [or else the space that exists before feature selection]). We also assume no *a priori* correspondence between object ordering in K_1 and K_2 (simulating the 'black-box' multi-modal problem).

We then carry out an SVM classification of the full composite data using a Kernel matrix, K, derived from the inner product of the 20 pattern vectors in the 2D space. We use the LIBSVM [36] toolbox with default settings, *i.e.*, the trade-off parameter C in SVM is set to 1. This provides a benchmark of idealized performance, in which no decision information is lost though marginalization into the two distinct modalities. It also produces a direct illustration of the maximum margin decision boundary in the input domain (see Figure 3 right). Note that the Kernel matrix K equates to $K_1 + K_2$, when the orderings of K_1 and K_2 are assumed to be *correspondent*. The current experiment thus seeks to benchmark the two *non-correspondence*-based methods of augmented kernel fusion and tomographic kernel fusion against ground-truth correspondence-based kernel fusion in order to provide a measure of the degree to which correspondence information is recovered by the two methods.

The composite Kernel matrix, K_{tom}, is then computed from the K_1 and K_2 kernel matrices via the tomographic method. (We also compute the corresponding class-label vector for the meta-objects so formed).

These are then used to produce an SVM classifier. Figure 4 depicts the objects and decision hyper-plane defined in the input domain using this approach. (Of course, the embedding space is not identical to the input space *intrinsically*; in order to generate the decision hyper-plane in Figure 4, it is necessary to identify the specific object correspondences implicit in

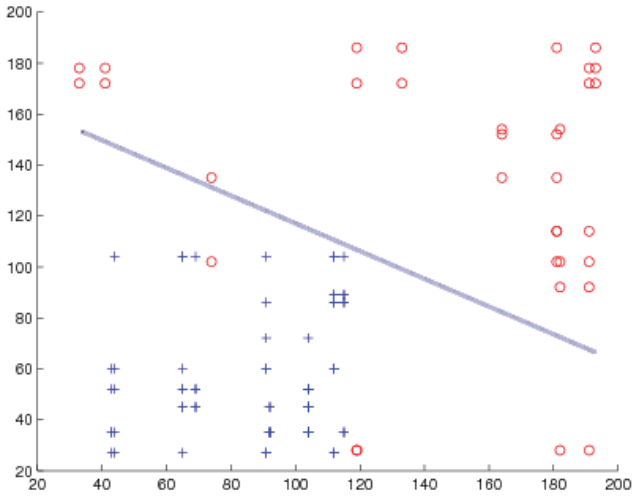

Figure 4: Decision boundaries in the composite space obtained via tomographic kernel fusion

each element of K_{tom} such that the support objects can be attributed the corresponding x and y ordinates). The relevant information is obtainable from the final correspondence set $\mathcal{S}^l(n)$ in conjunction with the original pattern vector ordinates.

Finally, we compute the Augmented Kernel [26] as the block diagonal matrix obtained by appending K_1 and K_2 as a direct sum, *i.e.*, $K_{aug} = K_1 \bigoplus K_2$. We then train an SVM on this kernel matrix (and also plot the decision boundary in the input domain: *cf.* Figure 5; this is calculated by assuming that $K_1 \bigoplus K_2 = ([v_1] \bigoplus [v_2])([v_1] \bigoplus [v_2])^\top$, where v_1 and v_2 are respectively the ordered column vectors of object ordinates x and y that exist within K_1 and K_2 [i.e we recover the 'missing' object correspondences purely in order to visualize the decision boundary; note that the actual decision boundary is calculated purely from the augmented kernel]. This ensures that the support objects exist within the original space with coordinates $(v_1^i, 0)$ or $(0, v_2^i)$ with $v_1^i \in v_1$, $v_2^i \in v_2$).

We thus obtain three performances figures for the three combined kernel matrices K, K_{aug}, K_{tom}. For the distribution given in Figure 3, we obtain, after 20 trials, training accuracy figures of: $accuracy(K) = 90\%$, $accuracy(K_{tom}) = 90\%$, $accuracy(K_{aug}) = 60\%$. (The training error is given here rather than the test error for clarity of methodological illustra-

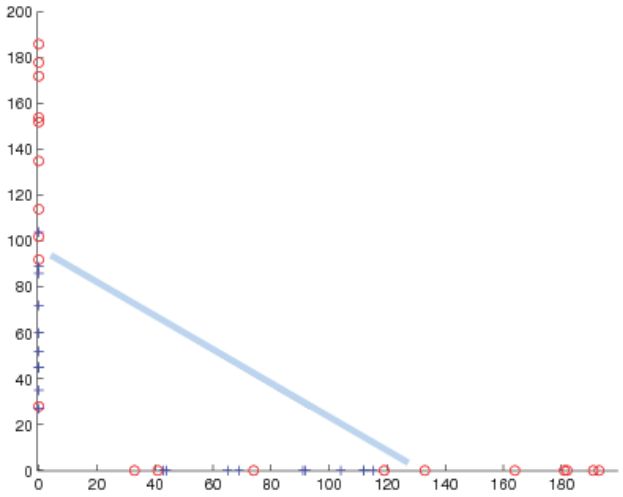

Figure 5: Decision boundaries in the composite space obtained via the Augmented Kernel (note the pattern vectors on the margins).

tion at this stage, i.e with no test set sampling issues). It is clear that despite the creation of some additional 'ambiguity-set' objects, the tomographic kernel procedure produces an extremely similar hyper-plane to the ground-truth. Moreover, the performance obtained with respect to the training set is *identical* to that of the ground truth; the tomographic method has thus reconstructed all of the key classification-relevant missing object correspondence information.

Having thus illustrated the result of kernel combination on the decision boundary using artificial data, we now turn to an evaluation of the method on real data.

4.2. Experiments on UCI data sets

To further evaluate the method on standard data we utilize the UCI data sets *Ionosphere*, *Heart*, *Iris*, *Pima*, *Sonar*, and *Wine*. In order to generate comparable two-class problems across data sets, we identify the class label 'Iris-virginica' with the class label 'Iris-setosa' for the Iris data set, and the class label '3' with class label '1' for the Wine data set.

Implicit within tomographic kernel fusion is a principled methodology for building composite kernels via object correspondences, which can thus be considered as a 'wrapper' to the underlying kernel combination aspect.

There is consequently the same freedom of linear kernel composition that applies in standard MKL (*i.e.*, when object correspondence information *is* available). There hence thus exists freedom to introduce arbitrary coefficients into the linear combination. In the following experiments, however, we wish to focus on the purely tomographic aspects of combination, without consideration of these factors. We thus work with the inner product kernel on feature spaces that have been normalized to a range of ± 1, so that coefficients with unity value are a suitable (*i.e.*, near optimal) choice of linear combination.

We carry out 1400 trials on each of the UCI data sets; for each experimental trial, we divide the feature dimensions into two sets at random. Experimental inputs are the kernel matrices associated with each feature set, along with their associated class labeling (*i.e.*, the kernel matrices are formed via the inner product of the object feature vectors; $K_n = X_n^T X_n$). We again use LIBSVM with default settings as the base classifier and compute the following quantities: the Tomographic Kernel Fusion training and test errors, the Augmented Kernel training and test errors, the training and test errors of the individual feature kernel matrices (as a measure of baseline performance), and finally training and test errors for the kernel formed by summing the individual feature matrices *with all of the ground-truth object associations in place* (*i.e.*, so that the composite kernel matrix is of the same size as individual feature kernel matrices). This latter would thus represent the ideal outcome of tomographic kernel fusion, in which there is no ambiguity of object association, and all objects are correctly associated. It thus serves as an overall performance benchmark for the proposed method.

The calculation of the training error serves to provide a direct window on the relative morphological bias that has been removed from the training set by the proposed procedure, aside from generalization considerations. Note that, in the case of the tomographic fusion strategy this error measure is calculated in relation to the tomographically generated kernel matrices. The test error, of course, remains the principle performance indicator.

Within each of the 1400 trials we perform 3-fold cross-validation to arrive at the accuracy figure for each of the selected feature pairs, *i.e.*, so there are three sub-trials per trial, with a 2:1 training-to-test-set split for each sub-trial. Random sub-sampling is applied for the Cartesian product generation aspect of the tomographic fusion algorithm. The sample frequency for this is specified in terms of the quantity S_{min}; the minimum possible coverage required to generate the same marginal object ambiguity sets (see above). Thus, for two object ambiguity sets, S_1, S_2 associated with feature sets 1 and 2, this value is $S_{min} = sup(|S_1||S_2|)$. We thus select a sampling probability

Table 1: Training Accuracy Rates for the 5 Classification Scenarios (in %). From column 2 to column 6 are SVM training accuracy of: tomographic kernel, kernel from feature set 1 only, kernel from feature set 2 only, augmented kernel, and composite kernel (*i.e.*, with full object correspondence), respectively.

Data set	Tomog.	Ftr. Set 1	Ftr. Set 2	Aug.	Comp.
Ionosphere	96.8 ± 1.1	86.1 ± 8.1	86.9 ± 7.2	85.5 ± 3.9	94.4 ± 1.2
Heart	93.8 ± 1.9	78.3 ± 7.4	80.0 ± 6.0	78.6 ± 2.8	86.4 ± 1.7
Iris	80.1 ± 5.7	68.7 ± 3.9	70.3 ± 4.1	69.5 ± 2.2	73.5 ± 3.1
Pima	81.9 ± 2.0	70.8 ± 5.1	72.6 ± 5.0	71.1 ± 1.1	77.9 ± 1.2
Sonar	97.3 ± 1.3	83.5 ± 8.6	84.0 ± 8.0	83.6 ± 3.6	92.5 ± 1.9
Wine	99.4 ± 0.6	88.7 ± 11.7	91.3 ± 9.5	89.3 ± 5.0	99.1 ± 0.7

Table 2: Testing Accuracy Rates for the 5 Classification Scenarios (in %). From column 2 to column 6 are SVM training accuracy of: tomographic kernel, kernel from feature set 1 only, kernel from feature 2 set only, augmented kernel, and composite kernel (*i.e.*, with full object correspondence), respectively.

Data set	Tomog.	Ftr. Set 1	Ftr. Set 2	Aug.	Comp.
Ionosphere	85.5 ± 3.1	82.4 ± 7.2	83.2 ± 6.5	81.4 ± 4.7	87.7 ± 2.6
Heart	82.6 ± 3.5	76.3 ± 7.9	77.8 ± 6.5	76.7 ± 3.7	83.3 ± 3.3
Iris	65.2 ± 7.2	67.8 ± 6.1	68.6 ± 6.2	68.2 ± 5.8	69.9 ± 6.3
Pima	74.8 ± 2.4	70.1 ± 5.5	71.7 ± 5.5	70.6 ± 2.1	76.8 ± 2.2
Sonar	76.1 ± 4.7	73.0 ± 6.8	73.3 ± 6.3	72.9 ± 4.6	75.9 ± 4.5
Wine	96.4 ± 2.6	86.9 ± 12.3	89.5 ± 10.1	87.5 ± 5.5	97.3 ± 2.1

of factor of $S_{min}/|S_1| \times |S_2|$ to ensure reasonable sampling, and at the same time guarantee non-excessive tomographic kernel matrix sizes.

A characteristic of the UCI data sets is that there are large numbers of similar kernel matrix values in a number of the data sets. We therefore need an additional method to deal with the associated matrix element degeneracy, when this leads to very large Cartesian product sets. For the sortal processes within the tomographic fusion algorithm we thus introduce a sub-sampling of the indices of the sorted density bins. This is performed via the bisection method, which terminates with a maximum density bin index gap of 5. (This is an $O(log(n))$ process and so does not generate significant computational overhead).

Finally, we compute the mean Mahalanobis distance of class centroids over the marginal features as a measure of intrinsic classification difficulty, and also the tomographic performance deficit (defined as the performance of the composite kernel with full object correspondence minus the tomographic

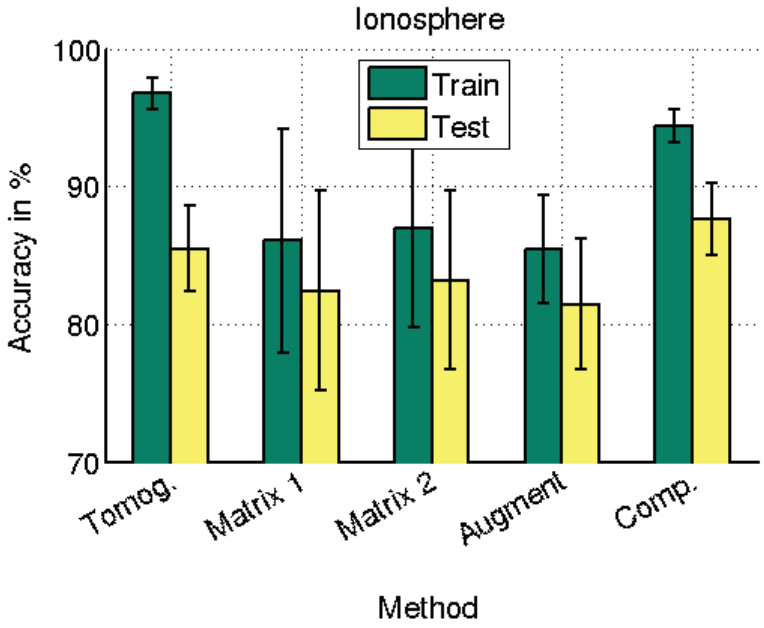

Figure 6: Ionosphere Data Set Accuracy Results

test performance). This relationship is plotted in 12. Results are reported for the six data sets in Tables 1 and 2, and Figures 6 to 11.

4.3. Experiments on CAL500 data sets

We now consider a realistic experimental setting, in which kernel matrices are derived from very different modalities. The CAL500 multi-kernel dataset (http://cosmal.ucsd.edu/cal/) consists of 6 distinct kernel matrices derived from various features that describe retail music [37].

Specifically, it contains 500 distinct songs human-labelled with 174 tags in 8 different semantic categories. These categories include genre, prevailing emotion, instrumentation, presence of solos, and vocal style. Tags are given as Binary labels (*i.e.*, relevant or irrelevant). The 6 kernels matrices are derived as follows:

K_subsamplePPK: this is the Probability Product Kernel (PPK) between Gaussian mixture models (GMMs) of sub-samples of individual songs' Delta-MFCC feature vectors.

K_30sec_PPK: this is a temporal coarse-graining of the above (the PPK between GMMs accumulated over 30 seconds intervals of individual song's Delta-MFCC feature vectors).

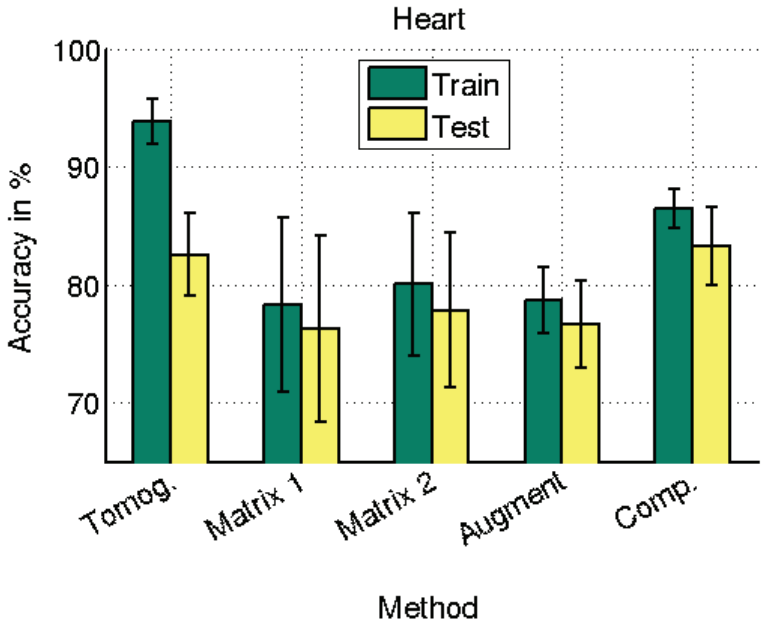

Figure 7: Heart Data Set Accuracy Results

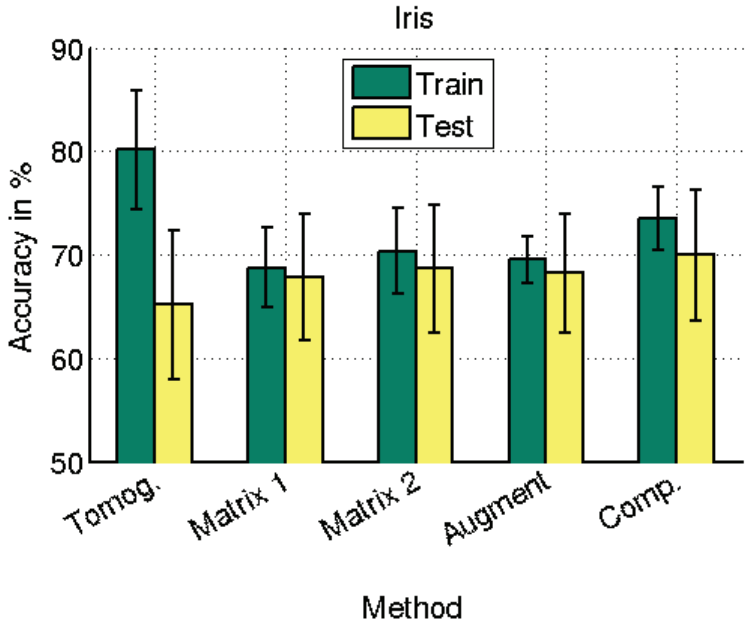

Figure 8: Iris Data Set Accuracy Results

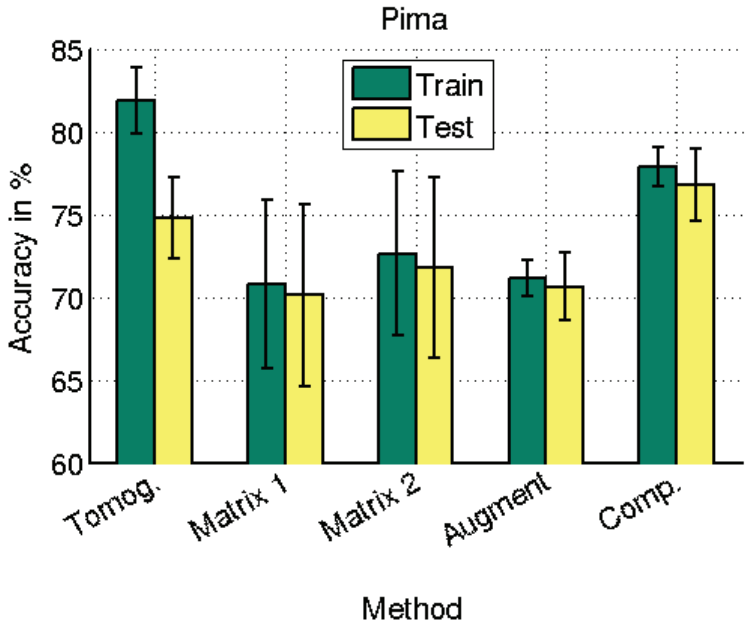

Figure 9: Pima Data Set Accuracy Results

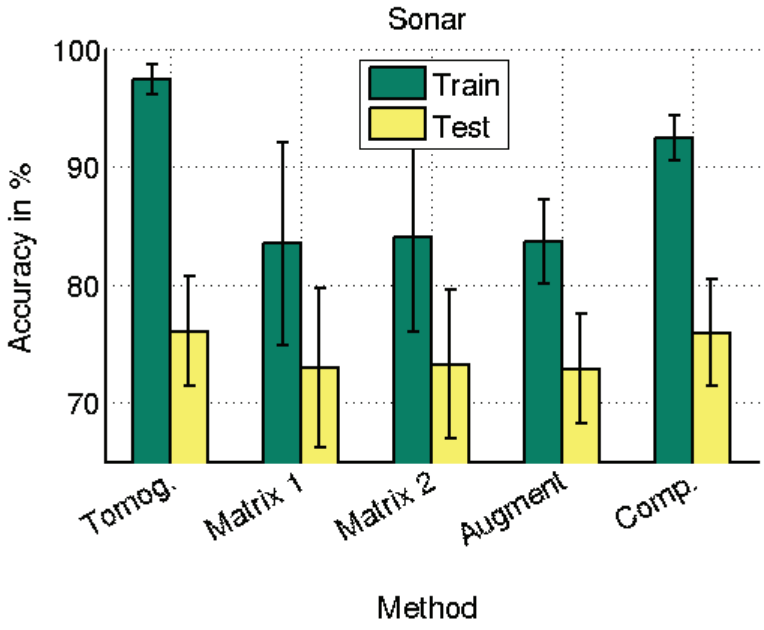

Figure 10: Sonar Data Set Accuracy Results

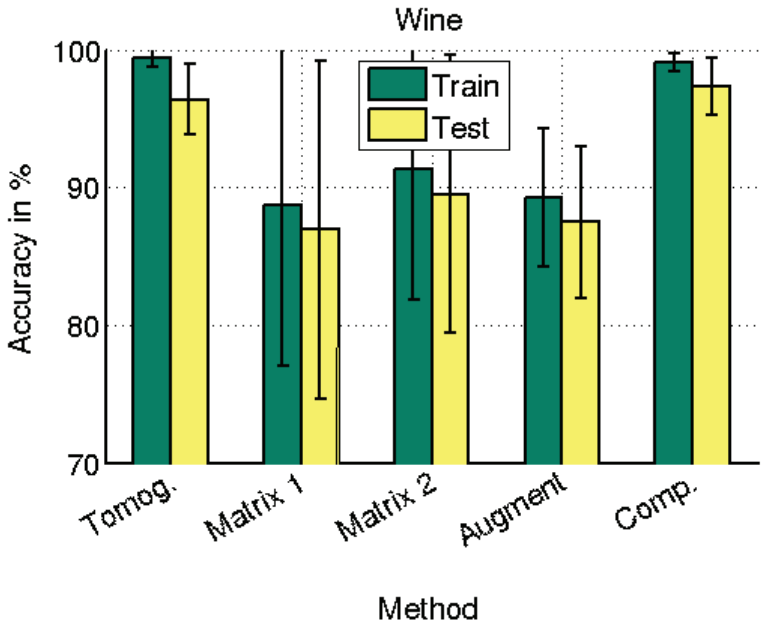

Figure 11: Wine Data Set Accuracy Results

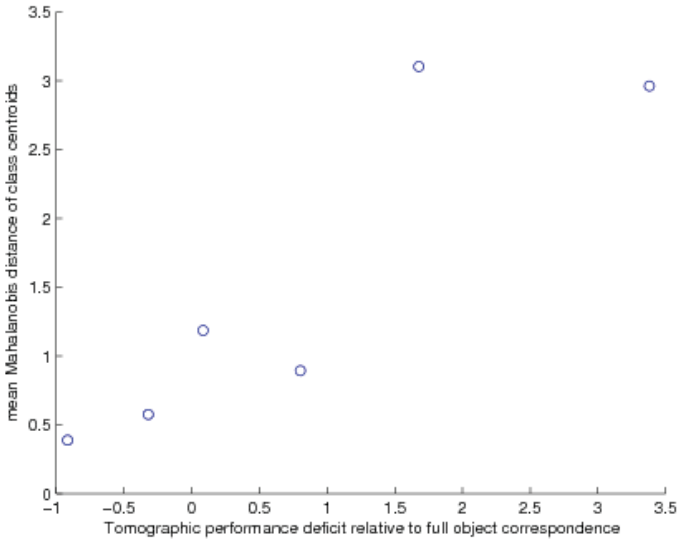

Figure 12: Tomographic performance deficit relative to mean Mahalanobis distance of class centroids

K_30sec_CHROMAPPK: this is the 30 second temporally coarse-grained PPK between GMMs of individual song's Chroma feature vectors (consisting of vectors of pitch-histograms).
K_fpRBF: this is the RBF (Radial basis function) kernel between features representing individual songs.
K_lastfm: this is the RBF kernel of comparisons between document vectors obtained from (Last.fm's) social tagging.
K_webdoc: this is the RBF kernel of comparisons between document vectors describing web pages returned by Google searching on the song title.

For our experiment, we choose the binary tag 'Instrument - Male Lead Vocals' as the class label so as to give the most equally-proportioned binary partition of the possible labels available (most of the other classes exhibit very large imbalances). There are thus 339 positive class labels out of a total 502. We perform 100 trials of two-fold cross validation on the data (that is, for each of the 100 trials we randomly partition the objects into test and training objects, which are then reversed so that objects in the test partition are reallocated to the training partition and vice versa). Following this partitioning of the data we obtain the following accuracy scores for the tomographic and augmented kernels:

Tomographic kernel matrix accuracy %
-test accuracy: 64.70 ± 0.80 %
-train accuracy: $100.0 \pm 0\%$
Augmented kernel matrix accuracy %
-test accuracy: $63.67 \pm 0.86\%$
-train accuracy: $100.0 \pm 0\%$
Average kernel matrix accuracy %
-test accuracy: 63.70 ± 0.84 %
-train accuracy: 100.0 ± 0 %

The latter performance indicator, an average performance of the 6 individual kernel matrices, performs similarly to the augmented kernel, suggesting that the latter is failing to add significant information to the classification problem. The tomographic kernel, however performs substantially better, suggesting that the method is indeed finding some appropriate correspondences between different objects within each modalities' kernel matrix, despite their very different nature. This result is confirmed by applying the same experimental methodology to a very different class label category, in this case the alternative binary partitioning "Emotion-Powerful Strong (yes/no)", which has a total of 160 positive labels, and gives the following results:

Tomographic kernel matrix accuracy %
-test accuracy: 66.55 ± 0.57 %
-train accuracy: $100.0 \pm 0\%$
Augmented kernel matrix accuracy %
-test accuracy: 65.79 ± 0.51 %
-train accuracy: $100.0 \pm 0\%$
Average kernel matrix accuracy %
-test accuracy: 65.79 ± 0.52 %
-train accuracy: 100.0 ± 0 %

5. Discussion

In both the simulated and real experimental environments the tomographic kernel fusion method gives very similar performance to the benchmark composite kernel matrix that incorporates full ground-truth object correspondence information across all of the data sets (with the exception of the *Iris* dataset, the accuracy figure for is derived from a very small number of samples in consequence of the overall paucity of its feature set). The tomographic Kernel method is thus very nearly equivalent to the sum of kernels when object correspondence information is available.

The tomographic kernel fusion method thus finds nearly all of the significant object correspondences for discriminative purposes. This result would, by analogy with the original density-based tomographic fusion approach, presumably scale advantageously with increasing numbers of Kernel input spaces (the favorable scaling of accuracy with input dimensionality is shown for non-kernelized tomographic fusion in [38]).

It is therefore an ideal 'first resort' strategy in situations for which object correspondence information is unavailable, such as those typically found in multi-modal problems (especially biometrics). It is also capable of being deployed in cases where there are significant omissions in the feature values for some fraction of objects (as for instance happens frequently in census data returns). In this case each individual pattern space can be treated as having distinct sets of objects associated with it, only some of which correspond to the objects in other feature spaces.

We have, in the above, used random sub-sampling to avoid the generation of large kernel matrices. A further possibility unexplored here is to employ just the outer vertices of the region defined by the set Cartesian products, *i.e.*, the hypercube corner vertices. This would scale computationally as 2^n rather than $|S|^n$ (S being the number of objects in the ambiguity set for

some iteration of the algorithm). This would give identical results for SVM classification if the classes are linearly separable in the embedding space; however random sub-sampling is likely to be statistically more reliable if classes are not well separated.

In this regard, we might also ask what effect random sampling has in comparison with the non-subsampled alternative (which was omitted from the UCI tests in consequence of the large kernel matrices generated for certain data sets). If the random sub-sampling were significantly affecting performance, it might be expected that an inverse correlation would exist between the tomographic performance deficit (relative to kernel summation with full object correspondence) and the mean Mahalanobis distance of class centroids: randomly-sampled feature pairs often have very poor class separability on the UCI data sets, so that it may be thought that more significance would thereby attach to the sub-sampling data omissions. However, Figure 12 demonstrates the opposite trend, in common with the augmented kernel approach. We therefore conclude that sub-sampling is not the most significant factor affecting performance.

A final observation is that the method has some conceptual similarities with the pyramid kernel matching method [39] used predominantly for image processing, in that it combines unordered feature-sets, and produces an output kernel matrix with an efficiency similar to that of pyramid matching. However, the current method is more principled, in that it proceeds directly from the notion of filtered bias removal in tomographic combination. Ambiguity classes of matches are thus handled in the most optimal fashion. However, the main difference of the current method over pyramid kernel matching is that it takes kernel matrices as inputs, not features, and produces an output kernel matrix using an appropriate linear combination (Mercer properties are guaranteed by the sum of PSD kernels, so that the output defines a definite feature space, *i.e.*, $K(x_i, x_j) =< \Phi(x_i), \Phi(x_j) >, \forall x_i, x_j)$.

Finally, as we noted before, the choice of β coefficients of the linear combination remains free, so an additional optimization along the lines of Lanckriet *et al.* [1] can also be introduced.

6. Conclusions

We have presented a novel methodology for multiple kernel learning that utilizes a tomographic approach to remove the implicit bias from linear combination methods. In kernelizing this approach, the method essentially becomes one of identifying object correspondences between kernel matrices associated with differing modalities. We therefore anticipate that the

method has wide application in areas such as biometrics, for which proprietary data sets often preclude the possibility of object correspondence, and also missing feature problems, which can be treated as a combination over Kernel matrices with incomplete object correspondence.

In the illustrative experiment (the first of our experimental evaluations), the tomographic method gives rise to a very similar decision boundary to the benchmark composite kernel matrix generated in the original composite space. Since this is equivalent to the sum of kernels when object correspondence information *is* available, it is clear that the tomographic kernel fusion method recovers nearly all of this information (at least in regard to the discrimination problem).

In the evaluation on UCI and CAL500 datasets we have shown that the same findings apply on real-world problems, giving superior performance in almost all cases to the augmented kernel alternative, and consequently conclude that the method is an ideal 'first resort' kernel combination strategy.

References

[1] G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, M. Jordan, Learning the kernel matrix with semidefinite programming, JMLR 5 (2004) 27–72.

[2] F. Bach, G. Lanckriet, Multiple kernel learning, conic duality, and the smo algorithm, in: ICML, 2004.

[3] S. Sonnenburg, G. Rätsch, C. Schafer, B. Schölkopf, Large scale multiple kernel learning, JMLR 7 (2006) 1531–1565.

[4] A. Zien, C. Ong, Multiclass multiple kernel learning, in: ICML, 2007.

[5] M. Gönen, E. Alpaydin, Localized multiple kernel learning, in: ICML, 2008, pp. 352–359.

[6] A. Rakotomamonjy, F. Bach, Y. Grandvalet, S. Canu, Simplemkl, JMLR 9 (2008) 2491–2521.

[7] C. Cortes, M. . Mohri, A. Rostamizadeh, Learning nonlinar combinations of kernels, in: NIPS, 2009.

[8] J. Aflalo, A. Ben-Tal, C. Bhattacharyya, J. Nath, S. Raman, Variable sparsity kernel learning, JMLR 12 (2011) 565–592.

[9] M. Kloft, U. Brefeld, S. Sonnenburg, A. Zien, Lp norm multiple kernel learning, JMLR 12 (2011) 953–997.

[10] M. Gönen, E. Alpaydin, Multiple kernel learning algorithms, JMLR 12 (2011) 2211–2268.

[11] M. Kloft, G. Blanchard, On the convergence rate of lp-norm multiple kernel learning, JMLR 13 (2012) 2465–2502.

[12] J. Shawe-Taylor, N. Cristianini, Kernel Methods for Pattern Analysis, Cambridge University Press, 2004.

[13] B. Scholkopf, A. Smola, Learning with Kernels, MIT Press, 2002.

[14] V. Vapnik, The Nature of Statistical Learning Theory, Springer-Verlag, 1999.

[15] S. Mika, Kernel fisher discriminants, PhD Thesis, University of Technology, Berlin, Germany (2002).

[16] G. Baudat, F. Anouar, Generalized discriminant analysis using a kernel approach, Neural Computation 12 (2000) 2385–2404.

[17] O. Chapelle, V. Vapnik, O. Bousquet, S. Mukherjee, Choosing multiple parameters for support vector machines, Machine Learning 46 (2002) 131–159.

[18] O. Bousquet, D. Herrmann, On the complexity of learning the kernel matrix, in: NIPS, 2003.

[19] M. Kloft, U. Brefeld, S. Sonnenburg, A. Zien, Efficient and accurate lp-norm mkl, in: NIPS, 2009.

[20] C. Ong, A. Smola, R. C. Williamson, Learning the kernel with hyperkernels, JMLR 6 (2005) 1043–1071.

[21] Y. Ying, K. Huang, C. Campbell, Information theoretic kernel integration, in: NIPS Workshop on Learning from Multiple Sources, 2009.

[22] S. Kim, A. Magnani, S. Boyd, Optimal kernel selection in kernel fisher discriminant analysis, in: ICML, 2006.

[23] J. Ye, S. Ji, J. Chen, Multi-class discriminant kernel learning via convex programming, JMLR 9 (2008) 719–758.

[24] S. Ji, L. Sun, R. Jin, J. Ye, Multilabel multiple kernel learning, in: NIPS, 2008.

[25] M. Varma, B. Babu, More generality in efficient multiple kernel learning, in: ICML, 2009.

[26] F. Yan, K. Mikolajczyk, J. Kittler, A. Tahir, Combining multiple kernels by augmenting the kernel matrix, in: International Workshop on Multiple Classifier Systems 2010, 2010.

[27] M. Awais, F. Yan, K. Mikolajczyk, J. Kittler, Augmented kernel matrix vs classifier fusion for object recognition, in: British Machine Vision Conference, 2011.

[28] N. Poh, D. Windridge, V. Mottl, A. Tatarchuk, A. Eliseyev, Addressing missing values in kernel-based multimodal biometric fusion using neutral point substitution, IEEE Trans. on Information Forensics and Security.

[29] D. Windridge, J. Kittler, A morphologically optimal strategy for classifier combination: Multiple expert fusion as a tomographic process, IEEE Trans. Pattern Anal. Mach. Intell. 25 (3) (2003) 343–353.

[30] D. Windridge, Tomographic considerations in ensemble bias/variance decomposition, in: Proc. Multiple Classifier Systems, 9th International Workshop, MCS 2010, 2010.

[31] G. Valentini, T. G. Dietterich, Bias-variance analysis of support vector machines for the development of svm-based ensemble methods, J. Mach. Learn. Res. 5 (2004) 725–775.

[32] G. Brown, J. L. Wyatt, P. Tiňo, Managing diversity in regression ensembles, J. Mach. Learn. Res. 6 (2005) 1621–1650.

[33] N. Uedar, R. Nakano, Generalization error of ensemble estimators, In Proceedings of International Conference on Neural Networks (1996) 90–95.

[34] A. Frank, A. Asuncion, UCI machine learning repository (2010).
URL `http://archive.ics.uci.edu/ml`

[35] D. Windridge, J. Kittler, A morphologically optimal strategy for classifier combination, PAMI 25(3) (2003) 343–353.

[36] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, software available at `http://www.csie.ntu.edu.tw/ cjlin/libsvm` (2001).

[37] D. Turnbull, L. Barrington, D. Torres, G. Lanckriet, Towards musical query-by-semantic-description using the cal500 data set, in: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2007, pp. 439–446.

[38] D. Windridge, J. Kittler, Performance measures of the tomographic classifier fusion methodology, Intern. Jrnl. of Pattern Recognition and Artificial Intelligence 19 (6).

[39] K. Grauman, T. Darrell, The pyramid match kernel: Discriminative classification with sets of image features, in: In ICCV, 2005, pp. 1458–1465.