LETTER **Communicated by Erkki Oja**

Canonical Correlation Analysis: An Overview with Application to Learning Methods

David R. Hardoon
drh@ecs.soton.ac.uk
Sandor Szedmak
ss03v@ecs.soton.ac.uk
John Shawe-Taylor
jst@ecs.soton.ac.uk
School of Electronics and Computer Science, Image, Speech and Intelligent Systems Research Group, University of Southampton, Southampton S017 1BJ, U.K.

We present a general method using kernel canonical correlation analysis to learn a semantic representation to web images and their associated text. The semantic space provides a common representation and enables a comparison between the text and images. In the experiments, we look at two approaches of retrieving images based on only their content from a text query. We compare orthogonalization approaches against a standard cross-representation retrieval technique known as the generalized vector space model.

1 Introduction

During recent years, there have been advances in data learning using kernel methods. Kernel representation offers an alternative learning to nonlinear functions by projecting the data into a high-dimensional feature space to increase the computational power of the linear learning machines, though this still leaves unresolved the issue of how best to choose the features or the kernel function in ways that will improve performance. We review some of the methods that have been developed for learning the feature space.

Principal component analysis (PCA) is a multivariate data analysis procedure that involves a transformation of a number of possibly correlated variables into a smaller number of uncorrelated variables known as principal components. PCA makes use of only the training inputs while making no use of the labels.

Independent component analysis (ICA), in contrast to correlation-based transformations such as PCA, not only decorrelates the signals but also reduces higher-order statistical dependencies, attempting to make the signals as independent as possible. In other words, ICA is a way of finding a linear, not only orthogonal, coordinate system in any multivariate data. The

directions of the axes of this coordinate system are determined by both the second- and higher-order statistics of the original data. The goal is to perform a linear transform that makes the resulting variables as statistically independent from each other as possible.

Partial least squares (PLS) is a method similar to canonical correlation analysis. It selects feature directions that are useful for the task at hand, though PLS uses only one view of an object and the label as the corresponding pair. PLS could be thought of as a method that looks for directions that are good at distinguishing the different labels.

Canonical correlation analysis (CCA) is a method of correlating linear relationships between two multidimensional variables. CCA can be seen as using complex labels as a way of guiding feature selection toward the underlying semantics. CCA makes use of two views of the same semantic object to extract the representation of the semantics.

Proposed by Hotelling in 1936, CCA can be seen as the problem of finding basis vectors for two sets of variables such that the correlations between the projections of the variables onto these basis vectors are mutually maximized. In an attempt to increase the flexibility of the feature selection, kernelization of CCA (KCCA) has been applied to map the hypotheses to a higher-dimensional feature space. KCCA has been applied in some preliminary work by Fyfe and Lai (2001) and Akaho (2001) and the recent Vinokourov, Shawe-Taylor, and Cristianini (2002) with improved results.

During recent years, there has been a vast increase in the amount of multimedia content available both off-line and online, though we are unable to access or make use of these data unless they are organized in such a way as to allow efficient browsing. To enable content-based retrieval with no reference to labeling, we attempt to learn the semantic representation of images and their associated text. We present a general approach using KCCA that can be used for content-based retrieval (Hardoon & Shawe-Taylor, 2003) and mate-based retrieval (Vinokourov, Hardoon, & Shawe-Taylor, 2003; Hardoon & Shawe-Taylor, 2003). In both cases, we compare the KCCA approach to the generalized vector space model (GVSM), which aims at capturing some term-term correlations by looking at co-occurrence information.

This study aims to serve as a tutorial and give additional novel contributions in several ways. We follow the work of Borga (1999) where we represent the eigenproblem as two eigenvalue equations, as this allows us to reduce the computation time and dimensionality of the eigenvectors. Further to that, we follow the idea of Bach and Jordan (2002) to compute a new correlation matrix with reduced dimensionality. Though Bach and Jordan address a very different problem, they use the same underlining technique of Cholesky decomposition to re-represent the kernel matrices. We show that using partial Gram-Schmidt orthogonalization (Cristianini, Shawe-Taylor, & Lodhi, 2001) is equivalent to incomplete Cholesky decomposition, in the sense that incomplete Cholesky decomposition can be seen

as a dual implementation of partial Gram-Schmidt. We show that the general approach can be adapted to two different types of problems, content, and mate retrieval by changing only the selection of eigenvectors used in the semantic projection. And to simplify the learning of the KCCA, we explore a method of selecting the regularization parameter a priori such that it gives a value that performs well in several different tasks.

In this study, we also present a generalization of the framework for canonical correlation analysis. Our approach is based on the work of Gifi (1990) and Ketterling (1971). The purpose of the generalization is to extend the canonical correlation as an associativity measure between two set of variables to more than two sets, while preserving most of its properties. The generalization starts with the optimization problem formulation of canonical correlation. By changing the objective function, we will arrive at the multiset problem. Applying similar constraint sets in the optimization problems, we find that the feasible solutions are singular vectors of matrices, which are derived in the same way for the original and generalized problem.

In section 2, we present the theoretical background of CCA. In section 3, we present the CCA and KCCA algorithm. Approaches to deal with the computational problems that arose in section 3 are presented in section 4. Our experimental results are presented in section 5, and section 6 draws final conclusions.

2 Theoretical Foundations

Proposed by Hotelling in 1936, Canonical correlation analysis can be seen as the problem of finding basis vectors for two sets of variables such that the correlation between the projections of the variables onto these basis vectors is mutually maximized. Correlation analysis is dependent on the coordinate system in which the variables are described, so even if there is a very strong linear relationship between two sets of multidimensional variables, depending on the coordinate system used, this relationship might not be visible as a correlation. Canonical correlation analysis seeks a pair of linear transformations, one for each of the sets of variables, such that when the set of variables is transformed, the corresponding coordinates are maximally correlated.

Definition 1. *$\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product of the vectors $\mathbf{x}$,$\mathbf{y}$, and it is equal to $\mathbf{x}'\mathbf{y}$, where we use A' to denote the transpose of a vector or matrix A.*

Consider a multivariate random vector of the form $(\mathbf{x}, \mathbf{y})$. Suppose we are given a sample of instances $S = ((\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n))$ of $(\mathbf{x}, \mathbf{y})$. Let S_x denote $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ and similarly S_y denote $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$. We can consider defining a new coordinate for $\mathbf{x}$ by choosing a direction $\mathbf{w}_x$ and projecting

$\mathbf{x}$ onto that direction,

$$\mathbf{x} \to \langle \mathbf{w}_x, \mathbf{x} \rangle.$$

If we do the same for $\mathbf{y}$ by choosing a direction $\mathbf{w}_y$, we obtain a sample of the new $\mathbf{x}$ coordinate. Let

$$S_{x,\mathbf{w}_x} = (\langle \mathbf{w}_x, \mathbf{x}_1 \rangle, \ldots, \langle \mathbf{w}_x, \mathbf{x}_n \rangle),$$

with the corresponding values of the new $\mathbf{y}$ coordinate being

$$S_{y,\mathbf{w}_y} = \left(\langle \mathbf{w}_y, \mathbf{y}_1 \rangle, \ldots, \langle \mathbf{w}_y, \mathbf{y}_n \rangle\right).$$

The first stage of canonical correlation is to choose $\mathbf{w}_x$ and $\mathbf{w}_y$ to maximize the correlation between the two vectors. In other words, the function's result to be maximized is

$$\begin{aligned} \rho &= \max_{\mathbf{w}_x, \mathbf{w}_y} corr(S_x \mathbf{w}_x, S_y \mathbf{w}_y) \\ &= \max_{\mathbf{w}_x, \mathbf{w}_y} \frac{\langle S_x \mathbf{w}_x, S_y \mathbf{w}_y \rangle}{\|S_x \mathbf{w}_x\| \|S_y \mathbf{w}_y\|}. \end{aligned}$$

If we use $\hat{\mathbb{E}}[f(\mathbf{x}, \mathbf{y})]$ to denote the empirical expectation of the function $f(\mathbf{x}, \mathbf{y})$, where

$$\hat{\mathbb{E}}[f(\mathbf{x}, \mathbf{y})] = \frac{1}{m} \sum_{i=1}^{m} f(\mathbf{x}_i, \mathbf{y}_i),$$

we can rewrite the correlation expression as

$$\begin{aligned} \rho &= \max_{\mathbf{w}_x, \mathbf{w}_y} \frac{\hat{\mathbb{E}}[\langle \mathbf{w}_x, \mathbf{x} \rangle \langle \mathbf{w}_y, \mathbf{y} \rangle]}{\sqrt{\hat{\mathbb{E}}[\langle \mathbf{w}_x, \mathbf{x} \rangle^2] \hat{\mathbb{E}}[\langle \mathbf{w}_x, \mathbf{x} \rangle^2]}} \\ &= \max_{\mathbf{w}_x, \mathbf{w}_y} \frac{\hat{\mathbb{E}}[\mathbf{w}_x' \mathbf{x} \mathbf{y}' \mathbf{w}_y]}{\sqrt{\hat{\mathbb{E}}[\mathbf{w}_x' \mathbf{x} \mathbf{x}' \mathbf{w}_x] \hat{\mathbb{E}}[\mathbf{w}_y' \mathbf{y} \mathbf{y}' \mathbf{w}_y]}}. \end{aligned}$$

It follows that

$$\rho = \max_{\mathbf{w}_x, \mathbf{w}_y} \frac{\mathbf{w}_x' \hat{\mathbb{E}}[\mathbf{x} \mathbf{y}'] \mathbf{w}_y}{\sqrt{\mathbf{w}_x' \hat{\mathbb{E}}[\mathbf{x} \mathbf{x}'] \mathbf{w}_x \mathbf{w}_y' \hat{\mathbb{E}}[\mathbf{y} \mathbf{y}'] \mathbf{w}_y}}.$$

Now observe that the covariance matrix of $(\mathbf{x}, \mathbf{y})$ is

$$C(\mathbf{x}, \mathbf{y}) = \hat{\mathbb{E}}\left[\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}'\right] = \begin{bmatrix} C_{\mathbf{x}\mathbf{x}} & C_{\mathbf{x}\mathbf{y}} \\ C_{\mathbf{y}\mathbf{x}} & C_{\mathbf{y}\mathbf{y}} \end{bmatrix} = C. \tag{2.1}$$

The total covariance matrix C is a block matrix where the within-sets covariance matrices are $C_{\mathbf{xx}}$ and $C_{\mathbf{yy}}$ and the between-sets covariance matrices are $C_{\mathbf{xy}} = C'_{\mathbf{yx}}$, although equation 2.1 is the covariance matrix only in the zero-mean case.

Hence, we can rewrite the function ρ as

$$\rho = \max_{\mathbf{w}_x, \mathbf{w}_y} \frac{\mathbf{w}'_x C_{\mathbf{xy}} \mathbf{w}_y}{\sqrt{\mathbf{w}'_x C_{\mathbf{xx}} \mathbf{w}_x \mathbf{w}'_y C_{\mathbf{yy}} \mathbf{w}_y}}. \tag{2.2}$$

The maximum canonical correlation is the maximum of ρ with respect to $\mathbf{w}_x$ and $\mathbf{w}_y$.

3 Algorithm

In this section, we give an overview of the CCA and KCCA, algorithms where we formulate the optimization problem as a standard eigenproblem.

3.1 Canonical Correlation Analysis. Observe that the solution of equation 2.2 is not affected by rescaling $\mathbf{w}_x$ or $\mathbf{w}_y$ either together or independently, so that, for example, replacing $\mathbf{w}_x$ by $\alpha \mathbf{w}_x$ gives the quotient

$$\frac{\alpha \mathbf{w}'_x C_{\mathbf{xy}} \mathbf{w}_y}{\sqrt{\alpha^2 \mathbf{w}'_x C_{\mathbf{xx}} \mathbf{w}_x \mathbf{w}'_y C_{\mathbf{yy}} \mathbf{w}_y}} = \frac{\mathbf{w}'_x C_{\mathbf{xy}} \mathbf{w}_y}{\sqrt{\mathbf{w}'_x C_{\mathbf{xx}} \mathbf{w}_x \mathbf{w}'_y C_{\mathbf{yy}} \mathbf{w}_y}}.$$

Since the choice of rescaling is therefore arbitrary, the CCA optimization problem formulated in equation 2.2 is equivalent to maximizing the numerator subject to

$$\mathbf{w}'_x C_{\mathbf{xx}} \mathbf{w}_x = 1$$
$$\mathbf{w}'_y C_{\mathbf{yy}} \mathbf{w}_y = 1.$$

The corresponding Lagrangian is

$$L(\lambda, \mathbf{w}_x, \mathbf{w}_y) = \mathbf{w}'_x C_{\mathbf{xy}} \mathbf{w}_y - \frac{\lambda_x}{2}(\mathbf{w}'_x C_{\mathbf{xx}} \mathbf{w}_x - 1) - \frac{\lambda_y}{2}(\mathbf{w}'_y C_{\mathbf{yy}} \mathbf{w}_y - 1).$$

Taking derivatives in respect to $\mathbf{w}_x$ and $\mathbf{w}_y$, we obtain

$$\frac{\partial L}{\partial \mathbf{w}_x} = C_{\mathbf{xy}} \mathbf{w}_y - \lambda_x C_{\mathbf{xx}} \mathbf{w}_x = \mathbf{0} \tag{3.1}$$
$$\frac{\partial L}{\partial \mathbf{w}_y} = C_{\mathbf{yx}} \mathbf{w}_x - \lambda_y C_{\mathbf{yy}} \mathbf{w}_y = \mathbf{0}. \tag{3.2}$$

Subtracting $\mathbf{w}'_y$ times the second equation from $\mathbf{w}'_x$ times the first, we have

$$\begin{aligned} 0 &= \mathbf{w}'_x C_{\mathbf{xy}} \mathbf{w}_y - \mathbf{w}'_x \lambda_x C_{\mathbf{xx}} \mathbf{w}_x - \mathbf{w}'_y C_{\mathbf{yx}} \mathbf{w}_x + \mathbf{w}'_y \lambda_y C_{\mathbf{yy}} \mathbf{w}_y \\ &= \lambda_y \mathbf{w}'_y C_{\mathbf{yy}} \mathbf{w}_y - \lambda_x \mathbf{w}'_x C_{\mathbf{xx}} \mathbf{w}_x, \end{aligned}$$

which together with the constraints implies that $\lambda_y - \lambda_x = 0$, let $\lambda = \lambda_x = \lambda_y$. Assuming $C_{\mathbf{yy}}$ is invertible, we have

$$\mathbf{w}_y = \frac{C_{\mathbf{yy}}^{-1} C_{\mathbf{yx}} \mathbf{w}_x}{\lambda}, \tag{3.3}$$

and so substituting in equation 3.1 gives

$$C_{\mathbf{xy}} C_{\mathbf{yy}}^{-1} C_{\mathbf{yx}} \mathbf{w}_x = \lambda^2 C_{\mathbf{xx}} \mathbf{w}_x. \tag{3.4}$$

We are left with a generalized eigenproblem of the form $A\mathbf{x} = \lambda B\mathbf{x}$. We can therefore find the coordinate system that optimizes the correlation between corresponding coordinates by first solving for the generalized eigenvectors of equation 3.4 to obtain the sequence of $\mathbf{w}_x$'s and then using equation 3.3 to find the corresponding $\mathbf{w}_y$'s.

If $C_{\mathbf{xx}}$ is invertible, we are able to formulate equation 3.4 as a standard eigenproblem of the form $B^{-1}A\mathbf{x} = \lambda\mathbf{x}$, although to ensure a symmetric standard eigenproblem, we do the following. As the covariance matrices $C_{\mathbf{xx}}$ and $C_{\mathbf{yy}}$ are symmetric positive definite, we are able to decompose them using a complete Cholesky decomposition,

$$C_{\mathbf{xx}} = R_{\mathbf{xx}} \cdot R'_{\mathbf{xx}},$$

where $R_{\mathbf{xx}}$ is a lower triangular matrix. If we let $\mathbf{u}_x = R'_{\mathbf{xx}} \cdot \mathbf{w}_x$, we are able to rewrite equation 3.4 as follows:

$$\begin{aligned} C_{\mathbf{xy}} C_{\mathbf{yy}}^{-1} C_{\mathbf{yx}} R_{\mathbf{xx}}^{-1'} \mathbf{u}_x &= \lambda^2 R_{\mathbf{xx}} \mathbf{u}_x \\ R_{\mathbf{xx}}^{-1} C_{\mathbf{xy}} C_{\mathbf{yy}}^{-1} C_{\mathbf{yx}} R_{\mathbf{xx}}^{-1'} \mathbf{u}_x &= \lambda^2 \mathbf{u}_x. \end{aligned}$$

We are therefore left with a symmetric standard eigenproblem of the form $A\mathbf{x} = \lambda\mathbf{x}$.

3.2 Generalization of the Canonical Correlation. In this section, we show a possible framework of the generalization of the canonical correlation for more than two samples. We show the relationship between the canonical correlation and the minimization of the total distance. (For details, see Gifi, 1990, and Ketterling, 1971.)

3.2.1 The Simultaneous Formulation of the Canonical Correlation. Instead of using the successive formulation of the canonical correlation, we can join the subproblems into one. We use subscripts $1, 2, \ldots$ instead of x, y to denote more than two classes. The matrices $W^{(1)}$ and $W^{(2)}$ contain the vectors $(\mathbf{w}_1^{(1)}, \ldots, \mathbf{w}_p^{(1)})$ and $(\mathbf{w}_p^{(2)}, \ldots, \mathbf{w}_p^{(2)})$. The simultaneous formulation

is the optimization problem assuming p iteration steps:

$$\max_{(\mathbf{w}_1^{(1)},\ldots,\mathbf{w}_p^{(1)}),\ldots,(\mathbf{w}_1^{(2)},\ldots,\mathbf{w}_p^{(2)})} \quad \sum_{i=1}^{p} \mathbf{w}_i^{(1)T} C_{12} \mathbf{w}_i^{(2)} \tag{3.5}$$

$$\text{s.t. } \mathbf{w}_i^{(1)T} C_{11} \mathbf{w}_j^{(1)} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases}$$

$$\mathbf{w}_i^{(2)T} C_{22} \mathbf{w}_j^{(2)} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise,} \end{cases}$$

$$i, j = 1, \ldots, p,$$

$$\mathbf{w}_i^{(1)T} C_{12} \mathbf{w}_j^{(2)} = 0,$$

$$i, j = 1, \ldots, p, \ j \neq i.$$

Based on equation 3.5, we can give a compact form to this optimization,

$$\max_{W^{(1)}, W^{(2)}} \quad Tr\left(W^{(1)T} C_{12} W^{(2)}\right) \tag{3.6}$$

$$\text{s.t.} \quad W^{(k)T} C_{kk} W^{(k)} = I,$$

$$\mathbf{w}_i^{(k)T} C_{kl} \mathbf{w}_j^{(l)} = 0,$$

$$k, l = \{1, 2\}, \ l \neq k, \ i, j = 1, \ldots, p, \ j \neq i.$$

where I is an identity matrix with size $p \times p$.

The canonical correlation problem can be transformed into a distance minimization problem where the distance between two matrices is measured by the Frobenius norm:

$$\min_{W^{(1)}, W^{(2)}} \quad \left\| S^{(1)} W^{(1)} - S^{(2)} W^{(2)} \right\|_F \tag{3.7}$$

$$\text{s.t.} \quad W^{(k)T} C_{kk} W^{(k)} = I,$$

$$\mathbf{w}_i^{(k)T} C_{kl} \mathbf{w}_j^{(l)} = 0,$$

$$k, l = 1, \ldots, 2, \ l \neq k, \ i, j = 1, \ldots, p, \ j \neq i.$$

Unfolding the objective function of equation 3.7 shows this optimization problem is the same as equation 3.6.

Exploiting the distance problem, we can give a generalization of the canonical correlation for more than two known samples. Let us give a set of samples in matrix form $\{S^{(1)}, \ldots, S^{(K)}\}$ with dimension $m \times n_1, \ldots, m \times n_K$. We are looking for the linear combinations of the columns of these matrices in the matrix form $W^{(1)}, \ldots, W^{(K)}$ such that they give the optimum solution

of the problem:

$$\min_{W^{(1)},\ldots W^{(K)}} \quad \sum_{k,l=1,k\neq l}^{K} \left\| S^{(k)}W^{(k)} - S^{(l)}W^{(l)} \right\|_F \tag{3.8}$$

$$\text{s.t.} \quad W^{(k)T}C_{kk}W^{(k)} = I,$$

$$\mathbf{w}_i^{(k)T}C_{kl}\mathbf{w}_j^{(l)} = 0,$$

$$k, l = 1, \ldots, K,\ l \neq k,\ i, j = 1, \ldots, p,\ j \neq i.$$

Unfolding the objective function, we have the sum of the squared Euclidean distances between all of the pairs of the column vectors of the matrices $S^{(k)}W^{(k)},\ k = 1, \ldots, K$. One can show this problem can be solved by using singular value decomposition for arbitrary K.

3.3 Kernel Canonical Correlation Analysis. CCA may not extract useful descriptors of the data because of its linearity. Kernel CCA offers an alternative solution by first projecting the data into a higher-dimensional feature space (Cristianini & Shawe-Taylor, 2000):

$$\phi\colon \mathbf{x} = (x_1, \ldots x_m) \mapsto \phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \ldots, \phi_N(\mathbf{x}))\ (m < N),$$

before performing CCA in the new feature space, essentially moving from the primal to the dual representation approach. Kernels are methods of implicitly mapping data into a higher-dimensional feature space, a method known as the kernel trick. A kernel is a function K, such that for all $\mathbf{x}, \mathbf{z} \in X$,

$$K(\mathbf{x}, \mathbf{z}) = \langle \phi(\mathbf{x}) \cdot \phi(\mathbf{z}) \rangle, \tag{3.9}$$

where ϕ is a mapping from X to a feature space F. Kernels offer a great deal of flexibility, as they can be generated from other kernels. In the kernel, the data appears only through entries in the Gram matrix. Therefore, this approach gives a further advantage as the number of tuneable parameters and updating time does not depend on the number of attributes being used.

Using the definition of the covariance matrix in equation 2.1, we can rewrite the covariance matrix C using the data matrices (of vectors) X and Y, which have the sample vector as rows and are therefore of size $m \times N$; we obtain

$$C_{\mathbf{xx}} = X'X$$
$$C_{\mathbf{xy}} = X'Y.$$

The directions $\mathbf{w}_x$ and $\mathbf{w}_y$ (of length N) can be rewritten as the projection of the data onto the direction α and β (of length m):

$$\mathbf{w}_x = X'\alpha$$
$$\mathbf{w}_y = Y'\beta.$$

Substituting into equation 2.2, we obtain the following:

$$\rho = \max_{\alpha,\beta} \frac{\alpha' XX' YY' \beta}{\sqrt{\alpha' XX' XX' \alpha \cdot \beta' YY' YY' \beta}}. \tag{3.10}$$

Let $K_x = XX'$ and $K_y = YY'$ be the kernel matrices corresponding to the two representation. We substitute into equation 3.10,

$$\rho = \max_{\alpha,\beta} \frac{\alpha' K_x K_y \beta}{\sqrt{\alpha' K_x^2 \alpha \cdot \beta' K_y^2 \beta}}. \tag{3.11}$$

We find that in equation 3.11, the variables are now represented in the dual form.

Observe that as with the primal form presented in equation 2.2, equation 3.11 is not affected by rescaling of α and β either together or independently. Hence, the KCCA optimization problem formulated in equation 3.11 is equivalent to maximizing the numerator subject to

$$\alpha' K_x^2 \alpha = 1$$
$$\beta' K_y^2 \beta = 1.$$

The corresponding Lagrangian is

$$L(\lambda, \alpha, \beta) = \alpha' K_x K_y \beta - \frac{\lambda_\alpha}{2}(\alpha' K_x^2 \alpha - 1) - \frac{\lambda_\beta}{2}(\beta' K_y^2 \beta - 1).$$

Taking derivatives in respect to α and β, we obtain

$$\frac{\partial L}{\partial \alpha} = K_x K_y \beta - \lambda_\alpha K_x^2 \alpha = \mathbf{0} \tag{3.12}$$

$$\frac{\partial L}{\partial \beta} = K_y K_x \alpha - \lambda_\beta K_y^2 \beta = \mathbf{0}. \tag{3.13}$$

Subtracting β' times equation 3.13 from α' times equation 3.12, we have

$$\begin{aligned} 0 &= \alpha' K_x K_y \beta - \alpha' \lambda_\alpha K_x^2 \alpha - \beta' K_y K_x \alpha + \beta' \lambda_\beta K_y^2 \beta \\ &= \lambda_\beta \beta' K_y^2 \beta - \lambda_\alpha \alpha' K_x^2 \alpha, \end{aligned}$$

which together with the constraints implies that $\lambda_\alpha - \lambda_\beta = 0$; let $\lambda = \lambda_\alpha = \lambda_\beta$. Considering the case where the kernel matrices K_x and K_y are invertible, we have

$$\begin{aligned} \beta &= \frac{K_y^{-1} K_y^{-1} K_y K_x \alpha}{\lambda} \\ &= \frac{K_y^{-1} K_x \alpha}{\lambda}. \end{aligned}$$

Substituting in equation 3.12, we obtain

$$K_x K_y K_y^{-1} K_x \alpha - \lambda^2 K_x K_x \alpha = 0.$$

Hence,

$$K_x K_x \alpha - \lambda^2 K_x K_x \alpha = 0$$

or

$$I\alpha = \lambda^2 \alpha. \tag{3.14}$$

We are left with a standard eigenproblem of the form $A\mathbf{x} = \lambda\mathbf{x}$. We can deduce from equation 3.14 that $\lambda = 1$ for every vector of α; hence, we can choose the projections α to be unit vectors j_i $i = 1, \ldots, m$ while β are the columns of $\frac{1}{\lambda} K_y^{-1} K_x$. Hence, when K_x or K_y is invertible, perfect correlation can be formed. Since kernel methods provide high-dimensional representations, such independence is not uncommon. It is therefore clear that a naive application of CCA in kernel-defined feature space will not provide useful results. This highlights the potential problem of overfitting that arises in high-dimensional feature spaces. Clearly these correlations are failing to distinguish spurious features from those capturing the underlying semantics. In the next section, we investigate how this problem can be avoided.

4 Computational Issues

We observe from equation 3.14 that if K_x is invertible, maximal correlation is obtained, suggesting learning is trivial. To force nontrivial learning, we introduce a control on the flexibility of the projections by penalizing the norms of the associated weight vectors by a convex combination of constraints based on partial least squares. Another computational issue that can arise is the use of large training sets, as this can lead to computational problems and degeneracy. To overcome this issue, we apply partial Gram-Schmidt orthogonalization (equivalently incomplete Cholesky decomposition) to reduce the dimensionality of the kernel matrices.

4.1 Regularization. To force nontrivial learning on the correlation by controlling the problem of overfitting and hence finding nonrelevant correlations, we introduce a control on the flexibility of the projection mappings using partial least squares (PLS) to penalize the norms of the associated weights. We convexly combine the PLS term with the KCCA term in the

denominator of equation 3.11, obtaining

$$\rho = \max_{\alpha,\beta} \frac{\alpha' K_x K_y \beta}{\sqrt{(\alpha' K_x^2 \alpha + \kappa \|\mathbf{w}_x\|^2) \cdot (\beta' K_y^2 \beta + \kappa \|\mathbf{w}_y\|^2))}}$$
$$= \max_{\alpha,\beta} \frac{\alpha' K_x K_y \beta}{\sqrt{(\alpha' K_x^2 \alpha + \kappa \alpha' K_x \alpha) \cdot (\beta' K_y^2 \beta + \kappa \beta' K_y \beta)}}.$$

We observe that the new regularized equation is not affected by rescaling of α or β; hence, the optimization problem is subject to

$$(\alpha' K_x^2 \alpha + \kappa \alpha' K_x \alpha) = 1$$
$$(\beta' K_y^2 \beta + \kappa \beta' K_y \beta) = 1.$$

We follow the same approach as in section 3.3, where we assume that K is invertible. We are able to find that

$$K_x K_y (K_y + \kappa I)^{-1} K_x \alpha = \lambda^2 K_x (K_x + \kappa I)\alpha$$
$$K_y (K_y + \kappa I)^{-1} K_x \alpha = \lambda^2 (K_x + \kappa I)\alpha$$
$$(K_x + \kappa I)^{-1} K_y (K_y + \kappa I)^{-1} K_x \alpha = \lambda^2 \alpha.$$

We obtain a standard eigenproblem of the form $A\mathbf{x} = \lambda\mathbf{x}$.

4.2 Incomplete Cholesky Decomposition. Complete decomposition (Golub & Loan, 1983) of a kernel matrix is an expensive step and should be avoided with real-world data. Incomplete Cholesky decomposition as described in Bach and Jordan (2002) differs from Cholesky decomposition in that all pivots below a certain threshold are skipped. If M is the number of nonskipped pivots, then we obtain a lower triangular matrix G^i with only M nonzero columns. Symmetric permutations of rows and columns are necessary during the factorization if we require the rank to be as small as possible (Golub & Loan, 1983).

In algorithm 1, we describe the algorithm from Bach and Jordan (2002) (with slight modification). The algorithm involves picking one column of K at a time, choosing the column to be added by greedily maximizing a lower bound on the reduction in the error of the approximation. After l steps, we have an approximation of the form $\tilde{K}_l = G_l^i G'^i_l$, where G_l^i is $N \times l$. The ranking of the $N - l$ vectors can be computed by comparing the diagonal elements of the remainder matrix $K - G_l^i G'^i_l$.

4.3 Partial Gram-Schmidt Orthogonalization. We explore the partial Gram-Schmidt orthogonalization (PGSO) algorithm, described in Cristianini et al. (2001), as our matrix decomposition approach. ICD could been

Algorithm 1: Pseudocode for ICD

Input matrix K of size $N \times N$, precision parameter η

Initialization: $i = 1$, $K' = K$, $P = I$, for $j \in [1, N]$, $G_{jj} = K_{jj}$
while $\sum_{j=1}^{N} G_{jj} > \eta$ and $i! = N + 1$
Find best new element: $j^* = \arg\max_{j \in [i,N]} G_{jj}$
Update $j^* = (j^* + i) - 1$
Update permutation P :
$Pnext = I$, $Pnext_{ii} = 0$, $Pnext_{j^*j^*} = 0$, $Pnext_{ij^*} = 1$, $Pnext_{j^*i} = 1$
$P = P \cdot Pnext$
Permute elements i and j^* in K':
$K' = Pnext \cdot K' \cdot Pnext$
Update (due to new permutation) the already calculated elements of G: $G_{i,1:i-1} \leftrightarrow G_{j^*,1:i-1}$
Permute elements j^*, j^* and i, i of G:
$G(i, i) \leftrightarrow G(j^*, j^*)$
Set $G_{ii} = \sqrt{G_{ii}}$
Calculate ith column of G:
$G_{i+1:n,i} = \frac{1}{G_{ii}}(K'_{i+1:n,i} - \sum_{j=1}^{i-1} G_{i+1:n,j} G_{ij})$
Update only diagonal elements: for $j \in [i + 1, N]$, $G_{jj} = K'_{jj} - \sum_{k=1}^{i} G_{jk}^2$
Update $i = i + 1$
end while
Output P, G and $M = i$

Output: $N \times M$ lower triangular matrix G, a permutation matrix P such that $\|P'KP - GG'\| \leq \eta$.

as equivalent to PGSO as ICD is the dual implementation of PGSO. PGSO works as follows. The projection is built up as the span of a subset of the projections of a set of m training examples. These are selected by performing a Gram-Schmidt orthogonalization of the training vectors in the feature space. We slightly modify the Gram-Schmidt algorithm so it will use a precision parameter as a stopping criterion, as shown in Bach and Jordan (2002). The PGSO pseudocode is as shown in algorithm 2. The pseudocode to classify a new example is given in algorithm 3. The advantage of using the PGSO in comparison to the incomplete Cholesky decomposition (as described in section 4.2) is that there is no need for a permutation matrix P.

4.4 Kernel-CCA with PGSO. So far we have considered the kernel matrices as invertible, although in practice this may not be the case. In this section, we address the issue of using large training sets, which may lead to computational problems and degeneracy. We use PGSO to approximate the kernel matrices such that we are able to re-represent the correlation with reduced dimensionality.

Algorithm 2: Pseudocode for PGSO

Given kernel K from a training set, precision parameter η

$m =$ size of K, a $N \times N$ matrix
$j = 1$
size and *index* are a vector with the same length as K
$feat$ a zeros matrix equal to the size of K
for $i = 1$ to m **do**
  norm2$[i] = K_{ii}$;
end for
while $\sum_i norm2[i] > \eta$ and $j! = N + 1$ **do**
  $i_j = \arg\max_i(\text{norm2}[i])$;
  index$[j] = i_j$;
  size$[j] = \sqrt{norm2[i_j]}$;
  for $i = 1$ to m **do**
    feat$[i, j] = \frac{(k(d_i, d_{i_j}) - \sum_{t=1}^{j-1} feat[i,t] \cdot feat[i_j,t])}{size[j]}$;
    norm2$[i]$ = norm2$[i]$− feat$(i, j)\cdot$ feat(i, j);
  end for
  $j = j + 1$
end while
$M = j$
return feat

Output:
$\|K - feat \cdot feat'\| \leq \eta$ where $feat$ is a $N \times M$ lower triangular matrix

Algorithm 3: Pseudocode to Classify a New Example at Location i

Given a kernel K from a testing set
  for $j = 1$ to M **do**
    newfeat$[j] = (K_{i,index[j]} - \sum_{t=1}^{j-1}$ newfeat$[t] \cdot$ feat$[index[j], t])$/size$[j]$;
  end for

Decomposing the kernel matrices K_x and K_y via PGSO, where R is a lower triangular matrix, gives

$$K_x \doteq R_x R_x'$$
$$K_y \doteq R_y R_y'.$$

Substituting the new representation into equations 3.12 and 3.13 and multiplying the first equation with R_x' and the second equation with R_y' gives

$$R_x' R_x R_x' R_y R_y' \beta - \lambda R_x' R_x R_x' R_x R_x' \alpha = 0 \tag{4.1}$$
$$R_y' R_y R_y' R_x R_x' \alpha - \lambda R_y' R_y R_y' R_y R_y' \beta = 0. \tag{4.2}$$

Let Z be the new correlation matrix with the reduced dimensionality:

$$\begin{aligned} R_x' R_x &= Z_{xx} \\ R_y' R_y &= Z_{yy} \\ R_x' R_y &= Z_{xy} \\ R_y' R_x &= Z_{yx}. \end{aligned}$$

Let $\tilde{\alpha}$ and $\tilde{\beta}$ be the reduced directions, such that

$$\begin{aligned} \tilde{\alpha} &= R_x' \alpha \\ \tilde{\beta} &= R_y' \beta. \end{aligned}$$

Substituting in equations 4.1 and 4.2, we find that we return to the primal representation of CCA with a dual representation of the data:

$$\begin{aligned} Z_{xx} Z_{xy} \tilde{\beta} - \lambda Z_{xx}^2 \tilde{\alpha} &= 0 \\ Z_{yy} Z_{yx} \tilde{\alpha} - \lambda Z_{yy}^2 \tilde{\beta} &= 0. \end{aligned}$$

Assume that the Z_{xx} and Z_{yy} are invertible. We multiply the first equation with Z_{xx}^{-1} and the second with Z_{yy}^{-1}:

$$Z_{xy} \tilde{\beta} - \lambda Z_{xx} \tilde{\alpha} = 0 \tag{4.3}$$

$$Z_{yx} \tilde{\alpha} - \lambda Z_{yy} \tilde{\beta} = 0. \tag{4.4}$$

We are able to rewrite $\tilde{\beta}$ from equation 4.4 as

$$\tilde{\beta} = \frac{Z_{yy}^{-1} Z_{yx} \tilde{\alpha}}{\lambda},$$

and substituting in equation 4.3 gives

$$Z_{xy} Z_{yy}^{-1} Z_{yx} \tilde{\alpha} = \lambda^2 Z_{xx} \tilde{\alpha}. \tag{4.5}$$

We are left with a generalized eigenproblem of the form $A\mathbf{x} = \lambda B\mathbf{x}$. Let SS' be equal to the complete Cholesky decomposition of Z_{xx} such that $Z_{xx} = SS'$, where S is a lower triangular matrix, and let $\hat{\alpha} = S' \cdot \tilde{\alpha}$. Substituting in equation 4.5, we obtain

$$S^{-1} Z_{xy} Z_{yy}^{-1} Z_{yx} S^{-1'} \hat{\alpha} = \lambda^2 \hat{\alpha}.$$

We now have a symmetric standard eigenproblem of the form $A\mathbf{x} = \lambda \mathbf{x}$.

We combine the dimensionality reduction with the regularization parameter (see section 4.1). Following the same approach, it is easy to show that we are left with a generalized eigenproblem of the form $A\mathbf{x} = \lambda B\mathbf{x}$:

$$Z_{xy}(Z_{yy} + \kappa I)^{-1}Z_{yx}S^{-1'}\alpha = \lambda^2(Z_{xx} + \kappa I)\alpha.$$

Performing a complete Cholesky decomposition on $Z_{xx} + \kappa I = SS'$ where S is a lower triangular matrix, let $\hat{\alpha} = S'\tilde{\alpha}$:

$$S^{-1}Z_{xy}(Z_{yy} + \kappa I)^{-1}Z_{yx}S^{-1'}\hat{\alpha} = \lambda^2\hat{\alpha}.$$

We obtain a symmetric standard eigenproblem of the form $A\mathbf{x} = \lambda\mathbf{x}$.

5 Experimental Results

In the following experiments, the problem of learning semantics of multimedia content by combining image and text data is addressed. The synthesis is addressed by the kernel canonical correlation analysis described in section 4.4. We test the use of the derived semantic space in an image retrieval task that uses only image content. The aim is to allow retrieval of images from a text query but without reference to any labeling associated with the image. This can be viewed as a cross-modal retrieval task. We used the combined multimedia image-text web database, which was kindly provided by Kolenda, Hansen, Larsen, and Winther (2002), where we are trying to facilitate mate retrieval on a test set. The data were divided into three classes: Sport, Aviation, and Paintball (see Figure 1). There were 400 records each, consisting of jpeg images retrieved from the Internet with attached text. We randomly split each class into two halves, used as training and test data, accordingly. The extracted features of the data used were the same as in Kolenda et al. (2002). (A detailed description of the features used can be found in Kolenda et al.) They were image HSV color, image Gabor texture, and term frequencies in text. The first view (K_x) comprises a gaussian kernel (with σ set to be the minimum distance between images) on the image HSV color add with a linear kernel on the Gabor texture. The second view (K_y) is a linear kernel on the term frequencies.

We compute the value of κ for the regularization by running the KCCA with the association between image and text randomized. Let $\lambda(\kappa)$ be the spectrum without randomization, the database with itself, and $\lambda_R(\kappa)$ be the spectrum with randomization, the database with a randomized version of itself (by spectrum, we mean the vector whose entries are the eigenvalues). We would like to have the nonrandom spectrum as distant as possible from the randomized spectrum, as if the same correlation occurs for $\lambda(\kappa)$ and $\lambda_R(\kappa)$. Then overfitting clearly is taking place. Therefore, we expect that for $\kappa = 0$ (no regularization) and $\mathbf{j} = 1, \ldots, 1$ (the all ones vector) that we may have $\lambda(\kappa) = \lambda_R(\kappa) = \mathbf{j}$, since it is possible that the examples are linearly

Figure 1: Example of images in database.

independent. Though we find that only 50% of the examples are linearly independent, this does not affect the selection of κ through this method. We choose the value of κ so that the difference between the spectrum of the randomized set is maximally different (in the two norm) from the true spectrum:

$$\kappa = \arg\max \|\lambda_R(\kappa) - \lambda(\kappa)\|$$

We find that $\kappa = 7$ and set via a heuristic technique the Gram-Schmidt precision parameter $\eta = 0.5$.

To perform the test image retrieval, we compute the features of the images and text query using the Gram-Schmidt algorithm. Once we have obtained the features for the test query (text) and test images, we project them into the semantic feature space using $\tilde{\beta}$ and $\tilde{\alpha}$ (which are computed through training), respectively. Now we can compare them using an inner product of the semantic feature vector. The higher the value of the inner product, the more similar the two objects are. Hence, we retrieve the images whose inner products with the test query are highest.

We compared the performance of our methods with a retrieval technique based on the generalized vector space model (GVSM). This uses as a semantic feature vector the vector of inner products between either a text query and each training label or test image and each training image. For both methods, we used a gaussian kernel, with $\sigma = \max$ distance/20 for the image color component, and all experiments were an average of 10 runs. For convenience, we review the content-based and mate-based approaches in separate sections. Whereas in the content-based retrieval, we are interested in retrieving images of the same class (label), in mate based, we are inter-

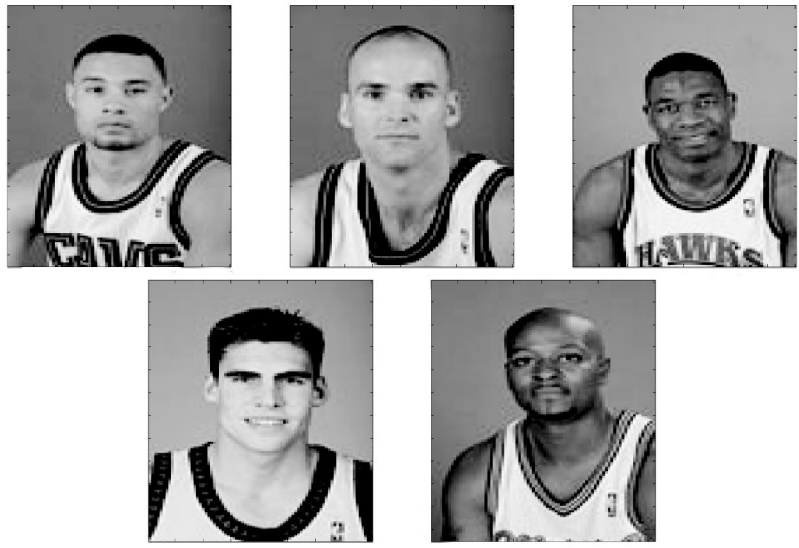

Figure 2: Images retrieved for the text query: "height: 6-11 weight: 235 lbs position: forward born: september 18, 1968, split, croatia college: none."

ested in the retrieval of the exact matching image. Although one could argue the relevance of the content-based retrieval experiment due to the lack of labels, we feel it is important in showing the duality and possibility of the application.

5.1 Content-Based Retrieval. In this experiment, we used the first 30 and 5 $\tilde{\alpha}$ eigenvectors and $\tilde{\beta}$ eigenvectors (corresponding to the largest eigenvalues). We computed the 10 and 30 images for which their semantic feature vector has the closest inner product with the semantic feature vector of the chosen text. Success is considered if the images contained in the set are of the same label as the query text (see Figure 2, the retrieval example for set of five images).

In Tables 1 and 2, we compare the performance of the kernel CCA algorithm and generalized vector space model. In Table 1, we present the performance of the methods over 10 and 30 image sets, and in Table 2, as plotted in Figure 3, we see the overall performance of the KCCA method against the GVSM for image sets (1–200). In the 200′th image set location, the maximum of 200×600 of the same labeled images over all text queries can be retrieved (we have only 200 images per label). The success rate in

Table 1: Success Cross-Results Between Kernel-CCA and Generalized Vector Space.

Image Set	GVSM Success	KCCA Success (30)	KCCA Success (5)
10	78.93%	85%	90.97%
30	76.82%	83.02%	90.69%

Table 2: Success Rate Over All Image Sets (1–200).

Method	Overall Success
GVSM	72.3%
KCCA (30)	79.12%
KCCA (5)	88.25%

Table 1 and Figure 3 is computed as follows:

$$\text{success \% for image set } i = \frac{\sum_{j=1}^{600} \sum_{k=1}^{i} count_k^j}{i \times 600} \times 100,$$

where $count_k^j = 1$ if the image k in the set is of the same label as the text query present in the set, else $count_k^j = 0$. The success rate in Table 2 is computed as above and averaged over all image sets.

As visible in Figure 4, we observe that when we add eigenvectors to the semantic projection, we will reduce the success of the content-based retrieval. We speculate that this may be the result of unnecessary detail in the semantic eigenvectors and that the semantic information needed is contained in the first few eigenvectors. Hence a minimal selection of 5 eigenvectors is sufficient to obtain a high success rate.

5.2 Mate-Based Retrieval. In the experiment, we used the first 150 and 30 $\tilde{\alpha}$ eigenvectors and $\tilde{\beta}$ eigenvectors (corresponding to the largest eigenvalues). We computed the 10 and 30 images for which their semantic feature vector has the closest inner product with the semantic feature vector of the chosen text. A successful match is considered if the image that actually matched the chosen text is contained in this set. We computed the success as the average of 10 runs (see Figure 5, the retrieval example for set of 5 images).

In Table 3, we compare the performance of the KCCA algorithm with the GVSM over 10 and 30 image sets, whereas in Table 4, we present the overall success over all image sets. In Figure 6, we see the overall performance of the KCCA method against the GVSM for all possible image sets.

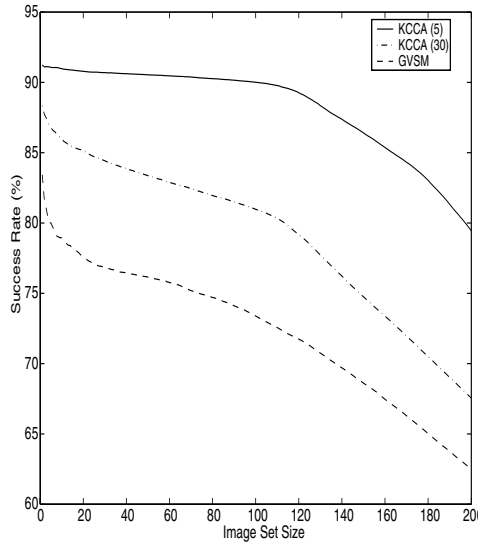

Figure 3: Success plot for content-based KCCA against GVSM.

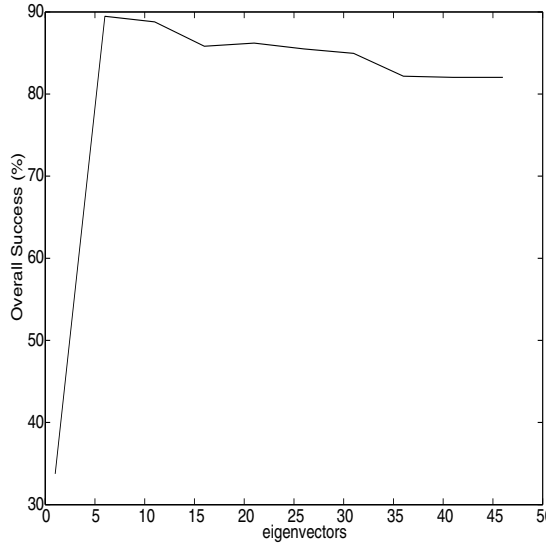

Figure 4: Content-based plot of eigenvector selection against overall success (%).

Figure 5: Images retrieved for the text query: "at phoenix sky harbor on july 6, 1997. 757-2s7, n907wa phoenix suns taxis past n902aw teamwork america west america west 757-2s7, n907wa phoenix suns taxis past n901aw arizona at phoenix sky harbor on july 6, 1997." The actual match is the middle picture in the top row.

Table 3: Success Cross-Results Between Kernel-CCA and Generalised Vector Space.

Image Set	GVSM Success	KCCA Success (30)	KCCA Success (150)
10	8%	17.19%	59.5%
30	19%	32.32%	69%

The success rate in Table 3 and Figure 6 is computed as follows:

$$\text{success \% for image set } i = \frac{\sum_{j=1}^{600} count_j}{600} \times 100,$$

where $count_j = 1$ if the exact matching image to the text query was present in the set, else $count_j = 0$. The success rate in Table 4 is computed as above and averaged over all image sets.

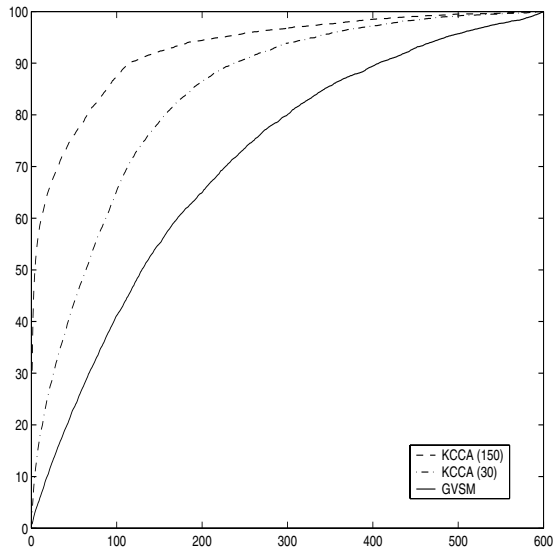

Figure 6: Success plot for KCCA mate-based against GVSM (success (%) against image set size).

Table 4: Success Rate Over All Image Sets.

Method	Overall Success
GVSM	70.6511%
KCCA (30)	83.4671%
KCCA (150)	92.9781%

As visible in Figure 7, we find that unlike the content-based retrieval, increasing the number of eigenvectors used will assist in locating the matching image to the query text. We speculate that this may be the result of added detail toward exact correlation in the semantic projection. We do not compute for all eigenvectors as this process would be expensive and the remaining eigenvectors would not necessarily add meaningful semantic information.

It is clear that the kernel CCA significantly outperforms the GVSM method in both content retrieval and mate retrieval.

5.3 Regularization Parameter. We next verify that the method of selecting the regularization parameter κ a priori gives a value that performs well. We randomly split each class into two halves, used as training and test data accordingly. We keep this divided set for all runs. We set the value of the incomplete Gram-Schmidt orthogonalization precision parameter $\eta = 0.5$

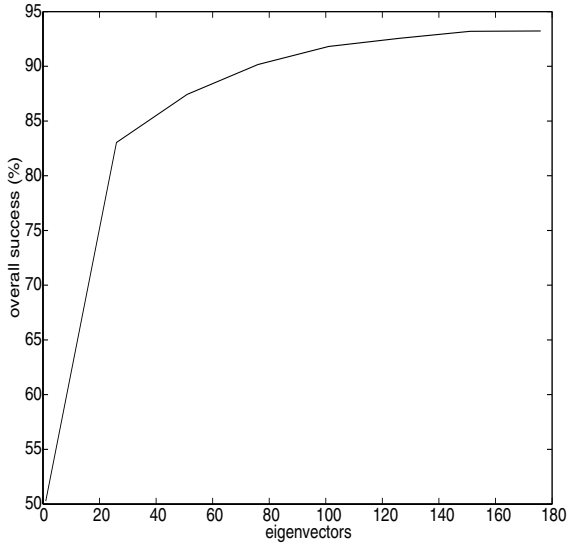

Figure 7: Mate-based plot of eigenvector selection against overall success (%).

Table 5: Overall Success of Content-Based (CB) KCCA with Respect to κ.

κ	CB-KCCA (30)	CB-KCCA (5)
0	46.278%	43.8374%
$\hat{\kappa}$	83.5238%	91.7513%
90	88.4592%	**92.7936**%
230	**88.5548**%	92.5281%

Note: Bold type signifies the best overall success achieved.

and run over possible values κ, where for each value, we test its content-based and mate-based retrieval performance.

Let $\hat{\kappa}$ be the previous optimal choice of the regularization parameter $\hat{\kappa} = \kappa = 7$. As we define the new optimal value of κ by its performance on the testing set, we can say that this method is biased (loosely, it is cheating). We will show that despite this, the difference between the performance of the biased κ and our a priori $\hat{\kappa}$ is slight.

In Table 5, we compare the overall performance of the content-based (CB) performance with respect to the different values of κ, and Figures 8 and 9 plot the comparison. We observe that the difference in performance between the a priori value $\hat{\kappa}$ and the newly found optimal value κ for 5 eigenvectors is 1.0423% and for 30 eigenvectors it is 5.031%. The more substantial increase in performance on the latter is due to the increase in the selection of the

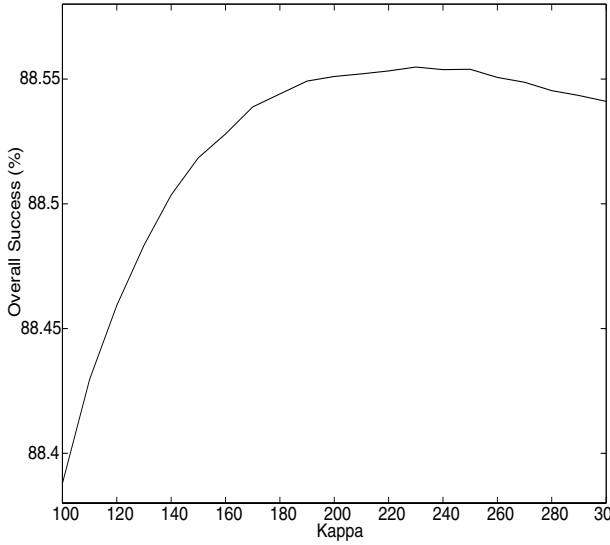

Figure 8: Content-based κ selection against overall success for 30 eigenvectors.

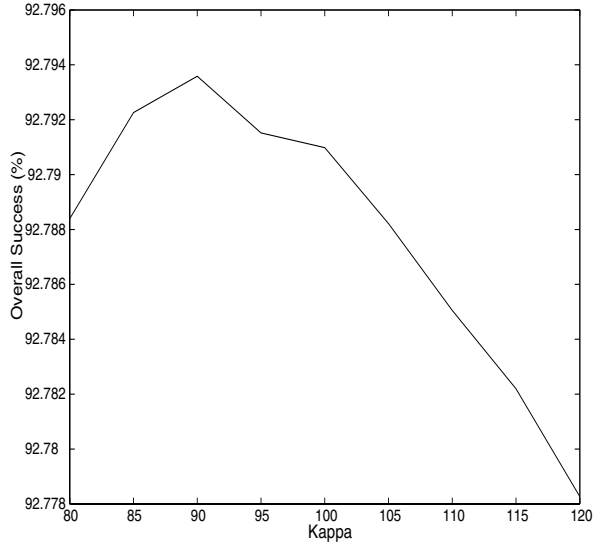

Figure 9: Content-based κ selection against overall success for five eigenvectors.

Table 6: Overall Success of Mate-Based (MB) KCCA with Respect to κ.

κ	MB-KCCA (30)	MB-KCCA (150)
0	73.4756%	83.46%
$\hat{\kappa}$	84.75%	92.4%
170	**85.5086**%	92.9975%
240	**85.5086**%	93.0083%
430	85.4914%	**93.027**%

Note: Bold type signifies the best overall success achieved.

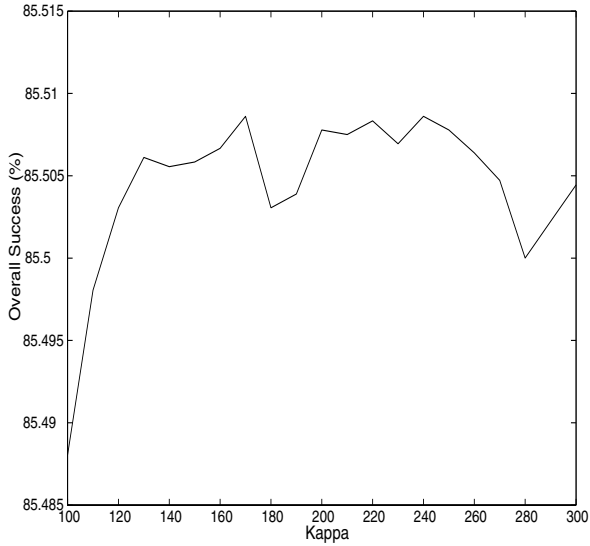

Figure 10: Mate-based κ selection against overall success for 30 eigenvectors.

regularization parameter, which compensates for the substantial decrease in performance (see Figure 6) of the content-based retrieval, when high-dimensional semantic feature space is used.

In Table 6, we compare the overall performance of the mate-based (MB) performance with respect to the different values of κ, and in Figures 10 and 11 we view a plot of the comparison. We observe that in this case, the difference in performance between the a priori value $\hat{\kappa}$ and the newly found optimal value κ is for 150 eigenvectors 0.627% and for 30 eigenvectors is 0.7586%.

Our observed results support our proposed method for selecting the regularization parameter κ in an a priori fashion, since the difference between the actual optimal κ and the a priori $\hat{\kappa}$ is very slight.

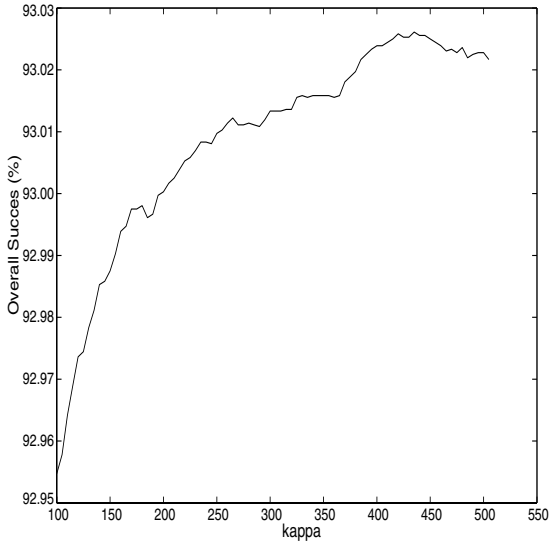

Figure 11: Mate-based κ selection against overall success for 150 eigenvectors.

6 Conclusions

In this study, we have presented a tutorial on canonical correlation analysis and have established a novel general approach to retrieving images based solely on their content. This is then applied to content-based and mate-based retrieval. Experiments show that image retrieval can be more accurate than with the generalized vector space model. We demonstrate that one can choose the regularization parameter κ a priori that performs well in very different regimes. Hence, we have come to the conclusion that kernel canonical correlation analysis is a powerful tool for image retrieval via content. In the future, we will extend our experiments to other data collections.

In the procedure of the generalization of the canonical correlation analysis, we can see that the original problem can be transformed and reinterpreted as a total distance problem or variance minimization problem. This special duality between the correlation and the distance requires more investigation to give more suitable descriptions of the structure of some special spaces generated by different kernels.

These approaches can provide tools to handle some problems in the kernel space where the inner products and the distances between the points are known but the coordinates are not. For some problems, it is sufficient to know only the coordinates of a few special points, which can be expressed from the known inner product (e.g., do cluster analysis in the kernel space and compute the coordinates of the cluster centers only).

Acknowledgments

We acknowledge the financial support of EU Projects KerMIT, No. IST-2000-25341 and LAVA, No. IST-2001-34405. The majority of work was done at Royal Holloway, University of London.

References

Akaho, S. (2001). A kernel method for canonical correlation analysis. In *International Meeting of Psychometric Society*. Osaka, Japan.

Bach, F., & Jordan, M. (2002). Kernel independent component analysis. *Journal of Machine Leaning Research*, *3*, 1–48.

Borga, M. (1999). *Canonical correlation*. Online tutorial. Available online at: http://www.imt.liu.se/∼magnus/cca/tutorial.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge: Cambridge University Press.

Cristianini, N., Shawe-Taylor, J., & Lodhi, H. (2001). Latent semantic kernels. In C. Brodley & A. Danyluk (Eds.), *Proceedings of ICML-01, 18th International Conference on Machine Learning* (pp. 66–73). San Francisco: Morgan Kaufmann.

Fyfe, C., & Lai, P. L. (2001). Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems, 10*, 365–374.

Gifi, A. (1990). *Nonlinear multivariate analysis*. New York: Wiley.

Golub, G. H., & Loan, C. F. V. (1983). *Matrix computations*. Baltimore: Johns Hopkins University Press.

Hardoon, D. R., & Shawe-Taylor, J. (2003). KCCA for different level precision in content-based image retrieval. In *Proceedings of Third International Workshop on Content-Based Multimedia Indexing*. Rennes, France: IRISA.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, *28*, 312–377.

Ketterling, J. R. (1971). Canonical analysis of several sets of variables. *Biometrika*, *58*, 433–451.

Kolenda, T., Hansen, L. K., Larsen, J., & Winther, O. (2002). Independent component analysis for understanding multimedia content. In H. Bourlard, T. Adali, S. Bengio, J. Larsen, & S. Douglas (Eds.), *Proceedings of IEEE Workshop on Neural Networks for Signal Processing XII* (pp. 757–766). Piscataway, NJ: IEEE Press.

Vinokourov, A., Hardoon, D. R., & Shawe-Taylor, J. (2003). Learning the semantics of multimedia content with application to web image retrieval and classification. In *Proceedings of Fourth International Symposium on Independent Component Analysis and Blind Source Separation*. Nara, Japan.

Vinokourov, A., Shawe-Taylor, J., & Cristianini, N. (2002). Inferring a semantic representation of text via cross-language correlation analysis. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems, 15*. Cambridge, MA: MIT Press.

Received June 24, 2003; accepted May 4, 2004.