
Multiview Fisher Discriminant Analysis

Tom Diethe, David R. Hardoon, John Shawe-Taylor
University College London
{t.diethe,d.hardoon,jst}@cs.ucl.ac.uk

Abstract

CCA can be seen as a multiview extension of PCA, in which information from two sources is used for learning by finding a subspace in which the two views are most correlated. However PCA, and by extension CCA, does not use label information. Fisher Discriminant Analysis uses label information to find informative projections, which can be more informative in supervised learning settings. We show that FDA and its dual can both be formulated as generalized eigenproblems, enabling a kernel formulation. We derive a regularised two-view equivalent of Fisher Discriminant Analysis and its corresponding dual, both of which can also be formulated as generalized eigenproblems. We then show that these can be cast as equivalent disciplined convex optimisation problems, and subsequently extended to multiple views. We show experimental results on an EEG dataset and part of the PASCAL 2007 VOC challenge dataset.

1 Introduction

The motivation for this paper comes from the desire to combine multiple sources of information in a learning framework where labels are known. Canonical correlation analysis (CCA), introduced by Hotelling in 1936 [1], is a method of correlating linear relationships between two sets of multidimensional variables. CCA makes use of two views of the same underlying semantic object to extract a common representation. Kernel CCA (KCCA) is a generalised form of kernel independent components analysis [2], resulting in a nonlinear version of CCA [3]. However both CCA and KCCA are effectively unsupervised techniques, and as such are not ideally suited to a classification setting. A common way of performing classification on two-view data using KCCA is to use the projected data from one of the views as input to a standard classification algorithm, such as a Support Vector Machine (SVM). However, the subspace that is learnt through such unsupervised methods may not always align well with the label space.

1.1 Subspace Learning

In standard single view subspace learning, a parallel can be drawn between subspace projections that are independent of the label space, such as Principal Components Analysis (PCA), and those that incorporate label information, such as Fisher Discriminant Analysis (FDA). PCA searches for directions in the data that have largest variance and project the data onto a subset of these directions. In this way, we obtain a lower dimensional representation of the data that captures most of the variance. PCA is an unsupervised technique and as such does not include label information of the data. For instance, if we are given 2-dimensional data from two classes forming two long and thin clusters, such that the clusters are positioned in parallel and very closely together, the total variance ignoring the labels would be in the lengthways direction of the clusters. For classification, this would be a poor projection, because the labels would be evenly mixed. A much more useful projection would be orthogonal to the clusters, i.e. in the direction of least overall variance, which would perfectly separate the two classes. We would then perform classification in this 1-dimensional space. FDA would find exactly this projection.

1.2 Multiview Learning

Multiview learning algorithms seek to find some common subspace between two sets of random variables, implicitly making the assumption that they are views of the same underlying semantic object. CCA can be viewed as finding basis vectors for two sets of variables such that the correlations between the projections onto these basis vectors $x_a = \mathbf{w}'_a \phi_a(x)$ and $x_b = \mathbf{w}'_b \phi_b(x)$ are mutually maximised. Kernel CCA (KCCA) uses the so called “kernel trick” to produce a nonlinear version of CCA. Each of the two views of the data are projected into distinct feature spaces before performing CCA in the new feature space. This leads to the kernelised form, KCCA

$$\begin{aligned} \max_{\alpha, \beta} \quad & \alpha' \mathbf{K}_a \mathbf{K}_b \beta \\ \text{s.t.} \quad & \alpha' \mathbf{K}_a^2 \alpha = 1 \text{ and } \beta' \mathbf{K}_b^2 \beta = 1 \end{aligned} \quad (1)$$

where \mathbf{K}_a and \mathbf{K}_b are the kernel matrices of the two views and α and β are the dual weight vectors.

When two views of the same phenomenon are available KCCA has been shown to be an effective preprocessing step that can improve the performance of classification algorithms such as the Support Vector Machine (SVM) [3]. SVM-2K [4] was an attempt to take this to its logical conclusion by combining this two stage learning into a single optimisation. The authors presented both experimental and theoretical analysis of the approach, showing encouraging results and insights. The algorithm combines the two steps by introducing the constraint of similarity between two 1-dimensional projections identifying two distinct SVMs the two feature spaces. The constraint slightly different to the 2-norm constraint of KCCA. However SVM-2K requires extra parameters (the C -parameter for each SVM, and another mixing parameter, along with any kernel parameters) that our method will not require. In addition, it is not easy to see how SVM-2K can be generalised to more than two views.

2 Regularised Kernel Fisher Discriminant Analysis

We will be basing the derivation on the form given by [5] for regularised Kernel Fisher Discriminant Analysis, which we review briefly here. Let $(\mathbf{x}, y) \sim S$ be an input-output pair from an m -sample S with $\mathbf{x} \in \mathbb{R}^n$ an n -dimensional vector and $y \in \mathbb{R}$. Given a matrix of inputs $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_m)'$ with the corresponding labels $\mathbf{y} = (y_1, \dots, y_m)'$ and $(\cdot)'$ denotes the transpose of a vector, the regularised Fisher discriminant chooses \mathbf{w} to solve the following optimisation problem

$$\rho = \max_{\mathbf{w}} \frac{\mathbf{w} \mathbf{X}' \mathbf{y} \mathbf{y}' \mathbf{X} \mathbf{w}}{\mathbf{w}' \mathbf{X}' \mathbf{B} \mathbf{X} \mathbf{w} + \mu \|\mathbf{w}\|^2} \quad (2)$$

where \mathbf{B} is a matrix incorporating the label information and the balance of the dataset as follows:

$$\mathbf{B} = \mathbf{D} - \mathbf{C}^+ - \mathbf{C}^-$$

where \mathbf{D} is a diagonal matrix with entries

$$\mathbf{D}_{ii} = \begin{cases} 2\ell^-/\ell & \text{if } y_i = +1 \\ 2\ell^+/\ell & \text{if } y_i = -1 \end{cases}$$

and \mathbf{C}^+ and \mathbf{C}^- are given by

$$\mathbf{C}_{ij}^+ = \begin{cases} 2\ell^-/(\ell\ell^+) & \text{if } y_i = +1 = y_j \\ 0 & \text{otherwise} \end{cases}$$

$$\mathbf{C}_{ij}^- = \begin{cases} 2\ell^+ / (\ell\ell^-) & \text{if } y_i = -1 = y_j \\ 0 & \text{otherwise} \end{cases}$$

Following Shawe-Taylor and Christianini [5], the form for Kernel FDA is given by

$$(\mathbf{BK} + \mu\mathbf{I})\boldsymbol{\alpha} = \mathbf{y} \quad (3)$$

3 Regularised Two view Fisher Discriminant Analysis (FDA2)

In this section we present the two view extension of FDA. The method by which this is achieved close resembles the way CCA can be viewed as a multiview extension of PCA. The key motivation for why this is feasible is as follows. Since standard FDA is Bayes optimal for two normal distributions with equal covariance (and therefore kernel FDA (KFDA) is Bayes optimal for two normal distributions in the feature space), it follows that the outputs of the classifier can be seen as probabilities. In fact there is a direct equivalence between KFDA and the Relevance Vector Machine given particular choices of noise models and regularisation [6]. This means that if we simply sum the outputs of FDA or KFDA trained on two different view of the same data, the output will always be that of the most confident classifier. We are now given examples drawn from two views of the same underlying semantic object, denoted as \mathbf{X}_a and \mathbf{X}_b respectively. These examples are assumed to be paired, and as such have the same labelling. The classification function is then,

$$f(x_i) = \text{sgn}(\langle \mathbf{w}_a, \phi_a(x_i) \rangle + \langle \mathbf{w}_b, \phi_b(x_i) \rangle + b) \quad (4)$$

where $b = b_a + b_b$ is chosen to bisect the two centres of mass of each view, i.e.

$$\begin{aligned} b_a &= 0.5\mathbf{w}_a'(\mathbf{X}_a'(\mathbf{y} \cdot \mathbf{t})), \\ b_b &= 0.5\mathbf{w}_b'(\mathbf{X}_b'(\mathbf{y} \cdot \mathbf{t})), \\ \mathbf{t} &= \{1/\ell^+ \quad \text{if } y_i = +1, 1/\ell^- \quad \text{if } y_i = -1\}. \end{aligned}$$

In fact it is easy to see that this can be generalised to k views as follows:

$$f(x_i) = \text{sgn}\left(\sum_{j=1}^k \langle \mathbf{w}_j, \phi_j(x_i) \rangle + b\right)$$

We use a single regularisation parameter μ to constrain both sets of weights in the denominator. The two view Fisher discriminant chooses two sets of weights \mathbf{w}_a and \mathbf{w}_b to solve the following optimisation problem

$$\rho = \frac{\mathbf{w}_a' \mathbf{X}_a' \mathbf{y} \mathbf{y}' \mathbf{X}_b \mathbf{w}_b}{\sqrt{(\mathbf{w}_a' \mathbf{X}_a' \mathbf{B} \mathbf{X}_a \mathbf{w}_a + \mu \|\mathbf{w}_a\|^2) \cdot (\mathbf{w}_b' \mathbf{X}_b' \mathbf{B} \mathbf{X}_b \mathbf{w}_b + \mu \|\mathbf{w}_b\|^2)}} \quad (5)$$

where \mathbf{w}_a and \mathbf{w}_b are the weight vectors for each view. Since the equation is not affected by rescaling of \mathbf{w}_a or \mathbf{w}_b , the optimisation can be subjected to the following constraints

$$\begin{aligned} \mathbf{w}_a' \mathbf{X}_a' \mathbf{B} \mathbf{X}_a \mathbf{w}_a + \mu \|\mathbf{w}_a\|^2 &= 1 \\ \mathbf{w}_b' \mathbf{X}_b' \mathbf{B} \mathbf{X}_b \mathbf{w}_b + \mu \|\mathbf{w}_b\|^2 &= 1 \end{aligned}$$

We then introduce the Lagrange multipliers λ_a and λ_b , the corresponding Lagrangian for this optimization is:

$$L = \mathbf{w}'_a \mathbf{X}'_a \mathbf{y} \mathbf{y}' \mathbf{X}_b \mathbf{w}_b - \frac{\lambda_a}{2} (\mathbf{w}'_a \mathbf{X}'_a \mathbf{B} \mathbf{X}_a \mathbf{w}_a + \mu \|\mathbf{w}_a\|^2 - 1) - \frac{\lambda_b}{2} (\mathbf{w}'_b \mathbf{X}'_b \mathbf{B} \mathbf{X}_b \mathbf{w}_b + \mu \|\mathbf{w}_b\|^2 - 1)$$

Differentiating with respect to the weight vectors \mathbf{w}_a and \mathbf{w}_b , we have:

$$\frac{\partial L}{\partial \mathbf{w}_a} = \mathbf{X}'_a \mathbf{y} \mathbf{y}' \mathbf{X}_b \mathbf{w}_b - \lambda_a (\mathbf{X}'_a \mathbf{B} \mathbf{X}_a \mathbf{w}_a + \mu \mathbf{w}_a) = 0 \quad (6)$$

$$\frac{\partial L}{\partial \mathbf{w}_b} = \mathbf{w}'_a \mathbf{X}'_a \mathbf{y} \mathbf{y}' \mathbf{X}_b - \lambda_b (\mathbf{X}'_b \mathbf{B} \mathbf{X}_b \mathbf{w}_b + \mu \mathbf{w}_b) = 0 \quad (7)$$

$$(8)$$

If we multiply 6 and 7 by \mathbf{w}_a and \mathbf{w}_b respectively, we find that $\lambda_a = \lambda_b$, and as such only one Lagrange multiplier is needed. Using this fact, we can rearrange 6 to give

$$\mathbf{w}_a = \frac{(\mathbf{X}'_a \mathbf{B} \mathbf{X}_a + \mu \mathbf{I})^{-1} \mathbf{X}'_a \mathbf{y} \mathbf{y}' \mathbf{X}_b \mathbf{w}_b}{\lambda} \quad (9)$$

Here we must be aware that the matrix $\mathbf{M} = (\mathbf{X}'_a \mathbf{B} \mathbf{X}_a + \mu \mathbf{I})$ may not be invertible. In such cases we use the pseudo-inverse. Substituting 9 into 7 gives,

$$\mathbf{X}'_b \mathbf{y} \mathbf{y}' \mathbf{X}_a (\mathbf{X}'_a \mathbf{B} \mathbf{X}_a + \mu \mathbf{I})^{-1} \mathbf{X}'_a \mathbf{y} \mathbf{y}' \mathbf{X}_b \mathbf{w}_b = \lambda^2 \mathbf{X}'_b \mathbf{B} \mathbf{X}_b \mathbf{w}_b \quad (10)$$

Defining $\mathbf{R}_{ab} = \mathbf{X}'_a \mathbf{y} \mathbf{y}' \mathbf{X}_b$ and $\mathbf{R}_{ba} = \mathbf{X}'_b \mathbf{y} \mathbf{y}' \mathbf{X}_a$, it can be clearly seen that equation 10 is a generalised eigenproblem of the form $Ax = \lambda Bx$ where $A = \mathbf{R}_{ba} \mathbf{M}^{-1} \mathbf{R}_{ab}$, $B = \mathbf{X}'_b \mathbf{B} \mathbf{X}_b$, and $x = \mathbf{w}_b$, and the resulting solutions for λ need to be square rooted. The full solution is then given by taking the maximal eigenvalue and substituting into 9.

4 Kernel Two View Fisher Discriminant Analysis (FDA-2K)

We will now derive the dual form of FDA-2. The same caveats regarding convexity apply as in the construction of the dual form of FDA. We introduce two dual weight vectors through the substitutions $\mathbf{w}_a = \mathbf{X}'_a \boldsymbol{\alpha}$ and $\mathbf{w}_b = \mathbf{X}'_b \boldsymbol{\beta}$, and a single regularisation parameter κ for both sets of weights, giving the following optimisation,

$$\rho = \frac{\boldsymbol{\alpha} \mathbf{X}_a \mathbf{X}'_a \mathbf{y} \mathbf{y}' \mathbf{X}_b \mathbf{X}'_b \boldsymbol{\beta}}{\sqrt{(\boldsymbol{\alpha} \mathbf{X}_a \mathbf{X}'_a \mathbf{B} \mathbf{X}_a \mathbf{X}'_a \boldsymbol{\alpha} + \kappa \|\mathbf{w}_a\|^2) \cdot (\boldsymbol{\beta} \mathbf{X}_b \mathbf{X}'_b \mathbf{B} \mathbf{X}_b \mathbf{X}'_b \boldsymbol{\beta} + \kappa \|\mathbf{w}_b\|^2)}} \quad (11)$$

The kernel form of this is then,

$$\rho = \frac{\boldsymbol{\alpha} \mathbf{K}_a \mathbf{y} \mathbf{y}' \mathbf{K}_b \boldsymbol{\beta}}{\sqrt{(\boldsymbol{\alpha} \mathbf{K}_a \mathbf{B} \mathbf{K}_a \boldsymbol{\alpha} + \kappa \boldsymbol{\alpha} \mathbf{K}_a \boldsymbol{\alpha}) \cdot (\boldsymbol{\beta} \mathbf{K}_b \mathbf{B} \mathbf{K}_b \boldsymbol{\beta} + \kappa \boldsymbol{\beta} \mathbf{K}_b \boldsymbol{\beta})}} \quad (12)$$

As in the primal form, the equation is not affected by rescaling of $\boldsymbol{\alpha}$ or $\boldsymbol{\beta}$, hence the optimisation can be subjected to the following constraints

$$\boldsymbol{\alpha} \mathbf{K}_a \mathbf{B} \mathbf{K}_a \boldsymbol{\alpha} + \kappa \boldsymbol{\alpha} \mathbf{K}_a \boldsymbol{\alpha} = 1 \quad (13)$$

$$\boldsymbol{\beta} \mathbf{K}_b \mathbf{B} \mathbf{K}_b \boldsymbol{\beta} + \kappa \boldsymbol{\beta} \mathbf{K}_b \boldsymbol{\beta} = 1 \quad (14)$$

where \mathbf{K}_a and \mathbf{K}_b are the kernels for each of the two views respectively. Given Lagrange multipliers λ_a and λ_b , the corresponding Lagrangian for this optimisation is

$$L = \alpha \mathbf{K}_a \mathbf{y} \mathbf{y}' \mathbf{K}_b \beta - \frac{\lambda_a}{2} (\alpha \mathbf{K}_a \mathbf{B} \mathbf{K}_a \alpha + \kappa \alpha \mathbf{K}_a \alpha - 1) - \frac{\lambda_b}{2} (\beta \mathbf{K}_b \mathbf{B} \mathbf{K}_b \beta + \kappa \beta \mathbf{K}_b \beta - 1) \quad (15)$$

Differentiating with respect to the weight vectors α and β , we have:

$$\frac{\partial L}{\partial \alpha} = \mathbf{K}_a \mathbf{y} \mathbf{y}' \mathbf{K}_b \beta - \lambda_a (\mathbf{K}_a \mathbf{B} \mathbf{K}_a \alpha + \kappa \mathbf{K}_a \alpha) = 0 \quad (16)$$

$$\frac{\partial L}{\partial \beta} = \mathbf{K}_b \mathbf{y} \mathbf{y}' \mathbf{K}_a \alpha - \lambda_b (\mathbf{K}_b \mathbf{B} \mathbf{K}_b \beta + \kappa \mathbf{K}_b \beta) = 0 \quad (17)$$

$$(18)$$

If we multiply 16 and 17 by α and β respectively, we find that as in the primal version, $\lambda_a = \lambda_b$, and as such only one Lagrange multiplier is needed. And as before, rearranging 16 gives

$$\alpha = \frac{(\mathbf{K}_a \mathbf{B} \mathbf{K}_a + \kappa \mathbf{K}_a)^{-1} \mathbf{K}_a \mathbf{y} \mathbf{y}' \mathbf{K}_b \beta}{\lambda} \quad (19)$$

Here we must be aware that the matrix $\mathbf{M} = (\mathbf{K}_a \mathbf{B} \mathbf{K}_a + \kappa \mathbf{K}_a)$ may not be invertible. In such cases we use the pseudo-inverse. Substituting 19 into 17 gives,

$$\mathbf{K}_b \mathbf{y} \mathbf{y}' \mathbf{K}_a (\mathbf{K}_a \mathbf{B} \mathbf{K}_a + \kappa \mathbf{K}_a)^{-1} \mathbf{K}_a \mathbf{y} \mathbf{y}' \mathbf{K}_b \beta = \lambda^2 \mathbf{K}_b \mathbf{B} \mathbf{K}_b \beta \quad (20)$$

Defining $\mathbf{K}_{ab} = \mathbf{K}_a \mathbf{y} \mathbf{y}' \mathbf{K}_b$ and $\mathbf{K}_{ba} = \mathbf{K}_b \mathbf{y} \mathbf{y}' \mathbf{K}_a$, it can be clearly seen that equation 20 is a generalised eigenproblem of the form $Ax = \lambda Bx$ where $A = \mathbf{K}_{ba} \mathbf{M}^{-1} \mathbf{K}_{ab}$, $B = \mathbf{K}_b \mathbf{B} \mathbf{K}_b$, and $x = \beta$, and the resulting solutions for λ need to be square rooted. The full solution is then given by taking the maximal eigenvalue and substituting into 19.

5 Convex Formulations

We follow the approach outlined by [6] and provide a disciplined convex form of Multiview Fisher Discriminant Analysis. Since we are trying to minimise the variance of the data along the projection whilst maximising the distance between the average outputs for each class over all of the views. This can be done in two ways, which we have denoted as the block method and concatenation method.

5.1 Block Method

In this formulation we are effectively learning individual FDA classifiers constrained to have equal slacks. The resulting formulation is as follows,

$$\begin{aligned} \min_{\alpha, \beta, b, \xi} \quad & \|\xi\|_2^2 + \lambda P(\alpha, \beta) \\ \text{s.t.} \quad & \mathbf{K}_a \alpha + \mathbf{1}b = \mathbf{y} + \xi \\ & \mathbf{K}_b \beta + \mathbf{1}b = \mathbf{y} + \xi \\ & \mathbf{1}'_i \xi = 0 \text{ for } i = 1, 2 \end{aligned} \quad (21)$$

The first constraint ensures that the average loss between the output and its class label is minimised. The second constraint ensures that the average output for each class is each label.

5.2 Concatenation Method

In this formulation we are constraining the sum of the FDA classification outputs through the slack variables ξ . The formulation is as follows,

$$\begin{aligned} \min_{\alpha, \beta, b, \xi} \quad & \|\xi\|_2^2 + \lambda P(\alpha, \beta) & (22) \\ \text{s.t.} \quad & \mathbf{K}_a \alpha + \mathbf{K}_b \beta + \mathbf{1}b = \mathbf{y} + \xi \\ & \mathbf{1}'_i \xi = 0 \quad \text{for } i = 1, 2 \end{aligned}$$

Note that in this case there is a single offset parameter b . This second formulation requires less memory as the kernel matrices are simply concatenated, as opposed to the creation of a block diagonal matrix. However the problem is more tightly constrained, and as a result less flexible. We conducted experiments using both methods.

5.3 Regularisation and Loss functions

The natural choices for the regularisation function $P(\alpha, \beta)$ would either be the l_2 -norm of the dual weight vectors, i.e. $P(\alpha, \beta) = \|\alpha\|_2^2 + \|\beta\|_2^2$, or the l_2 -norm of the primal weight vector $P(\alpha, \beta) = \alpha' \mathbf{K}_a \alpha + \beta' \mathbf{K}_b \beta$. However more interesting is the l_1 -norm of the dual weight vector, $P(\alpha, \beta) = \|\alpha\|_1 + \|\beta\|_1$, as this choice leads to sparse solutions due to the fact that the l_1 -norm can be seen as an approximation to the l_0 -norm.

We can also follow [7] and remove the assumption of a Gaussian noise model, resulting in different loss functions on the slacks ξ . For example, if we choose a Laplacian noise model we can simply replace $\|\xi\|_2^2$ with $\|\xi\|_1$ in the objective function. The advantage of this is if the l_1 -norm regulariser from above is chosen, the resulting optimisation is a linear programme, which can be solved efficiently using methods such as column generation.

5.4 Extending to Multiple Views

The above formulations (equations 21 and 22) lead to natural extensions to more than two views. For k views, we denote \mathbf{K}_j as the kernel for the j^{th} view, and $\alpha_j, j = 1, \dots, k$ as the set of dual weight vectors. The optimisations are then,

$$\min_{\alpha_j, b, \xi} H(\xi) + \lambda P(\alpha_j), \quad j = 1, \dots, k \quad (23)$$

$$\text{s.t.} \quad \mathbf{K}_j \alpha_j + \mathbf{1}b = \mathbf{y} + \xi \quad j = 1, \dots, k \quad (24)$$

$$\mathbf{1}'_i \xi = 0 \quad i = 1, 2 \quad (25)$$

$$\text{for the block method and,} \quad (26)$$

$$(27)$$

$$\min_{\alpha_j, b, \xi} H(\xi) + \lambda P(\alpha_j), \quad j = 1, \dots, k \quad (28)$$

$$\text{s.t.} \quad \sum_{j=1}^k \mathbf{K}_j \alpha_j + \mathbf{1}b = \mathbf{y} + \xi \quad (29)$$

$$(30)$$

for the concatenation method, where in both cases $H(\cdot)$ is the loss function. If we then apply the Laplacian noise model and use the l_1 -norm regularisation as described above, we have Sparse Multiview Fisher Discriminant Analysis (SMFDA).

6 Experiments

6.1 EEG & Music Dataset

Here we present results on dataset produced by an EEG experiment analysed in [8]. The principal hypothesis was that neural patterns should reflect relative changes in the key of music that a listener is attending to. The classification results using SMFDA (concatenation method) are given in table 1. It can be seen that this method is able to classify between the tonal and atonal experimental conditions perfectly. As a comparison, we trained an SVM on the projection of the EEG data into the shared feature space, using a linear kernel and 5-fold cross validation to select the C parameter, and KCCA trained on both views followed by classification using an SVM on the projected EEG data.

Table 1: Test errors for within-subject classification for Tonal vs Atonal. The SVM classification results are on the EEG data alone. The KCCA + SVM, and SMFDA classification used a kernel on the music as a second view. ** denotes significance at the $p < 0.001$ level

Classifier	# Train	# Test	Linear
SVM (Linear)	1152	383	0.2298**
SVM (RBF)	1152	383	0.1175**
KCCA + SVM (linear)	1152	383	0.0157**
SMFDA	1152	393	0.0000**

6.2 VOC 2007 Dataset

The features that we used can be found in [9], with an extra feature extraction method known as Histogram of interest point (SIFT). We constructed rbf kernels for each of these features setting the width parameter using a heuristic method. We use the VOC2007 challenge database which contains 9963 images, each with at least 1 object. The number of objects in each image ranges from 1 to 20, with, for instance, objects of people, sheep, horses, cats, dogs etc. For a complete list of the objects, and description of the data set see the VOC2007 challenge website ¹.

Figure 6.2 shows Recall-Precision curves for SMFDA with 1, 2, 3 or 11 kernels and PicSOM. For the purposes of training, we used a random subset of 200 irrelevant images rather than the full training set. The results show that adding more kernels into the optimisation can assist in recall performance, although since the best kernel (based on SIFT features) performs well alone, the improvement is not that marked. Results are competitive with the PicSOM algorithm, which uses all 11 kernels, and all of the irrelevant images.

7 Conclusions

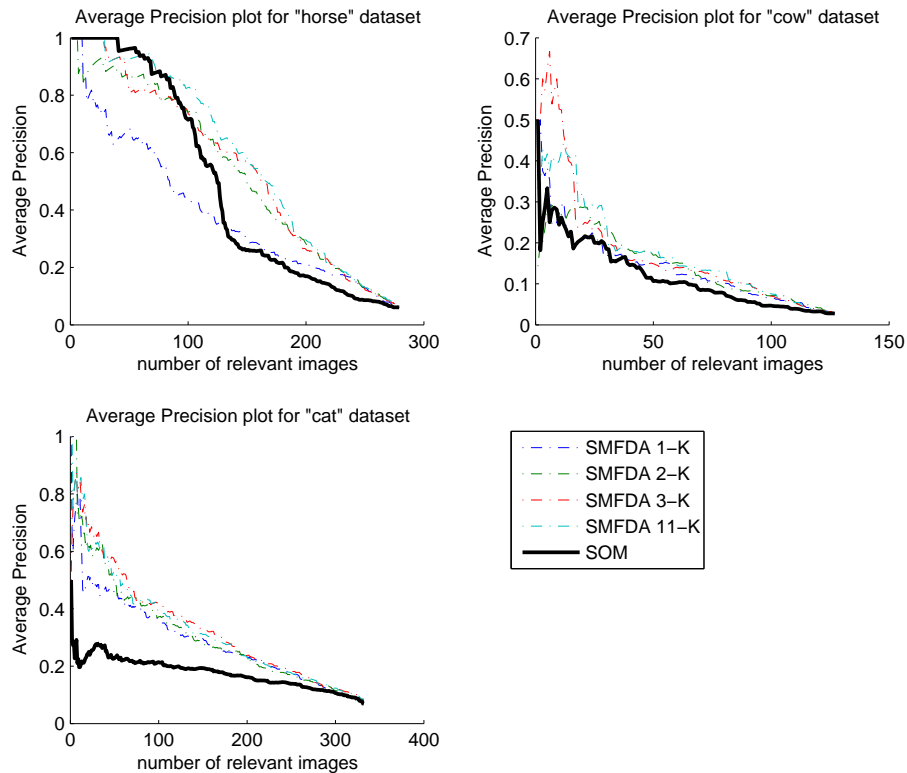
We have derived a regularised two-view equivalent of Fisher Discriminant Analysis and its corresponding dual, both of which can also be formulated as generalized eigenproblems. We then demonstrated that these can be cast as equivalent disciplined convex optimisation problems, and subsequently extended to multiple views. We show experimental results on an EEG dataset and part of the PASCAL 2007 VOC challenge dataset, which show the validity of the method. We are currently conducting further experiments, and examining ways to improve the efficiency of the algorithm.

References

- [1] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:312–377, 1936.
- [2] F. Bach and M. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

¹<http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html>

Figure 1: Average precision recall curves for 3 VOC 2007 datasets for Multiview Fisher Discriminant Analysis using the Block method plotted against SOM results



- [3] D.R. Hardoon, S.Szedmak, and J.Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.
- [4] Jason D. R. Farquhar, David Hardoon, Hongying Meng, John Shawe-Taylor, and Sandor Szedmak. Two view learning: Svm-2k, theory and practice. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 355–362. MIT Press, Cambridge, MA, 2006.
- [5] John Shawe-Taylor and Nello Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, U.K., 2004.
- [6] S. Mika, A. J. Smola, and B. Schölkopf. An improved training algorithm for kernel fisher discriminants. In *In AISTATS*, pages 98–104, 2001.
- [7] S. Mika, G. Rätsch, and K.-R. Müller. A mathematical programming approach to the kernel fisher algorithm. In T.K. Leen, T.G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13, pages 591–597, 2001.
- [8] T. Diethe, S. Durrant, J. Shawe-Taylor, and H. Neubauer. Semantic dimensionality reduction for the classification of EEG according to musical tonality. Technical report, Presented at the NIPS 2008 workshop Learning from Multiple Sources, 2008.
- [9] V. Viitaniemi and J. Laaksonen. Techniques for image classification, object detection and object segmentation applied to VOC challenge 2007. Technical report, Department of Information and Computer Science, Helsinki University of Technology (TKK), 2008.