# Non-Sparse Multiple Kernel Learning for Fisher Discriminant Analysis

Fei Yan Josef Kittler Krystian Mikolajczyk Atif Tahir
*Centre for Vision, Speech and Signal Processing*
*University of Surrey*
*Guidford, Surrey, United Kingdom, GU2 7XH*
*Email: {f.yan, j.kittler, k.mikolajczyk, m.tahir}@surrey.ac.uk*

***Abstract*—We consider the problem of learning a linear combination of pre-specified kernel matrices in the Fisher discriminant analysis setting. Existing methods for such a task impose an $\ell_1$ norm regularisation on the kernel weights, which produces sparse solution but may lead to loss of information. In this paper, we propose to use $\ell_2$ norm regularisation instead. The resulting learning problem is formulated as a semi-infinite program and can be solved efficiently. Through experiments on both synthetic data and a very challenging object recognition benchmark, the relative advantages of the proposed method and its $\ell_1$ counterpart are demonstrated, and insights are gained as to how the choice of regularisation norm should be made.**

***Keywords*-Fisher Discriminant Analysis; Multiple Kernel Learning; Semi-Infinite Programming; Object Recognition**

## I. Introduction

It is well recognised that the choice of kernel is critically important for a kernel-based learning method, since the kernel completely determines the embedding of the data in the feature space. In many classification problems, multiple kernels capturing different "views" of the problem are available. In such a situation, one naturally wants to use an "optimal" combination of the kernels.

In [1], Lanckriet et al. proposed a method for learning the optimal linear combination of kernels for the support vector machine (SVM) [2], [3], [4], where the key idea is to optimise the SVM criterion (the margin) with respect not only to the training samples, but also to the kernel weights. In [1], the kernel weights are regularised with an $\ell_1$ norm, which enforces sparsity but may lead to a loss of information. Kloft et al. recently extended this work by replacing the $\ell_1$ norm regularisation with an $\ell_2$ norm version [5]. Another extension to [1] is made by Kim et al. in [6] and Ye et al. in [7], where the SVM criterion is replaced by the one used in the Fisher discriminant analysis (FDA) [8], [9], [10]. This results in a multiple kernel FDA (MK-FDA). Similar to [1], in [6], [7], $\ell_1$ norm regularisation is used.

In this paper we combine these two extensions made to [1]. We propose to use $\ell_2$ norm to regularise MK-FDA. We formulate $\ell_2$ MK-FDA as a semi-infinite program (SIP), which can be solved efficiently. We show that as in the MK-SVM case, $\ell_2$ regularisation tends to produce non-sparse solutions. As a results, less information is lost during the kernel learning process, and the performance is improved over $\ell_1$ MK-FDA as well as the uniform weighting scheme.

The rest of this paper is organised as follows. In Section II, we introduce previous work that is related to this paper. We then present our non-sparse version of MK-FDA, $\ell_2$ MK-FDA, in Section III. Experimental evidence showing the advantage of $\ell_2$ MK-FDA over $\ell_1$ MK-FDA is provided in Section IV. Finally conclusions are given in Section V.

## II. Related Work: Multiple Kernel Learning

In multiple kernel learning (MKL), one is given $n$ $m \times m$ training kernel matrices $K_k, k = 1, \cdots, n$ and $m$ class labels $y_i \in \{1, -1\}, i = 1, \cdots, m$, where $m$ is the number of training sample. In [1], a linear combination of these $n$ kernel matrices is considered: $K = \sum_{k=1}^{n} \beta_k K_k, \quad \beta_i \geq 0, \quad ||\boldsymbol{\beta}||_1 = 1$. Geometrically, taking the sum of two kernel matrices can be interpreted as taking the Cartesian product of the two associated feature spaces. Scaling the feature spaces prior to taking the product leads to different embeddings of the data in the augmented feature spaces. The goal of MKL is then to learn the "optimal" scaling of the feature spaces, such that the "separability" of the two classes in the augmented feature space is maximised.

[1] proposes to use the margin as a measure of separability, i.e., to learn $\boldsymbol{\beta}$ by maximising the margin between the two classes. This maximisation problem has been formulated as different mathematical programs [1], [11], [12], [13], [14]. The original semi-definite programming (SDP) formulation [1] becomes intractable when $m$ is in the order of thousands, while the semi-infinite linear programming (SILP) formulation [12] and the reduced gradient descent algorithm [14] can deal with much larger problems.

The learning problem in [1] imposes an $\ell_1$ regularisation on the kernel weights. It has been known that $\ell_1$ norm regularisation tends to produce sparse solutions (e.g. [15]), which means during the learning most kernels are assigned zero weights. This behaviour may not always be desirable, since the information carried in the zero-weighted kernels is lost. Recently, a non-sparse version of MK-SVM was proposed in [5], where an $\ell_2$ norm regularisation is imposed instead of $\ell_1$ norm. Experiments in [5] show that the $\ell_2$ regularised multiple kernel SVM (MK-SVM) may be advantageous over its $\ell_1$ counterpart.

Another extension to [1] is made in [6], [7], where the (kernel) FDA is considered instead of the SVM. The basic idea is to use the FDA criterion as the measure of separability, i.e., to maximise the ratio of the projected between class scatter and projected within class scatter with respect to kernel weights. Since [6], [7] use an $\ell_1$ regularisation on $\boldsymbol{\beta}$, they also have the "over-selective" problem of $\ell_1$ MK-SVM. Note that in the rest of this paper we do not distinguish between conventional FDA and kernel FDA, and refer to both of them as FDA.

## III. Non-Sparse Multiple Kernel FDA

In this section we first formulate our non-sparse MK-FDA based on the $\ell_2$ regularisation of kernel weights. We then solve the associated optimisation problem using SIP.

### A. Problem Formulation

We consider a binary classification problem. Our goal is to learn optimal kernel weights $\boldsymbol{\beta} \in \mathbb{R}^n$ for the linear combination of $n$ kernels under the $\ell_2$ constraint:

$$K = \sum_{k=1}^{n} \beta_k K_k, \quad \beta_i \geq 0, \quad ||\boldsymbol{\beta}||_2 = 1 \tag{1}$$

such that the ratio criterion of FDA is maximised.

We assume each kernel is centred in its feature space. The centring in the feature space can be performed implicitly [16] by $\tilde{K}_k = PK_kP$, where $P$ is the $m \times m$ centring matrix defined as $P = I - \frac{1}{m}\mathbf{1} \cdot \mathbf{1}^T$, where $I$ is the $m \times m$ identity matrix.

Let $m^+$ be the number of positive training samples, and $m^- = m - m^+$ be the number of negative training samples. For a given kernel $\tilde{K}$, let $\phi(x_i^+)$ be the $i^{\text{th}}$ positive training point in the implicit feature space associated with $\tilde{K}$, $\phi(x_i^-)$ be the $i^{\text{th}}$ negative training point in the feature space. Here $x_i^+$ and $x_i^-$ can be thought of as training samples in some input space, and $\phi$ is the mapping to the feature space. Also let $\mu^+ = \frac{1}{m^+}\sum_{i=1}^{m^+}\phi(x_i^+)$ be the centroid of the positive samples in the feature space and $\mu^- = \frac{1}{m^-}\sum_{i=1}^{m^-}\phi(x_i^-)$ be the centroid of the negative samples. The within class covariance matrices of the two classes are:

$$C^+ = \frac{1}{m^+}\sum_{i=1}^{m^+}(\phi(x_i^+) - \mu^+)(\phi(x_i^+) - \mu^+)^T \tag{2}$$

$$C^- = \frac{1}{m^-}\sum_{i=1}^{m^-}(\phi(x_i^-) - \mu^-)(\phi(x_i^-) - \mu^-)^T \tag{3}$$

The between class scatter $S_B$ and within class scatter $S_w$ are defined as:

$$S_B = \frac{m^+m^-}{m}(\mu^+ - \mu^-)(\mu^+ - \mu^-)^T \tag{4}$$

$$S_W = m^+C^+ + m^-C^- \tag{5}$$

The objective of single kernel FDA is then to find the projection direction $\mathbf{w}$ in the feature space, such that the ratio of the projected between class and within class scatter is maximised. In other words, we want to maximise $\frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}}$. It is easy to show that this is equivalent to maximising $\frac{\mathbf{w}^T \frac{m}{m^+m^-} S_B \mathbf{w}}{\mathbf{w}^T S_T \mathbf{w}}$, where $S_T = S_B + S_W$ is the total scatter matrix. In practice a regularised version,

$$J_1(\mathbf{w}) = \frac{\mathbf{w}^T \frac{m}{m^+m^-} S_B \mathbf{w}}{\mathbf{w}^T (S_T + \lambda I)\mathbf{w}} \tag{6}$$

is maximised to improve generalisation and numerical stability [17], where $\lambda$ is a small positive number.

From Theorem 2.1 of [7], for a given kernel $\tilde{K}$, the maximal value of (6) is:

$$J_1^* = \mathbf{a}^T\mathbf{a} - \mathbf{a}^T(I + \frac{1}{\lambda}\tilde{K})^{-1}\mathbf{a} \tag{7}$$

where $\mathbf{a} = (\frac{1}{m^+}, \cdots, \frac{1}{m^+}, \frac{-1}{m^-}, \cdots, \frac{-1}{m^-})^T \in \mathbb{R}^m$ contains the centred labels. On the other hand, Lemma 2.1 of [7] states that the $\mathbf{w}$ that maximises (6) also minimises:

$$J_2(\mathbf{w}) = ||(\phi(X)P)^T\mathbf{w} - \mathbf{a}||^2 + \lambda||\mathbf{w}||^2 \tag{8}$$

where $\phi(X) = (\phi(x_1^+), \cdots, \phi(x_{m^+}^+), \phi(x_1^-), \cdots, \phi(x_{m^-}^-))$ and the minimum of (8) is given by:

$$J_2^* = \mathbf{a}^T(I + \frac{1}{\lambda}\tilde{K})^{-1}\mathbf{a} \tag{9}$$

Due to strong duality, the minimal value of (8) is equal to the maximal value of its Lagrangian dual problem (Theorem 2.2 of [7]), i.e.: $J_2^* = \max_{\boldsymbol{\alpha}} \boldsymbol{\alpha}^T\mathbf{a} - \frac{1}{4}\boldsymbol{\alpha}^T(I + \frac{1}{\lambda}\tilde{K})$, or equivalently

$$J_2^* = -(\min_{\boldsymbol{\alpha}} \frac{1}{4}\boldsymbol{\alpha}^T(I + \frac{1}{\lambda}\tilde{K})\boldsymbol{\alpha} - \boldsymbol{\alpha}^T\mathbf{a}) \tag{10}$$

where $\boldsymbol{\alpha} \in \mathbb{R}^m$. By combining (7), (9) and (10), it directly follows that the maximal value of the original FDA objective (6) is given by:

$$J_1^* = \mathbf{a}^T\mathbf{a} + (\min_{\boldsymbol{\alpha}} \frac{1}{4}\boldsymbol{\alpha}^T(I + \frac{1}{\lambda}\tilde{K})\boldsymbol{\alpha} - \boldsymbol{\alpha}^T\mathbf{a}) \tag{11}$$

Now instead of using a fixed single kernel, consider the case where the kernel $\tilde{K}$ can be chosen from a set of centred kernels $\tilde{\mathcal{K}}$. It easily follows that the optimal $\tilde{K}$ that maximises (11) is found by solving:

$$\max_{\tilde{K} \in \tilde{\mathcal{K}}} \min_{\boldsymbol{\alpha}} \quad \frac{1}{4}\boldsymbol{\alpha}^T(I + \frac{1}{\lambda}\tilde{K})\boldsymbol{\alpha} - \boldsymbol{\alpha}^T\mathbf{a} \tag{12}$$

We consider $\tilde{\mathcal{K}}$ as linear combinations of $n$ pre-specified kernels $\tilde{K}_1, \cdots, \tilde{K}_n$. The kernel weights must be regularised somehow to make sure (9) remains meaningful and does not become arbitrarily small. We propose to impose an $\ell_2$ regularisation on the kernel weights:

$$\tilde{\mathcal{K}} = \{\tilde{K} = \sum_{k=1}^{n} \beta_k \tilde{K}_k : \boldsymbol{\beta} \geq \mathbf{0}, ||\boldsymbol{\beta}||_2 = 1\} \tag{13}$$

Substituting (13) into (12) we arrive at the $\ell_2$ MK-FDA problem:

$$\begin{aligned} \max_{\boldsymbol{\beta}} \min_{\boldsymbol{\alpha}} \quad & \tfrac{1}{4\lambda}\boldsymbol{\alpha}^T \textstyle\sum_{k=1}^{n} \beta_k \tilde{K}_k \boldsymbol{\alpha} + \tfrac{1}{4}\boldsymbol{\alpha}^T\boldsymbol{\alpha} - \boldsymbol{\alpha}^T\mathbf{a} \\ \text{s.t.} \quad & \boldsymbol{\beta} \geq \mathbf{0}, \quad ||\boldsymbol{\beta}||_2 = 1 \end{aligned} \tag{14}$$

### B. Solving the Optimisation Problem with Semi-Infinite Programming

A semi-infinite program is an optimisation problem with finite number of variables $\mathbf{x} \in \mathbb{R}^{\mathbf{d}}$ on a feasible set described by infinitely many constraints [18], [19]:

$$\min_{\mathbf{x}} f(\mathbf{x}) \qquad \text{s.t. } g(\mathbf{x}, u) \geq 0 \ \ \forall u \in \mathcal{U} \tag{15}$$

where $\mathcal{U}$ is an infinite index set. We show in Theorem 1 that the $\ell_2$ MK-FDA problem (14) can be formulated as an SIP.

**Theorem 1.** *Given a set of $n$ centred kernel matrices $\tilde{K}_1, \cdots, \tilde{K}_n$, the kernel weights $\boldsymbol{\beta}$ that optimise* (14) *are given by solving the following SIP problem:*

$$\max_{\theta, \boldsymbol{\beta}} \quad \theta \tag{16}$$
$$\text{s.t.} \quad \boldsymbol{\beta} \geq \mathbf{0}, \quad ||\boldsymbol{\beta}||_2 = 1, \quad S(\boldsymbol{\alpha}, \boldsymbol{\beta}) \geq \theta \quad \forall \boldsymbol{\alpha} \in \mathbb{R}^m$$

*where*

$$S(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{4\lambda} \boldsymbol{\alpha}^T \sum_{k=1}^{n} \beta_k \tilde{K}_k \boldsymbol{\alpha} + \frac{1}{4} \boldsymbol{\alpha}^T \boldsymbol{\alpha} - \boldsymbol{\alpha}^T \mathbf{a} \tag{17}$$

*(Proof omitted due to lack of space.)*

As in the $\ell_2$ MK-SVM case, the $\ell_2$ MK-FDA problem (16) is non-convex since the set defined by $\{\boldsymbol{\beta} : \boldsymbol{\beta} \geq \mathbf{0}, ||\boldsymbol{\beta}||_2 = 1\}$ is non-convex. To remedy this, we relax the constraint $\boldsymbol{\beta} \geq \mathbf{0}, ||\boldsymbol{\beta}||_2 = 1$ to $\boldsymbol{\beta} \geq \mathbf{0}, ||\boldsymbol{\beta}||_2 \leq 1$:

$$\max_{\theta, \boldsymbol{\beta}} \quad \theta \tag{18}$$
$$\text{s.t.} \quad \boldsymbol{\beta} \geq \mathbf{0}, \quad ||\boldsymbol{\beta}||_2 \leq 1, \quad S(\boldsymbol{\alpha}, \boldsymbol{\beta}) \geq \theta \quad \forall \boldsymbol{\alpha} \in \mathbb{R}^m$$

According to Theorem 2, the approximation error of this relaxation is zero when $\tilde{K}_1, \cdots, \tilde{K}_n$ are positive definite.

**Theorem 2.** *Let $(\theta^*, \boldsymbol{\beta}^*)$ be optimal points of optimisation problem* (18) *and $\tilde{K}_1, \cdots, \tilde{K}_n$ be positive definite. Then we always have $||\boldsymbol{\beta}^*||_2 = 1$. (Proof omitted due to lack of space.)*

We use the wrapper algorithm proposed in [12] to solve (18). This algorithm is based on a technique called column generation, where the basic idea is to divide an SIP into an inner sub-problem and an outer sub-problem. The algorithm alternates between solving the two sub-problems until convergence. At step $t$, the inner sub-problem identifies constraints that maximises the constraint violation for an intermediate solution $(\theta^{(t)}, \boldsymbol{\beta}^{(t)})$:

$$\boldsymbol{\alpha}^{(t)} := \arg\min_{\boldsymbol{\alpha}} S(\boldsymbol{\alpha}, \boldsymbol{\beta}^{(t)}) \tag{19}$$

Observing that (19) is an unconstrained quadratic program, $\boldsymbol{\alpha}^{(t)}$ is obtained by solving the following linear system [7]:

$$(\frac{1}{2} I + \frac{1}{2\lambda} \sum_{k=1}^{n} \beta_k^{(t)} \tilde{K}_k) \boldsymbol{\alpha}^{(t)} = \mathbf{a} \tag{20}$$

If $\boldsymbol{\alpha}^{(t)}$ satisfies constraint $S(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)}) \geq \theta^{(t)}$ then solution $(\theta^{(t)}, \boldsymbol{\beta}^{(t)})$ is optimal. Otherwise, the constraint is added to the set of constraints and the algorithm switches to the outer sub-problem.

Table I
AN ITERATIVE ALGORITHM FOR SOLVING THE SIP PROBLEM (18)

- **Initialisation**: $S^{(0)} = 1$, $\theta^{(1)} = -\infty$, $\beta_k^{(1)} = n^{-1/2}$ for $k = 1, \cdots, n$
- **for** $t = 1, 2, \cdots$ **do**
  - Compute $\boldsymbol{\alpha}^{(t)} = \arg\min_{\boldsymbol{\alpha}} S(\boldsymbol{\alpha}, \boldsymbol{\beta}^{(t)})$ using (20)
  - Compute $S^{(t)} := S(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)})$
  - **if** $|1 - \frac{S^{(t)}}{\theta^{(t)}}| \leq \epsilon$ **break**
  - Compute $(\theta^{(t+1)}, \boldsymbol{\beta}^{(t+1)}) = \arg\max_{\theta, \boldsymbol{\beta}} \theta$, with respect to $\theta \in \mathbb{R}$ and $\boldsymbol{\beta} \in \mathbb{R}^n$, subject to $\boldsymbol{\beta} \geq \mathbf{0}$, $||\boldsymbol{\beta}||_2 \leq 1$ and $S(\boldsymbol{\alpha}^{(r)}, \boldsymbol{\beta}) \geq \theta$ for $r = 1, \cdots, t$.
- **end for**

The outer sub-problem is also called the restricted master problem. At step $t$, it computes the optimal $(\theta^{(t)}, \boldsymbol{\beta}^{(t)})$ in (18) for a restricted subset of constraints:

$$\max_{\theta, \boldsymbol{\beta}} \theta \tag{21}$$
$$\text{s.t.} \quad \boldsymbol{\beta} \geq \mathbf{0}, \ ||\boldsymbol{\beta}||_2 \leq 1, \ S(\boldsymbol{\alpha}^{(r)}, \boldsymbol{\beta}) \geq \theta \ \forall r = 1, \cdots, t$$

This turns out to be a quadratically constrained linear program (QCLP) with one quadratic constraint (the norm constraint) and $t + 1$ linear constraints, and can be solved by off-the-shelf optimisation tools such as Mosek (http://www.mosek.com).

Normalised maximal constraint violation is used as a convergence criterion. The algorithm stops when $|1 - \frac{S^{(t)}}{\theta^{(t)}}| \leq \epsilon$, where $S^{(t)} := S(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)})$ and $\epsilon$ is a pre-defined accuracy parameter. This iterative algorithm for solving the $\ell_2$ MK-FDA SIP problem is summarised in Table I. It is a special case of a set of SIP algorithms known as exchange methods, which are guaranteed to converge [18].

## IV. Experiments

### A. Simulation

We simulate two classes by sampling 100 points from two 2-dimensional Gaussian distributions, 50 points from each. The means of the two distributions in both dimensions are drawn from a uniform distribution between 1 and 2, and the covariances of the two distributions are also randomly generated. A radial basis function (RBF) kernel, which has proven to be positive definite [20], is then constructed using these 2-dimensional points. Similarly, 100 testing points are sampled from the same distributions, 50 from each, and an RBF kernel is built for the testing points. FDA is then applied to find the best projection direction in the feature space and compute the error rate on the testing set. Fig. 1 gives three examples of the simulated points. It shows that due to the parameters used in the two Gaussian distributions, the two classes are heavily, but not completely, overlapping. As a result, the error rate of such a problem is around 0.43: slightly better than a random guess.

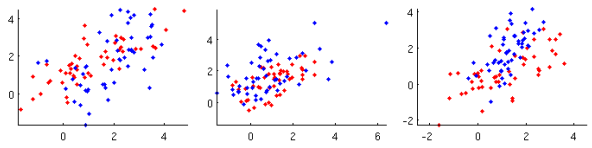

Figure 1. Three examples of the two Gaussian distributions.

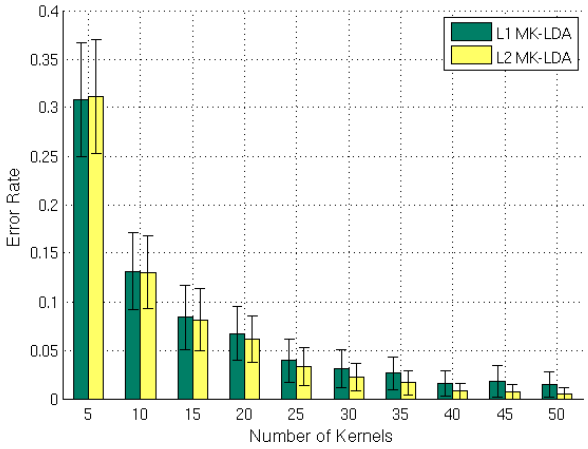


Figure 2. Mean and standard deviation of error rate of $\ell_1$ MK-FDA and $\ell_2$ MK-FDA using various number of kernels.

The process above is repeated $n$ times, resulting in $n$ training kernels (and $n$ corresponding testing kernels). These $n$ training kernels, although generated independently, can be thought of as kernels that capture different "views" of a single binary classification problem. With this interpretation in mind, we apply $\ell_1$ and $\ell_2$ MK-FDAs to learn optimal kernel weights for this classification problem. We vary the number $n$ from 5 to 50 at a step of 5. For each value of $n$, $\ell_1$ and $\ell_2$ MK-FDAs are applied and the resulting error rates are recorded. This process is repeated 100 time for each value of $n$ to compute the mean and standard deviation of error rate. The results for various $n$ values are plotted in Fig. 2.

It is clear in Fig. 2 that as the number of kernels increases, the error rates of both methods drop. This is expected, since more kernels bring more discriminative information. Another observation is that $\ell_1$ MK-FDA slightly outperforms $\ell_2$ MK-FDA when the number of kernels is 5, and vice versa when the number of kernels is 10 or 15. When there are 20 kernels, the advantage of the proposed $\ell_2$ MK-FDA becomes obvious. As the number of kernels keeps increasing, its advantage becomes more and more evident.

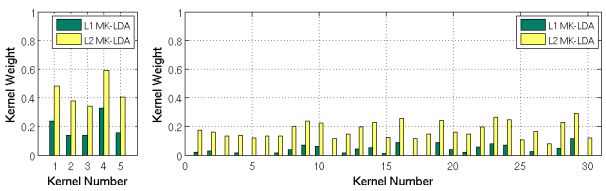


Figure 3. Comparing the kernel weights learnt from $\ell_1$ and $\ell_2$ MK-FDA. Left: using 5 kernels. Right: using 30 kernels.

The different behaviours of $\ell_1$ and $\ell_2$ MK-FDA can be explained by the different weights learnt from them. When the number of kernels is sufficiently small, the over-selectiveness of $\ell_1$ regularisation does not occur: as can be seen in the left plot of Fig. 3, when there are only 5 kernels, all of them get non-zero weights in both $\ell_1$ and $\ell_2$ MK-FDAs. As the number of kernels increases, eventually there are enough of them for the over-selectiveness of $\ell_1$ regularisation to exhibit itself. As the the right plot of Fig. 3 shows, when 30 kernels are used, some of the them are assigned zero weights by $\ell_1$ MK-FDA. This leads to loss of information. By contrast, the weights learnt in the proposed $\ell_2$ MK-FDA are non-sparse, hence the better performance.

### B. Object Recognition Dataset

*1) Experimental Setup:* In this section, we compare $\ell_1$ and $\ell_2$ MK-FDAs on the PASCAL visual object classes (VOC) challenge 2008 [21] development dataset. The VOC challenge provides a yearly benchmark for comparison of object classification methods, with one of the most challenging datasets in the object recognition / image classification community.

The classification of 20 object classes is treated as 20 independent binary problems. In our experiments, average precision [22] is used to measure the performance of each binary classifier. It is particularly suitable for measuring the performance of a retrieval system, since it emphasises higher ranked relevant instances. The mean of the APs of all classes in the dataset, MAP, is used as a measure of the overall performance.

We compare three learning algorithms that use multiple kernels: FDA with uniformly weighted kernel, $\ell_1$ MK-FDA with SILP formulation [7], and $\ell_2$ MK-FDA with SIP formulation proposed in this paper. The implementation of $\ell_1$ MK-FDA is available from the authors' website. For the proposed $\ell_2$ MK-FDA, the linear system in the inner sub-problem is solved using Matlab (http://www.mathworks.com), and the QCLP in the outer sub-problem is solved using the Mosek optimisation software (http://www.mosek.com).

SIFT descriptor [23] and codebook technique [24] are used to generate kernels. The combination of two sampling techniques (dense sampling and Harris-Laplace interest point sampling), five colour variants of SIFT descriptors [25], and three ways of dividing an image into spatial location grids results in $2 \times 5 \times 3 = 30$ "informative" kernels. We also generate 30 sets of 10-dimensional random vectors, and build 30 RBF kernels from them. These random kernels are then mixed with the informative ones, to study how the properties of kernels affect the performance of MK-FDAs. All the 60 kernels are positive definite, and are normalised to have unit trace.

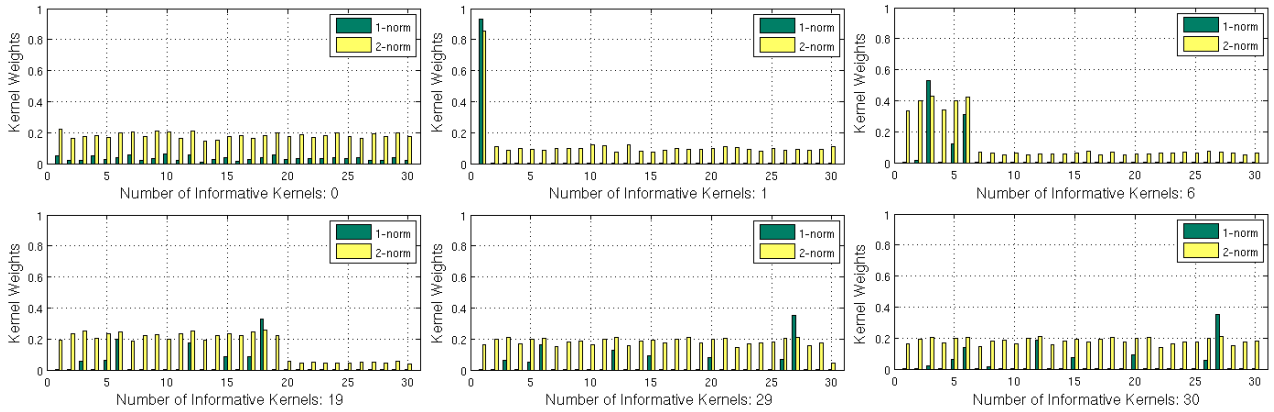


Figure 4. Learnt kernel weights in $\ell_1$ MK-FDA and $\ell_2$ MK-FDA. "motorbike" class.

The number of kernels used in each run is fixed to 30, but with varying composition. In the first run, only the 30 random kernels are used. In the following runs the number of informative kernels is increased and that of random kernels decreased, until the 31st run, where all 30 kernels are informative. In each run, we apply the three algorithms to the 20 binary problems, compute the MAP for each algorithm, and record the kernel weights learnt from the MK-FDAs.

*2) Experimental Results:* Fig. 4 plots the kernel weights learnt from $\ell_1$ MK-FDA and $\ell_2$ MK-FDA. In each sub-figure, the weights of the informative kernels are plotted towards the left end and those of random ones towards the right. We clearly observe again the "over-selective" behaviour of $\ell_1$ norm: it sets the weights of most kernels, including informative kernels, to zero. By contrast, the proposed $\ell_2$ MK-FDA always assigns non-zero weights to the informative kernels. However, $\ell_2$ MK-FDA is "under-selective": it assigns non-zero weights to the random kernels, which introduces noise to the augmented feature space. It is also worth noting that the kernels that do get selected by $\ell_1$ MK-FDA are usually the ones that get highest weights in $\ell_2$ MK-FDA.

Given the observation of the learnt weights, it is not surprising to see in Fig. 5 that $\ell_1$ MK-FDA outperforms $\ell_2$ MK-FDA when the noise level is high and vice versa when the noise level is low. Another observation is that $\ell_2$ MK-FDA consistently outperforms uniform weighting scheme regardless of noise level. However, when all kernels are informative, the improvement of $\ell_2$ MK-FDA over uniform FDA is small (0.463 over 0.462). Our explanation to this is that, first, due to the way they are constructed, the 30 informative kernels actually carry similar information. As a result, there is not much variance to be learnt in the kernel weights. Second, a difference of 0.001 in MAP is more significant than it may appear to be. For example, the leading methods in PASCAL VOC classification competitions typically differ only by a few tenths of a percent in MAP. Moreover, uniform FDA was used by the method

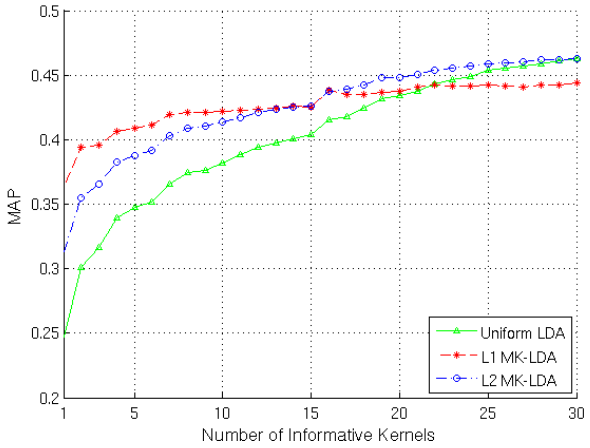


Figure 5. MAP of three weighting schemes for multiple kernel FDA.

that produced the highest MAP in PASCAL VOC 2008 classification challenge [21]. This means our $\ell_2$ MK-FDA improves over one of the state-of-the-art classifiers for object recognition. In the experiments, the parameter $\lambda$ is set to $10^{-4}$. This value is chosen using 3-fold cross-validation according to MAP, from 9 values that are logarithmically evenly spaced over $10^{-1}$ to $10^{-9}$. In fact, the performance of MK-FDAs is not sensitive to $\lambda$ when $\lambda$ is between $10^{-3}$ and $10^{-6}$.

*3) Time Complexity Analysis:* Both $\ell_1$ and $\ell_2$ MK-FDAs are based on the column generation technique. In the inner loop, the linear system (20) solved in both methods has a complexity of $O(m^3)$ [7]. In the outer loop, the Matlab linear program (LP) solver employed in $\ell_1$ MK-FDA is slightly faster than the Mosek QCLP solver in $\ell_2$ MK-FDA. However, we observe that it usually takes a few tens of iterations for the SILP in $\ell_1$ MK-FDA to converge, while less than 5 for the SIP in $\ell_2$ MK-FDA. This difference in the number of iterations overwhelms the advantage of LP over QCLP in each step of the outer loop, and results in big difference in total running time. On average, it takes 259.1

seconds to train $\ell_1$ MK-FDA for one object class, while only 15.6 seconds for the $\ell_2$ version. The stopping threshold $\epsilon$ is set to $5 \times 10^{-4}$ for both methods.

## V. Conclusions

In this paper we have presented a multiple kernel learning algorithm for Fisher discriminant analysis. We adopt the framework of learning a linear combination of a pre-specified set of kernels, and learn the kernel weights by maximising the FDA criterion. Instead of the $\ell_1$ norm used in previous work, we propose to regularise the kernel weights using the $\ell_2$ norm. This results in non-sparse solution, which avoids the over-selective problem of $\ell_1$ regularisation. Experiments on both synthetic data and a difficult object recognition benchmark show that $\ell_1$ MK-FDA is more resistant to noise, while in situations where kernels carry complementary information about the classification problem, the proposed $\ell_2$ MK-FDA offers better performance than the $\ell_1$ version. The proposed method also consistently outperforms the naive uniform weighting scheme.

## Acknowledgement

This work has been supported by EU IST-2-045547 VIDI-Video Project. We would also like to thank Marius Kloft for providing the proof of Theorem1 in [5], and thank Jianhui Chen and Prof. Jieping Ye for helpful discussions.

## References

[1] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semidefinite programming," *Journal of Machine Learning Research*, vol. 5, pp. 27–72, 2004.

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

[3] B. Scholkopf and A. Smola, *Learning with Kernels*. MIT Press, 2002.

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

[5] M. Kloft, U. Brefeld, P. Laskov, and S. Sonnenburg, "Non-sparse multiple kernel learning," in *NIPS Workshop on Kernel Learning: Automatic Selection of Optimal Kernels*, 2008.

[6] S. Kim, A. Magnani, and S. Boyd, "Optimal kernel selection in kernel fisher discriminant analysis," in *International Conference on Machine Learning*, 2006.

[7] J. Ye, S. Ji, and J. Chen, "Multi-class discriminant kernel learning via convex programming," *Journal of Machine Learning Research*, vol. 9, pp. 719–758, 2008.

[8] R. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179–188, 1936.

[9] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. Muller, "Fisher discriminant analysis with kernels," in *IEEE Signal Processing Society Workshop: Neural Networks for Signal Processing*, 1999.

[10] G. Baudat and F. Anouar, "Generalized discriminant analysis using a kernel approach," *Neural Computation*, vol. 12, pp. 2385–2404, 2000.

[11] F. R. Bach and G. R. G. Lanckriet, "Multiple kernel learning, conic duality, and the smo algorithm," in *International Conference on Machine Learning*, 2004.

[12] S. Sonnenburg, G. Ratsch, C. Schafer, and B. Scholkopf, "Large scale multiple kernel learning," *Journal of Machine Learning Research*, vol. 7, pp. 1531–1565, 2006.

[13] A. Zien and C. S. Ong, "Multiclass multiple kernel learning," in *International Conference on Machine Learning*, 2007, pp. 1191–1198.

[14] A. Rakotomamonjy, F. Bach, Y. Grandvalet, and S. Canu, "Simplemkl," *Journal of Machine Learning Research*, vol. 9, pp. 2491–2521, 2008.

[15] G. Ratsch, "Robust boosting via convex optimization," PhD Thesis, University of Potsdam, Potsdam, Germany, 2001.

[16] B. Scholkopf, A. Smola, and K. Muller, "Kernel principal component analysis," *Advances in Kernel Methods: Support Vector Learning*, pp. 327–352, 1999.

[17] S. Mika, "Kernel fisher discriminants," PhD Thesis, University of Techinology, Berlin, Germany, 2002.

[18] R. Hettich and K. Kortanek, "Semi-infinite programming: Theory, methods, and applications," *SIAM Review*, vol. 35(3), pp. 380–429, 1993.

[19] M. Lopez and G. Still, "Semi-infinite programming," *European Journal of Operational Research*, vol. 180, pp. 491–518, 2007.

[20] C. Micchelli, "Interpolation of scattered data: Distance matrices and conditionally positive definite functions," *Constructive Approximation*, vol. 2, pp. 11–22, 1986.

[21] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results," http://www.pascal-network.org/challenges/VOC/voc2008/workshop/index.html.

[22] C. Snoek, M. Worring, J. Gemert, J. Geusebroek, and A. Smeulders, "The challenge problem for automated detection of 101 semantic concepts in multimedia," in *ACM Multimedia Conference*, 2006, pp. 421–430.

[23] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27(10), pp. 1615–1630, 2005.

[24] J. Gemert, J. Geusebroek, C. Veenman, and A. Smeulders, "Kernel codebooks for scene categorization," in *European Conference on Computer Vision*, 2008.

[25] K. Sande, T. Gevers, and C. Snoek, "Evaluation of color descriptors for object and scene recognition," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2008.