Canonical correlation analysis; An overview with application to learning methods

David R. Hardoon , Sandor Szedmak and John Shawe-Taylor

Department of Computer Science
Royal Holloway, University of London
{davidh, sandor, john}@cs.rhul.ac.uk

Technical Report

CSD-TR-03-02

May 28, 2003

Department of Computer Science
Egham, Surrey TW20 0EX, England

Abstract

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 only on their content from a text query. We compare the approaches against a standard cross-representation retrieval technique known as the Generalised Vector Space Model.

Keywords: Canonical correlation analysis, kernel canonical correlation analysis, partial Gram-Schmidt orthogonolisation, Cholesky decomposition, incomplete Cholesky decomposition, kernel methods.

1 Introduction

During recent years there have been advances in data learning using kernel methods. Kernel representation offers an alternative learning to non-linear 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 open 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 only makes use of 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 co-ordinate system in any multivariate data. The directions of the axes of this co-ordinate system are determined by both the second and higher order statistics of the original data. The goal is to perform a linear transform which 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 only uses one view of an object and the label as the corresponding pair. PLS could be thought of as a method, which 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 towards the underling semantics. CCA makes use of two views of the same semantic object to extract the representation of the semantics. The main difference between CCA and the other three methods is that CCA is closely related to mutual information (Borga 1998 [3]). Hence CCA can be easily motivated in information based tasks and is our natural selection.

Proposed by H. Hotelling in 1936 [12], CCA 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 are mutually maximised. In an attempt to increase the flexibility of the feature selection, kernelisation 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 & Lai [8], Akaho [1] and the recently Vinokourov et al. [19] 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 this data unless it is organised 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 [11] to as well as mate based retrieval [18, 11]. In both cases we compare the KCCA approach to the Generalised 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 the following ways:

• In this study we follow the work of Borga [4] 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 & Jordan [2] to compute a new correlation matrix with reduced dimensionality. Though Bach & Jordan [2] address a very different problem, they use the same underlining technique of Cholesky decomposition to re-represent the kernel matrices. We show that by using partial Gram-Schmidt orthogonolisation [6] 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 only changing the selection of eigenvectors used in the semantic projection.

• 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 generalisation of the framework for canonical correlation analysis. Our approach is based on the works of Gifi (1990) and Ketterling (1971). The purpose of the generalisation is to extend the canonical correlation as an associativity measure between two set of variables to more than two sets, whilst preserving most of its properties. The generalisation starts with the optimisation problem formulation of canonical correlation. By changing the objective function we will arrive at the multi set problem. Applying similar constraint sets in the optimisation problems we find that the feasible solutions are singular vectors of matrices, which are derived the same way for the original and generalised 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. In Section 6 we present the generalisation framework for CCA while in Section 7 draws final conclusions.

2 Theoretical Foundations

Proposed by H. Hotelling in 1936 [12], 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 are mutually maximised. Correlation analysis is dependent on the co-ordinate 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 co-ordinate 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 are transformed the corresponding co-ordinates are maximally correlated.

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 $(\mathbf{x}, \mathbf{y})$, we use S_x to denote $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ and similarly S_y to denote $(\mathbf{y}_1, \ldots, \mathbf{y}_n)$. We can consider defining a new co-ordinate for $\mathbf{x}$ by choosing a direction $\mathbf{w}_x$ and projecting $\mathbf{x}$ onto that direction

$$\mathbf{x} \rightarrow \langle \mathbf{w}_x, \mathbf{x} \rangle$$

if we do the same for $\mathbf{y}$ by choosing a direction $\mathbf{w}_y$ we obtain a sample of the new $\mathbf{x}$ co-ordinate as

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

with the corresponding values of the new $\mathbf{y}$ co-ordinate being

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

The first stage of canonical correlation is to choose $\mathbf{w}_x$ and $\mathbf{w}_y$ to maximise the correlation between the two vectors. In other words the function to be maximised 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 $f(\mathbf{x}, \mathbf{y})$, were

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

we can rewrite the correlation expression as

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

follows that

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

Where we use A' to denote the transpose of a vector or matrix A.
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 $C_{\mathbf{x}\mathbf{x}}$ and $C_{\mathbf{y}\mathbf{y}}$ and the between-sets covariance matrices are $C_{\mathbf{x}\mathbf{y}} = C'_{\mathbf{y}\mathbf{x}}$

Hence, we can rewrite the function ρ as

$$\rho = \max_{\mathbf{w}_x,\mathbf{w}_y} \frac{\mathbf{w}_x' C_{\mathbf{x}\mathbf{y}}\mathbf{w}_y}{\sqrt{\mathbf{w}_x' C_{\mathbf{x}\mathbf{x}}\mathbf{w}_x\mathbf{w}_y' C_{\mathbf{y}\mathbf{y}}\mathbf{w}_y}} \tag{2.2}$$

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 will give an overview of the Canonical correlation analysis (CCA) and Kernel-CCA (KCCA) algorithms where we formulate the optimisation problem as a generalised eigenproblem.

3.1 Canonical Correlation Analysis

Observe that the solution of equation (2.2) is not affected by re-scaling $\mathbf{w}_x$ or $\mathbf{w}_y$ either together or independently, so that for example replacing $\mathbf{w}_x$ by $\alpha\mathbf{w}_x$ gives the quotient

$$\frac{\alpha\mathbf{w}_x' C_{\mathbf{x}\mathbf{y}}\mathbf{w}_y}{\sqrt{\alpha^2\mathbf{w}_x' C_{\mathbf{x}\mathbf{x}}\mathbf{w}_x\mathbf{w}_y' C_{\mathbf{y}\mathbf{y}}\mathbf{w}_y}} = \frac{\mathbf{w}_x' C_{\mathbf{x}\mathbf{y}}\mathbf{w}_y}{\sqrt{\mathbf{w}_x' C_{\mathbf{x}\mathbf{x}}\mathbf{w}_x\mathbf{w}_y' C_{\mathbf{y}\mathbf{y}}\mathbf{w}_y}}.$$

Since the choice of re-scaling is therefore arbitrary, the CCA optimisation problem formulated in equation (2.2) is equivalent to maximising the numerator

subject to

$$\begin{aligned} \mathbf{w}_x' C_{\mathbf{xx}} \mathbf{w}_x &= 1 \\ \mathbf{w}_y' C_{\mathbf{yy}} \mathbf{w}_y &= 1. \end{aligned}$$

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

$$\frac{\partial f}{\partial \mathbf{w}_x} = C_{\mathbf{xy}} \mathbf{w}_y - \lambda_x C_{\mathbf{xx}} \mathbf{w}_x = \mathbf{0} \tag{3.1}$$

$$\frac{\partial f}{\partial \mathbf{w}_y} = C_{\mathbf{yx}} \mathbf{w}_x - \lambda_y C_{\mathbf{yy}} \mathbf{w}_y = \mathbf{0}. \tag{3.2}$$

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

$$\frac{C_{\mathbf{xy}} C_{\mathbf{yy}}^{-1} C_{\mathbf{yx}} \mathbf{w}_x}{\lambda} - \lambda C_{\mathbf{xx}} \mathbf{w}_x = 0$$

or

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

We are left with a generalised eigenproblem of the form $A\mathbf{x} = \lambda B\mathbf{x}$. We can therefore find the co-ordinate system that optimises the correlation between corresponding co-ordinates by first solving for the generalised 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.

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 (more details on Cholesky decomposition can be found in section 4.2)

$$C_{\mathbf{xx}} = R_{\mathbf{xx}} \cdot R_{\mathbf{xx}}'$$

where R_{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 eigenproblem of the form $A\mathbf{x} = \lambda\mathbf{x}$.

3.2 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

$$\phi : \mathbf{x} = (\mathbf{x}_1, \dots \mathbf{x}_n) \mapsto \phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \dots, \phi_N(\mathbf{x})) \ (n < N)$$

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 $x, z \in X$

$$K(x, z) =< \phi(x) \cdot \phi(z) > \tag{3.5}$$

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 only appears 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

$$\begin{aligned} C_{\mathbf{xx}} &= X'X \\ C_{\mathbf{xy}} &= X'Y. \end{aligned}$$

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)

$$\begin{aligned} \mathbf{w}_x &= X'\alpha \\ \mathbf{w}_y &= Y'\beta. \end{aligned}$$

Substituting into equation (2.2) we obtain the following

$$\rho = \max_{\alpha,\beta} \frac{\alpha' XX'YY'\beta}{\sqrt{\alpha' XX'XX'\alpha \cdot \beta' YY'YY'\beta}} \tag{3.6}$$

Let $K_x = XX'$ and $K_y = YY'$ be the kernel matrices corresponding to the two representation. We substitute into equation (3.6)

$$\rho = \max_{\alpha,\beta} \frac{\alpha' K_x K_y \beta}{\sqrt{\alpha' K_x^2 \alpha \cdot \beta' K_y^2 \beta}}. \tag{3.7}$$

We find that in equation (3.7) the variables are now represented in the dual form.
Observe that as with the primal form presented in equation (2.2), equation (3.7) is not affected by re-scaling of α and β either together or independently. Hence

the KCCA optimisation problem formulated in equation (3.7) is equivalent to maximising the numerator subject to

$$\begin{aligned} \alpha' K_x^2 \alpha &= 1 \\ \beta' K_y^2 \beta &= 1 \end{aligned}$$

The corresponding Lagrangian is

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

Taking derivatives in respect to α and β we obtain

$$\frac{\partial f}{\partial \alpha} = K_x K_y \beta - \lambda_\alpha K_x^2 \alpha = \mathbf{0} \tag{3.8}$$

$$\frac{\partial f}{\partial \beta} = K_y K_x \alpha - \lambda_\beta K_y^2 \beta = \mathbf{0}. \tag{3.9}$$

Subtracting β' times the second equation from α' times the first 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 (3.8) 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

$$I\alpha = \lambda^2 \alpha. \tag{3.10}$$

We are left with a generalised eigenproblem of the form $A\mathbf{x} = \lambda\mathbf{x}$. We can deduce from equation 3.10 that $\lambda = 1$ for every vector of α; hence we can choose the projections $\mathbf{w}_x$ to be unit vectors j_i $i = 1, \ldots, m$ while $\mathbf{w}_y$ 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. In the next section we investigate how this problem can be avoided.

4 Computational Issues

We observe from equation (3.10) that if K_x is invertible maximal correlation is obtained, suggesting learning is trivial. To force non-trivial learning we introduce a control on the flexibility of the projections by penalising 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 orthogonolisation (equivalently incomplete Cholesky decomposition) to reduce the dimensionality of the kernel matrices.

4.1 Regularisation

To force non-trivial learning on the correlation we introduce a control on the flexibility of the projection mappings using Partial Least Squares (PLS) to penalise the norms of the associated weights. We convexly combine the PLS term with the KCCA term in the denominator of equation (3.7) obtaining

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

We observe that the new regularised equation is not affected by re-scaling of α or β, hence the optimisation problem is subject to

$$\begin{aligned}(\alpha' K_x^2\alpha + \kappa\alpha' K_x\alpha) &= 1 \\ (\beta' K_y^2\beta + \kappa\beta' K_y\beta) &= 1\end{aligned}$$

The corresponding Lagrangain is

$$\begin{aligned}L(\lambda_\alpha,\lambda_\beta,\alpha,\beta) &= \alpha' K_x K_y \beta \\ &\quad -\frac{\lambda_\alpha}{2}(\alpha' K_x^2\alpha + \kappa\alpha' K_x\alpha - 1) \\ &\quad -\frac{\lambda_\beta}{2}(\beta' K_y^2\beta + \kappa\beta' K_y\beta - 1).\end{aligned}$$

Taking derivatives in respect to α and β

$$\begin{aligned}\frac{\partial f}{\partial\alpha} &= K_x K_y\beta - \lambda_\alpha(K_x^2\alpha + \kappa K_x\alpha) &(4.1)\\ \frac{\partial f}{\partial\beta} &= K_y K_x\alpha - \lambda_\beta(K_y^2\beta + \kappa K_y\beta). &(4.2)\end{aligned}$$

Subtracting β' times the second equation from α' times the first we have

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

Which together with the constraints implies that $\lambda_\alpha - \lambda_\beta = 0$, let $\lambda = \lambda_\alpha = \lambda_\beta$. Consider the case where K_x and K_y are invertible, we have

$$\begin{aligned}\beta &= \frac{(K_y + \kappa I)^{-1} K_y^{-1} K_y K_x \alpha}{\lambda} \\ &= \frac{(K_y + \kappa I)^{-1} K_x \alpha}{\lambda}\end{aligned}$$

substituting in equation 4.1 gives

$$\begin{aligned}K_x K_y (K_y + \kappa I)^{-1} K_x \alpha &= \lambda^2 K_x (K_x + \kappa I)\alpha \\ K_y (K_y + \kappa I)^{-1} K_x \alpha &= \lambda^2 (K_x + \kappa I)\alpha \\ (K_x + \kappa I)^{-1} K_y (K_y + \kappa I)^{-1} K_x \alpha &= \lambda^2 \alpha\end{aligned}$$

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

4.2 Cholesky Decomposition

We describe some background information on direct factorisation methods on triangular decomposition [13].

$$LU = A \tag{4.3}$$

in which the diagonal elements of L are not necessarily unity. We consider $L \equiv (l_{ij})$ then equation (4.3) implies

$$l_{kk} u_{kk} = a_{kk} - \sum_{p=1}^{k-1} l_{kp} u_{pk} \qquad \text{for } k \geq 2 \tag{4.4}$$

$$u_{kj} = \frac{1}{l_{kk}} \left(a_{kj} - \sum_{p=1}^{k-1} l_{kp} u_{pj} \right) \quad \text{for } j > k \geq 2 \tag{4.5}$$

$$l_{ik} = \frac{1}{u_{kk}} \left(a_{ik} - \sum_{p=1}^{k-1} l_{ip} u_{pk} \right) \quad \text{for } i > k \geq 2 \tag{4.6}$$

Theorem 1. *Let A be symmetric. If the factorisation $LU = A$ is possible, then the choice $l_{kk} = u_{kk}$ implies $l_{ik} = u_{ki}$, that is, $LL^T = A$.*

Proof. Use equation (4.4) and induction on k.

A simple, non-singular, symmetric matrix for which the factorisation is not possible is

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

On the other hand, if the symmetric matrix A is positive definite (i.e., $x'Ax > 0$ if $x'x > 0$), then the factorisation is possible. We have

Theorem 2. *Let A be symmetric, positive definite. Then, A can be factored in the form*

$$LL' = A$$

Proof. If we define $l_{kk} = u_{kk} = \sqrt{b_{kk}}$ then we will obtain from the previous equations $LU = A$ where $l_{ik} = u_{ki}$

Incomplete Cholesky Decomposition

Complete decomposition of a kernel matrix is an expensive step and should be avoided with real world data. Incomplete Cholesky decomposition as described in [2] differs from Cholesky decomposition in that all pivots, which are below a certain threshold are skipped. If M is the number of non-skipped 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 factorisation if we require the rank to be as small as possible (Golub and loan, 1983).

We describe the algorithm from [2] (with slight modification) :

Input NxN matrix K
precision parameter η

1. Initialisation: $i = 1$, $K' = K$, $P = I$, for $j \in [1, N]$, $G_{jj} = K_{jj}$
2. While $\sum_{j=1}^{N} G_{jj} > \eta$ and $i! = N + 1$
   - Find best new element: $j^* = argmax_{j \in [i,N]} G_{jj}$
   - 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 i^{th} column of G:
     $G_{i+1:n,i} = \frac{1}{G_{ii}} \left(K'_{i+1:n,i} - \sum_{j=1}^{i-1} G_{i+1:n,j} G_{ij} \right)$
   - Update only diagonal elements: for $j \in [i+1, N]$, $G_{jj} = K'_{jj} - \sum_{k=1}^{i} G_{jk}^2$
   - Update $i = i + 1$
3. Output P, G and $M = i$

Output: an $N \times M$ lower triangular matrix G and a permutation matrix P such that $\|P'KP - GG'\| \leq \eta$ (appendix 1.2 for proof).

The algorithm involves picking one column of K at a time, choosing the column to be added by greedily maximising a lower bound on the reduction

in the error of the approximation. After l steps, we have an approximation of the form $\tilde{K}_l = G^i_l G'^i_l$, where G^i_l is $N \times l$. The ranking of the $N - l$ vectors can be computed by comparing the diagonal elements of the remainder matrix $K - G^i_l G'^i_l$.

Partial Gram-Schmidt Orthogonolisation

We explore the Partial Gram-Schmidt Orthogonolisation (PGSO) algorithm, described in [6], as our matrix decomposition approach. ICD could been 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 orthogonalisation 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 [2].

Given a kernel K from a training set, and precision parameter η:

Initialisations:

$m =$ 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$[i] = K_{ii}$;

Algorithm:

while $\sum_i norm2[i] > \eta$ and $j! = N + 1$ do
    $i_j = argmax_i(\text{norm2}[i])$;
    index$[j] = i_j$;
    size$[j] = \sqrt{norm2[i_j]}$;
    for $i = 1$ to m do
        feat$[i, j] = \frac{\left(k(d_i, d_{i_j}) - \sum_{t=1}^{j-1} feat[i,t] \cdot feat[i_j,t]\right)}{size[j]}$;
        norm2$[i]$ = norm2$[i]$− feat$(i, j)\cdot$ feat(i, j);
    end;
    $j = j + 1$
end;
return feat

Output:

$\|K - feat \cdot feat'\| \leq \eta$ where $feat$ is a $N \times M$ lower triangular matrix (appendix 1.2 for proof)

We observe that the output is equivalent to the output of ICD.

To classify a new example at location i:

Given a kernel K from a testing set

for $j = 1$ to M
    newfeat$[j] = (K_{i,index[j]} - \sum_{t=1}^{j-1}$ newfeat$[t] \cdot$ feat$[index[j], t])/$size$[j]$;
end;

The advantage of using the partial Gram-Schmidt orthonogolisation (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.

4.3 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 dimensiality.

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

$$\begin{aligned} K_x &\tilde{=} R_x R_x' \\ K_y &\tilde{=} R_y R_y' \end{aligned}$$

substituting the new representation into equations (3.8) and (3.9)

$$R_x R_x' R_y R_y' \beta - \lambda R_x R_x' R_x R_x' \alpha = 0 \quad (4.7)$$
$$R_y R_y' R_x R_x' \alpha - \lambda R_y R_y' R_y R_y' \beta = 0. \quad (4.8)$$

Multiplying the first equation with R_x' and the second equation with R_y' gives

$$R_x' R_x R_x' R_y R_y' \beta - \lambda R_x' R_x R_x' R_x R_x' \alpha = 0 \quad (4.9)$$
$$R_y' R_y R_y' R_x R_x' \alpha - \lambda R_y' R_y R_y' R_y R_y' \beta = 0. \quad (4.10)$$

Let Z be the new correlation matrix with the reduced dimensiality

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

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 (4.9) and (4.10) we find that we return to the primal representation of CCA with a dual representation of the data

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

Assuming 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}

$$\begin{aligned} Z_{xy}\tilde{\beta} - \lambda Z_{xx}\tilde{\alpha} &= 0 \quad (4.11) \\ Z_{yx}\tilde{\alpha} - \lambda Z_{yy}\tilde{\beta} &= 0. \quad (4.12) \end{aligned}$$

We are able to rewrite $\tilde{\beta}$ from equation (4.12) as

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

and substituting in equation (4.11) gives

$$Z_{xy}Z_{yy}^{-1}Z_{yx}\tilde{\alpha} = \lambda^2 Z_{xx}\tilde{\alpha} \quad (4.13)$$

we are left with a generalised 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 (4.13) we obtain

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

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

KCCA Regularisation with PGSO

We combine the dimensiality reduction introduced in the previous Section 4.3 with the regularisation parameter (Section 4.1) to maximise the learning. Following the same approach in the previous section we can rewrite equations (4.1) and (4.2) with the approximation of K_x and K_y as formulated in equations (4.7) and (4.8) respectively, in the following manner

$$\begin{aligned} R_x R_x' R_y R_y'\beta - \lambda(R_x R_x' R_x R_x' + \kappa R_x R_x')\alpha &= 0 \\ R_y R_y' R_x R_x'\alpha - \lambda(R_y R_y' R_y R_y' + \kappa R_y R_y')\beta &= 0 \end{aligned}$$

Multiplying the first equation with R_x' and the second equation with R_y' gives

$$\begin{aligned} R_x' R_x R_x' R_y R_y'\beta - \lambda R_x'(R_x R_x' R_x R_x' + \kappa R_x R_x')\alpha &= 0 \quad (4.14) \\ R_y' R_y R_y' R_x R_x'\alpha - \lambda R_y'(R_y R_y' R_y R_y' + \kappa R_y R_y')\beta &= 0 \quad (4.15) \end{aligned}$$

rewriting equation (4.14) with the new reduced correlation matrix Z as defined in the previous Section 4.3, we obtain

$$\begin{aligned} Z_{xx}Z_{xy}\tilde{\beta} - \lambda Z_{xx}(Z_{xx} + \kappa I)\tilde{\alpha} &= 0 \\ Z_{yy}Z_{yx}\tilde{\alpha} - \lambda Z_{yy}(Z_{yy} + \kappa I)\tilde{\beta} &= 0. \end{aligned}$$

Assuming 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}

$$\begin{aligned} Z_{xy}\tilde{\beta} - \lambda(Z_{xx} + \kappa I)\tilde{\alpha} &= 0 \quad (4.16) \\ Z_{yx}\tilde{\alpha} - \lambda(Z_{yy} + \kappa I)\tilde{\beta} &= 0. \quad (4.17) \end{aligned}$$

We are able to rewrite $\tilde{\beta}$ from equation (4.17) as

$$\tilde{\beta} = \frac{(Z_{yy} + \kappa I)^{-1} Z_{yx} \tilde{\alpha}}{\lambda}$$

substituting in equation 4.16 gives

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

We are left with a generalised eigenproblem of the form $A\mathbf{x} = \lambda B\mathbf{x}$. Performing a complete Cholesky decomposition on $Z_{xx} + \kappa I = SS'$ where S is a lower triangular matrix. and let $\hat{\alpha} = S'\tilde{\alpha}$, substituting in equation (4.18)

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

We obtain a symmetric generalised 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 4.3. 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 the authors of [15], where we are trying to facilitate mate retrieval on a test set. The data was divided into three classes (Figure 1) - Sport, Aviation and Paintball - 400 records each and consisted of jpeg images retrieved from the Internet with attached text. We randomly split each class into two halves which were used as training and test data accordingly. The extracted features of the data were used the same as in [15] (detailed description of the features used can be found in [15]: image HSV colour, image Gabor texture and term frequencies in text.

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 randomisation, the database with itself, and $\lambda_R(\kappa)$ be the spectrum with randomisation, the database with a randomised version of itself, (by spectrum it is meant that the vector whose entries are the eigenvalues). We would like to have the non-random spectrum as distant as possible from the randomised spectrum, as if the same correlation occurs for $\lambda(\kappa)$ and $\lambda_R\kappa$ then clearly overfitting is taking place. Therefor we expect for $\kappa = 0$ (no regularisation) and let $\mathbf{j} = 1, \ldots, 1$ (the all ones vector) that we may have $\lambda(\kappa) = \lambda_R(\kappa) = \mathbf{j}$, since it is very possible that the examples are linearly 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 κ so that the κ for which the

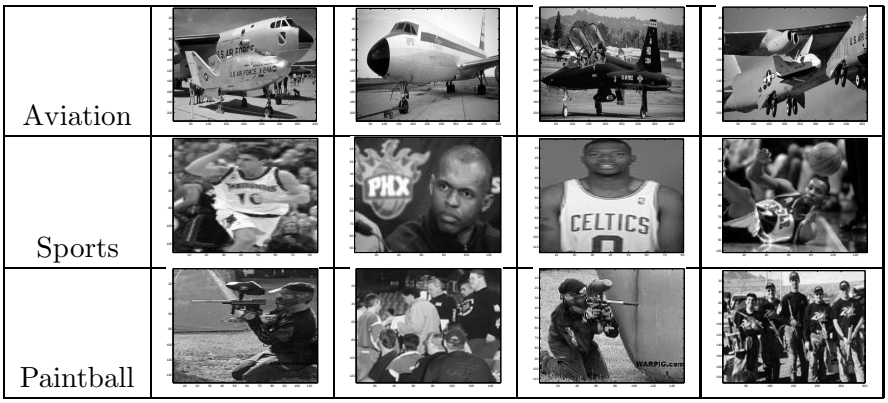

Figure 1 Example of images in database.

difference between the spectrum of the randomized set is maximally different (in the two norm) from the true spectrum.

$$\kappa = argmax \|\lambda_R(\kappa) - \lambda(\kappa)\|$$

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 Generalised 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 have used a Gaussian kernel, with $\sigma =$ max. distance/20, for the image colour component and all experiments were an average of 10 runs. For convenience we separate the content-based and mate-based approaches into the following Subsections 5.1 and 5.2 respectively.

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 (Figure 3 - retrieval

example for set of 5 images).

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

Table 1 Success cross-results between kernel-cca & generalised vector space.

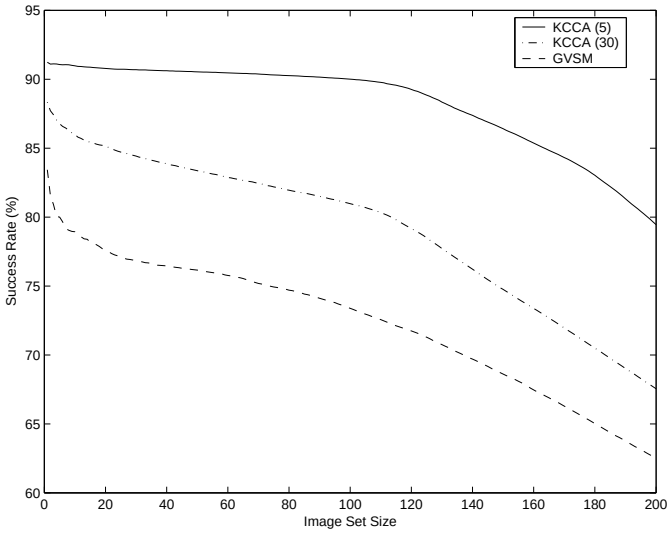

Figure 2 Success plot for content-based KCCA against GVSM

In Tables 1 and 2 we compare the performance of the kernel-CCA algorithm and generalised vector space model. In Table 1 we present the performance of the methods over 10 and 30 image sets where in Table 2 as plotted in Figure 2 we see the overall performance of the KCCA method against the GVSM for image sets $(1 - 200)$, as in the $200'th$ image set location the maximum of 200×600 of the same labelled images over all text queries can be retrieved (we only have 200 images per label). The success rate in Table 1 and Figure 2 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 projection. and as the semantic information needed is contained in the first

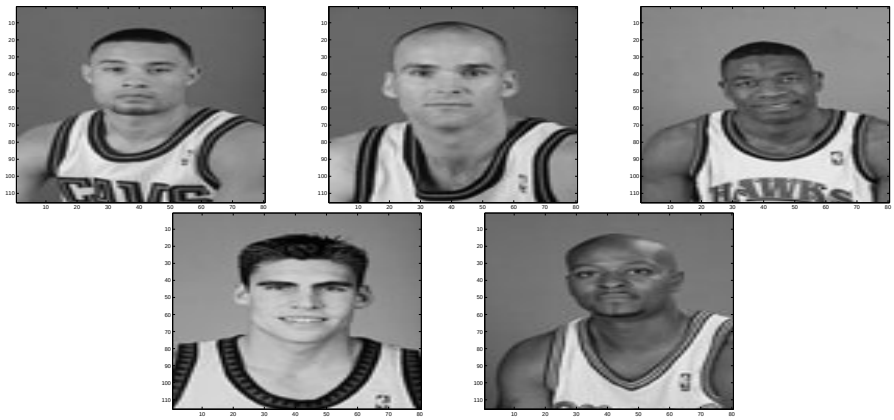

Figure 3 Images retrieved for the text query: "height: 6-11 weight: 235 lbs position: forward born: september 18, 1968, split, croatia college: none"

Method	overall success
GVSM	72.3%
KCCA (30)	79.12%
KCCA (5)	88.25%

Table 2 Success rate over all image sets $(1 - 200)$.

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 compute the success as the average of 10 runs (Figure 5 - retrieval example for set of 5 images).

Image set	GVSM success	KCCA success (30)	KCCA success (150)
10	8%	17.19%	59.5%
30	19%	32.32%	69%

Table 3 Success cross-results between kernel-cca & generalised vector space.

In Table 3 we compare the performance of the KCCA algorithm with the GVSM

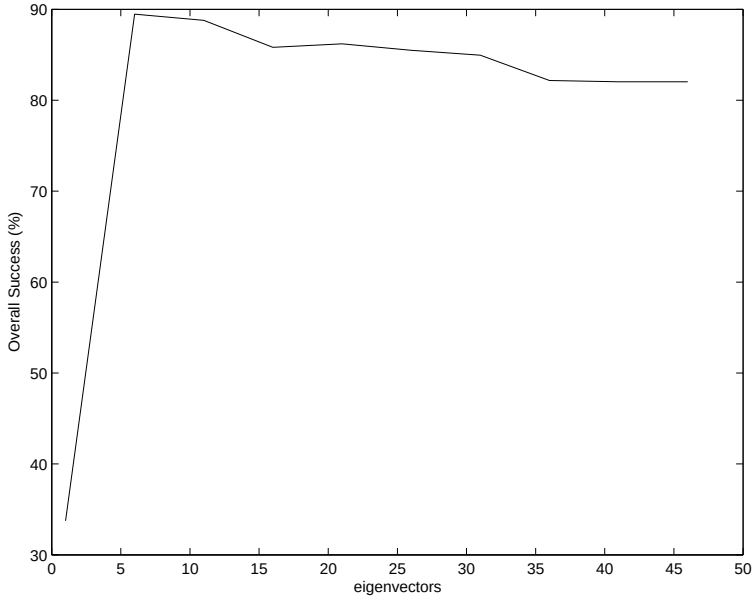

Figure 4 Content-Based plot of eigenvector selection against overall success (%).

over 10 and 30 image sets where 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.
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.

Method	overall success
GVSM	70.6511%
KCCA (30)	83.4671%
KCCA (150)	92.9781%

Table 4 Success rate over all image sets.

As visible in Figure 7 we find that unlike the Content-Based (Section 5.1) 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 towards exact correlation in the semantic projection. Though we do not compute for all eigenvectors as this process would be expensive

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 first row.

and the reminding eigenvectors would not necessarily add meaningful semantic information.

It is visible that the kernel-CCA significantly outperformes the GVSM method both in content retrieval and in mate retrieval.

5.3 Regularisation Parameter

We next verify that the method of selecting the regularisation parameter κ a priori gives a value performed well. We randomly split each class into two halves which were used as training and test data accordingly, we keep this divided set for all runs. We set the value of the incomplete Gram-Schmidt orthogonolisation precision parameter $\eta = 0.5$ 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 regularisation 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 its cheating). Though 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 in respect to the different values of κ and in figures 8 and 9 we view the plotting of the comparison. We observe that the difference in

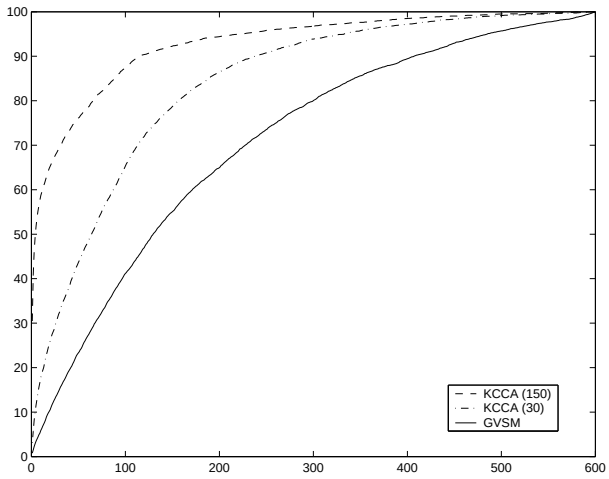

Figure 6 Success plot for KCCA mate-based against GVSM (success (%) against image set size).

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%

Table 5 Overall success of Content-Based (CB) KCCA with respect to κ.

performance between the a priori value $\hat{\kappa}$ and the new found optimal value κ for 5 eigenvectors is 1.0423% and for 30 eigenvectors is 5.031%. The more substantial increase in performance on the latter is due to the increase in the selection of the regularisation parameter, which compensates for the substantial decrease in performance (figure 6) of the content based retrieval, when high dimensional semantic feature space is used.

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**%

Table 6 Overall success of Mate-Based (MB) KCCA with respect to κ.

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

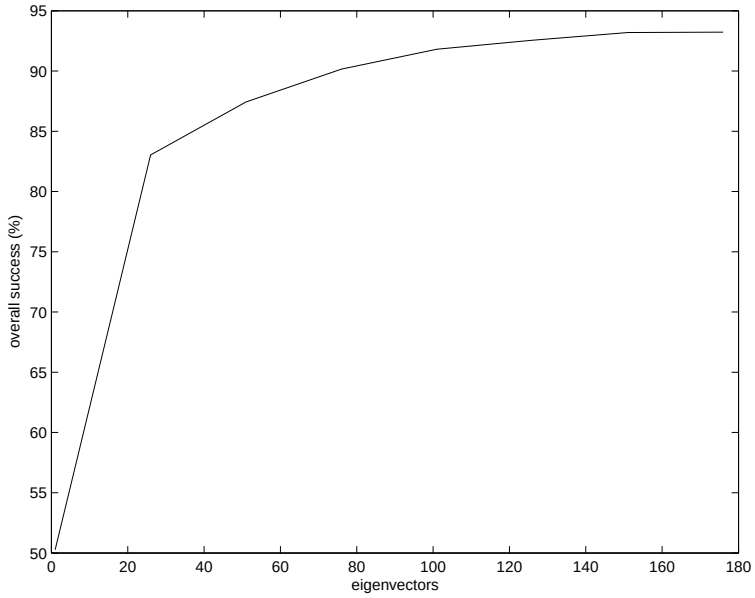

Figure 7 Mate-Based plot of eigenvector selection against overall success (%).

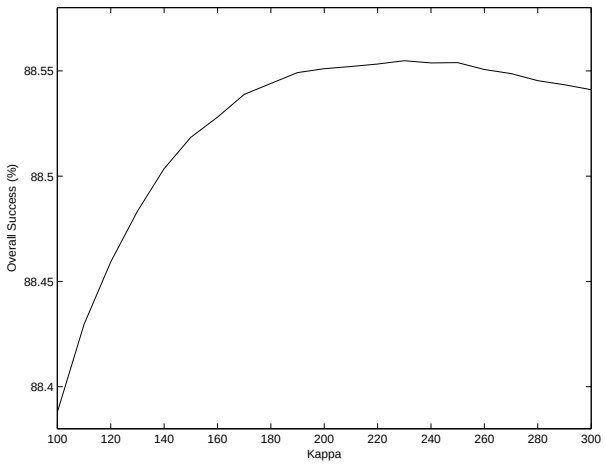

Figure 8 Content-Based. κ selection over overall success for 30 eigenvectors.

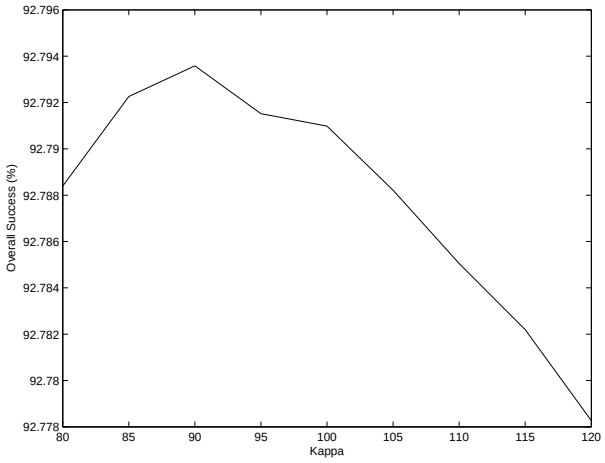

Figure 9 Content-Based. κ selection over overall success for 5 eigenvectors.

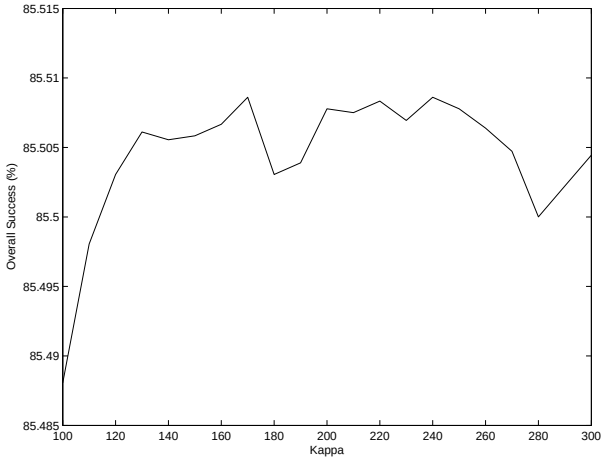

Figure 10 Mate-Based. κ selection over overall success for 30 eigenvectors.

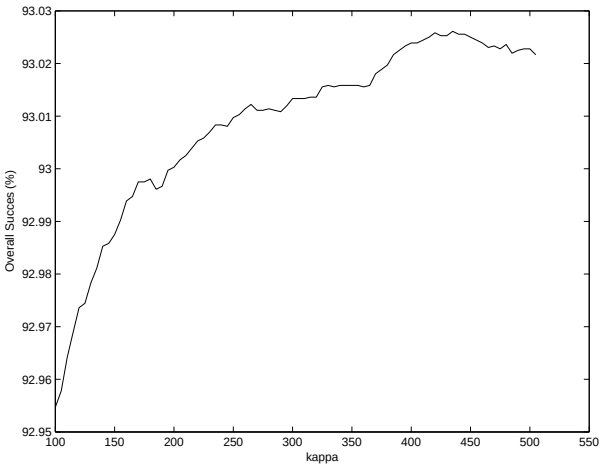

Figure 11 Mate-Based. κ selection over overall success for 150 eigenvectors.

performance between the a priori value $\hat{\kappa}$ and the new 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 regularisation parameter κ in an a priori fashion, since the difference between the actual optimal κ and the a priori $\hat{\kappa}$ is very slight.

6 Generalisation of Canonical Correlation Analysis

In this section we follow A. Gifi's book "Nonlinear Multivariate Analysis" (1990) and partially J. R. Ketterling "Canonical analysis of several sets of variables" (1971).

6.1 Some notations

For an $n \times n$ square matrix A having elements $\{a_{ij}\}$, $i, j = 1, \ldots, n$ we can define the trace by the formula

$$Tr(A) = \sum_i a_{ii} \tag{6.1}$$

the norm $\| \ \|_F$, so called the Frobenius norm, defined by

$$\|A\|_F = Tr\left(A'A\right) = \sum_{ij} a_{ij}^2 \tag{6.2}$$

and if a_i denotes the ith column(row) of A then we have

$$\|A\|_F = \sum_i \|a_i\|_2^2 = \sum_i \langle a_i, a_i \rangle \tag{6.3}$$

the notation $\| \ \|_2$ means the Euclidean, l_2, norm of a vector.

6.2 Some propositions

Proposition 3. *Let an optimisation problem be given in the form*

$$\min_{x,y} f(x,y) \tag{6.4}$$
$$\textit{subject to} \tag{6.5}$$
$$g(y) = 0, \tag{6.6}$$
$$x \in R^m, y \in R^n. \tag{6.7}$$

Let the set $Y \subseteq R^n$ the feasibility domain for y determined by the constrain $g(y) = 0$.

Assume the function f is convex in both variables x and y, the optimal solution of x can be expressed by the function $h(y)$ of the optimal solution of y, where the function of h is defined on the whole set Y and the functions f, g, h are twice continuously differentiable on $R^m \times Y$.

Then the optimisation problem with the same constrain

$$\min_{y} f(h(y), y) \tag{6.8}$$
$$\textit{subject to} \tag{6.9}$$
$$g(y) = 0, \tag{6.10}$$
$$y \in R^n, \tag{6.11}$$

has the same optimal solution in y than equation (6.4) has.

Proof. Let the optimal solution of equation (6.4) be denoted by x_1, y_1 and for equation (6.8) be denoted by y_2.

Based on the condition of the proposition we have $x_1 = h(y_1)$. Because y_1 is a feasible solution for equation (6.8) thus $f(x_1, y_1) = f(h(y_1), y_1) \geq f(h(y_2), y_2)$, but the objective function of equation (6.4) is not restricted in the first variable, thus the inequality $f(x_1, y_1) \leq f(h(y_2), y_2)$ holds, hence $f(x_1, y_1) = f(h(y_2), y_2)$.

From the convexity of f and the same feasibility domains the optimum solutions have to be the same. □

6.3 Formulation of the Canonical Correlation

Let $H^{(1)}, H^{(2)}$ be matrices with size $m \times n_1, m \times n_2$ respectively and assume the sum of the elements in the columns of these matrices are equal to 0, they are centralised and they are linearly independent vectors within one matrix. We consider arbitrary linear combinations of the columns of these matrices in the form $H^{(1)}a_i^{(1)}, H^{(2)}a_i^{(2)}, i = 1, \ldots, p$. Let $A^{(1)} = a_1^{(1)}, \ldots, a_{n_1}^{(1)}$ and $A^{(2)} = a_1^{(2)}, \ldots, a_{n_2}^{(2)}$ be matrices comprising the vectors of the linear combinations as

columns. Introducing notations for the product of the matrices to simplify the formulas:

$$\Sigma_{ij} = H'_{(i)} H_{(j)},\ i, j = 1, 2. \tag{6.12}$$

We are looking for linear combinations of the columns of these matrices such that the first pair of the vectors $(a_1^{(1)}, a_1^{(2)})$ are optimal solution of the optimisation problem:

$$\max_{a_1^{(1)}, a_1^{(2)}} a_1^{(1)'} \Sigma_{12} a_1^{(2)} \tag{6.13}$$

$$subject\ to \tag{6.14}$$

$$a_1^{(1)'} \Sigma_{11} a_1^{(1)} = 1, \tag{6.15}$$

$$a_1^{(2)'} \Sigma_{22} a_1^{(2)} = 1. \tag{6.16}$$

The meaning of this optimisation problem is to find the maximum correlation between the linear combinations of the columns of the matrices $H^{(1)}, H^{(2)}$, subject to the length of the vectors corresponding to these linear combinations normalised to 1.

To determinate the remaining pairs of the vectors, columns in $A^{(1)}$ and $A^{(2)}$, a series of optimisation problems are solved successively. For the pair of the vectors $(a_r^{(1)}, a_r^{(2)}),\ r = 2, \ldots, p$ we have

$$\max_{a_r^{(1)}, a_r^{(2)}} a_r^{(1)'} \Sigma_{12} a_r^{(2)} \tag{6.17}$$

$$subject\ to \tag{6.18}$$

$$a_r^{(k)'} \Sigma_{kk} a_r^{(k)} = 1, \tag{6.19}$$

$$a_r^{(k)'} \Sigma_{kk} a_j^{(k)} = 0, \tag{6.20}$$

$$a_r^{(k)'} \Sigma_{kl} a_j^{(l)} = 0, \tag{6.21}$$

$$k, l = 1, 2,\ j = 1, \ldots, r-1. \tag{6.22}$$

The problem (6.13) expanded by the orthogonality constrains (6.17), namely the components of every new pair in the iteration have to be orthogonal to the components of the previous pairs.

The upper limit p of the iteration has to be $\leq \min(rank(H_{(1)}), rank(H_(2)))$.

Applying the Karush-Kuhn-Tucker conditions we can express the optimal solutions of the problem (6.13) and the problems (6.17) for $r = 2, \ldots, p$. Let's begin with the problem (6.17).

First we apply a substitution such that

$$a_i^{(k)} = \Sigma_{kk}^{-\frac{1}{2}} y_i^{(k)}, \tag{6.23}$$

$$D_{kl} = \Sigma_{kk}^{-\frac{1}{2}} \Sigma_{kl} \Sigma_{ll}^{-\frac{1}{2}}, \tag{6.24}$$

$$k, l = 1, 2,\ i = 1, \ldots, p, \tag{6.25}$$

Thus we have the problem

$$\max_{y_1^{(1)}, y_1^{(2)}} y_1^{(1)'} D_{12} y_1^{(2)} \tag{6.26}$$

$$subject\ to \tag{6.27}$$

$$y_1^{(k)'} y_1^{(k)} = 1, k = 1, 2. \tag{6.28}$$

$$\tag{6.29}$$

The Lagrangian of this problem has the form

$$L_1 = y_1^{(1)'} D_{12} y_1^{(2)} + \frac{1}{2}\lambda_1(1 - y_1^{(1)'} y_1^{(1)}) + \frac{1}{2}\lambda_2(1 - y_2^{(2)'} y_2^{(2)}), \tag{6.30}$$

where λ_1 and λ_2 are the Lagrangian multipliers. The vectors of the partial derivatives of L_1respect to the vectors $y_1^{(1)}, y_1^{(2)}$ are equal to 0 by the KKT conditions, thus we get

$$\frac{\partial L_1}{\partial y_1^{(1)}} = 2D_{12} a_1^{(2)} - 2\lambda_1 y_1^{(1)} = \mathbf{0}, \tag{6.31}$$

$$\frac{\partial L_1}{\partial y_1^{(2)}} = 2D_{21} y_1^{(1)} - 2\lambda_2 y_1^{(2)} = \mathbf{0}. \tag{6.32}$$

Multiplying equation (6.31) by $y_1^{(1)'}$ and equation (6.32) by $y_1^{(2)'}$ and dividing by the constant 2 provides

$$y_1^{(1)'} D_{12} y_1^{(2)} - \lambda_1 y_1^{(1)'} y_1^{(1)} = \mathbf{0}, \tag{6.33}$$

$$y_1^{(2)'} D_{12}' y_1^{(1)} - \lambda_2 y_1^{(2)'} y_1^{(2)} = \mathbf{0}. \tag{6.34}$$

Based on the constrains of the optimisation problem (6.26) and the identity $D_{21}' = D_{12}$ we have

$$\lambda_1 = \lambda_2 = y_1^{(1)'} D_{12} y_1^{(2)}. \tag{6.35}$$

After replacing λ_1 and λ_2 with λ the following equality system can be formulated

$$\begin{pmatrix} -\lambda I & D_{12} \\ D_{21} & -\lambda I \end{pmatrix} \begin{pmatrix} y_1^{(1)} \\ y_1^{(2)} \end{pmatrix} = \mathbf{0}. \tag{6.36}$$

It is not too hard to realise this equality system is a singular vector and value problem of the matrix D_{12} having $y_1^{(1)}$ and $y_1^{(2)}$ are a left and a right singular vectors and the value of the Lagrangian λ is equal to the corresponding singular value. Based on this statements we can claim that the optimal solutions are the singular vectors belonging to the greatest singular value of the matrix D_{12}.

Considering the successive optimisation problem and applying similar substitution for the all variables $a_i^{(k)}$ as introduced in equation (6.23), a problem with the greatest r singular values and the corresponding left and right singular vectors arises.

6.4 The simultaneous formulation of the canonical correlation

Instead of using the successive formulation of the canonical correlation we can join the subproblems into one. The simultaneous formulation is the optimisation problem

$$\max_{(a_1^{(1)},a_1^{(2)}),\ldots,(a_p^{(1)},a_p^{(2)})} \sum_{i=1}^{p} a_i^{(1)'} \Sigma_{12} a_i^{(2)} \tag{6.37}$$

$$subject\ to \tag{6.38}$$

$$a_i^{(1)'} \Sigma_{11} a_j^{(1)} = \begin{cases} 1 & if\ i = j, \\ 0 & otherwise, \end{cases} \tag{6.39}$$

$$a_i^{(2)'} \Sigma_{22} a_j^{(2)} = \begin{cases} 1 & if\ i = j, \\ 0 & otherwise, \end{cases} \tag{6.40}$$

$$i, j = 1, \ldots, p, \tag{6.41}$$

$$a_i^{(1)'} \Sigma_{12} a_j^{(2)} = 0, \tag{6.42}$$

$$i, j = 1, \ldots, p,\ j \neq i. \tag{6.43}$$

Based on equation (6.37) and the definition of the Frobenius norm we have a compact formulation of the canonical correlation problem:

$$\max_{A^{(1)},A^{(2)}} Tr\left(A^{(1)'} \Sigma_{12} A^{(2)}\right) \tag{6.44}$$

$$subject\ to \tag{6.45}$$

$$A^{(k)'} \Sigma_{kk} A^{(k)} = I, \tag{6.46}$$

$$a_i^{(k)'} \Sigma_{kl} a_j^{(l)} = 0, \tag{6.47}$$

$$k, l = \{1, 2\},\ l \neq k, i, j = 1, \ldots, p,\ j \neq i. \tag{6.48}$$

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

Repeating the substitution in equation (6.23) the set of feasible vectors for the simultaneous problem is equal to the left and right singular vectors of matrix D_{12}, hence the optimal solution is compatible to the successive problems.

6.5 Correlation versus Distance

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

$$\min_{A^{(1)},A^{(2)}} \left\| H^{(1)} A^{(1)} - H^{(2)} A^{(2)} \right\|_F \tag{6.49}$$

$$subject\ to \tag{6.50}$$

$$A^{(k)'} \Sigma_{kk} A^{(k)} = I, \tag{6.51}$$

$$a_i^{(k)'} \Sigma_{kl} a_j^{(l)} = 0, \tag{6.52}$$

$$k, l = 1, \ldots, 2,\ l \neq k, i, j = 1, \ldots, p,\ j \neq i. \tag{6.53}$$

Unfolding the objective function of the minimisation problem (6.49) shows the optimisation problem is the same as the maximisation problem (6.44).

6.6 The generalisation of canonical correlation

Exploiting the distance problem we can give a generalisation of the canonical correlation for more than two known matrices. Given a set of matrices $\{H^{(1)}, \ldots, H^{(K)}\}$ 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 $A(1), \ldots, A(K)$ such that they gives the optimum solution of the problem

$$\min_{A^{(1)},\ldots A^{(K)}} \sum_{k,l=1}^{K} \left\| H^{(k)}A^{(k)} - H^{(l)}A^{(l)} \right\|_F \tag{6.54}$$

$$\textit{subject to} \tag{6.55}$$

$$A^{(k)\prime}\Sigma_{kk}A^{(k)} = I, \tag{6.56}$$

$$a_i^{(k)\prime}\Sigma_{kl}a_j^{(l)} = 0, \tag{6.57}$$

$$k, l = 1, \ldots, K,\ l \neq k, i, j = 1, \ldots, p,\ j \neq i. \tag{6.58}$$

In the forthcoming sections we will show how to simplify this problem.

6.7 Total Distance versus Variance

Given a set of vectors $X = x_1, \ldots, x_m \subseteq R^n$. The notation x_{ki} means the ith component of the vector x_k.

The total squared distance, the sum of the squared Euclidean distance of all possible pairs of vectors in X is equal to

$$\frac{1}{2}\sum_{k=1}^{m}\sum_{l=1,l\neq k}^{m} \|x_k - x_l\|_2^2 = \tag{6.59}$$

as for any k, $\|x_k - x_k\| = 0$ we can drop the constrain $l \neq k$, thus we have

$$= \frac{1}{2}\sum_{k=1,l=1}^{m} \|x_k - x_l\|_2^2 = \tag{6.60}$$

$$= \frac{1}{2}\sum_{k=1,l=1}^{m}\sum_{i=1}^{n}(x_{ki} - x_{li})^2 = \tag{6.61}$$

$$= \frac{1}{2}\sum_{k=1,l=1}^{m}\sum_{i=1}^{n}(x_{ki}^2 + x_{li}^2 - 2x_{ki}x_{li}) = \tag{6.62}$$

$$= \frac{1}{2}\sum_{i=1}^{n}\left(\sum_{k=1,l=1}^{m} x_{ki}^2 + \sum_{k=1,l=1}^{m} x_{li}^2 - \sum_{k=1,l=1}^{m} 2x_{ki}x_{li}\right) \tag{6.63}$$

$$= \frac{1}{2}\sum_{i=1}^{n}\left(m\sum_{k=1}^{m} x_{ki}^2 + m\sum_{l=1}^{m} x_{li}^2 - 2\sum_{k=1}^{m} x_{ki}\sum_{l=1}^{m} x_{li}\right) \tag{6.64}$$

to simplify the formula we introduce

$$M_1^{(i)} = \frac{1}{m}\sum_{k=1}^{m} x_{ki},\ M_2^{(i)} = \frac{1}{m}\sum_{k=1}^{m} x_{ki}^2, \tag{6.65}$$

we can reformulate equation (6.64)

$$= \sum_{i=1}^{n}\left(m^2 M_2^{(i)} - m^2 (M_1^{(i)})^2\right) = \tag{6.66}$$

applying the well-known identity of the variance for the vectors $(x_{11},\ldots,x_{m1}),\ldots,(x_{1n},\ldots x_{mn})$ gives

$$= m^2 \sum_{i=1}^{n}\sum_{k=1}^{m}(x_{ki} - M_1^{(i)})^2. \tag{6.67}$$

Hence the total squared distance turns to be equal to the sum of the component-wise variances of the vectors in X multiplied by the square of the number of the vectors.

Another statement about the variance is introduced. If we have the following optimisation problem

$$\min_z \|z - x_k\|_2^2,\ x_k \in X \text{ and } z \in R^n, \tag{6.68}$$

then the optimal solution can be expressed by

$$z_i = \frac{1}{n}\sum_{k=1}^{m} x_{ki}. \tag{6.69}$$

The components of the optimal solution are equal to the mean values of the corresponding components of the known vectors.

6.8 General form

Let $H^{(1)},\ldots,H^{(K)}$ be a set of known matrices with size $m \times n_1,\ldots,m \times n_K$ and X be an unknown matrix with size $m \times p$. The columns of the matrices $H^{(1)},\ldots,H^{(K)}$ are centralised, i.e. the mean of every column in every matrix is equal to 0. We assume the columns of every matrix $H^{(k)}, k = 1,\ldots,K$ are linearly independent. A notation to simplify the formulas, is introduced; $\Sigma_{kl} = H^{(k)T}H^{(l)}$. We are looking for linear combinations of the columns of the known matrices and a corresponding X such that they are the optimal solution of the optimisation problem given by

$$\min_{X,A^{(1)},\ldots,A^{(K)}} \frac{1}{K}\sum_{k=1}^{K}\left\|X - H^{(k)}A^{(k)}\right\|_F \tag{6.70}$$

$$\text{subject to} \tag{6.71}$$

$$a_i^{(k)\prime}\Sigma_{kl}a_j^{(l)} = \begin{cases} 1 & if\ k = l \text{ and } i = j, \\ 0 & if\ (k = l \text{ and } i \neq j) \text{ or } (k \neq l \text{ and } i \neq j), \end{cases} \tag{6.72}$$

$$k,l = 1,\ldots,K,\ i,j = 1,\ldots,p,\ \text{except when } k \neq l \text{ and } i = j, \tag{6.73}$$

where $a_i^{(k)}$ denotes the ith column of the matrix $A^{(k)}$ containing the possible linear combinations.

Applying substitutions for all $k = 1, \ldots, K,\ i = 1, \ldots, p$

$$a_i^{(k)} = \Sigma_{kk}^{-\frac{1}{2}} y_i^{(k)}, \tag{6.74}$$

where we can compute the inverse because the columns of the matrix $H^{(k)}$ are independent meaning Σ_{kk} has full rank. We can transform this optimisation problem into a more simply form. First, we modify the set of constrains. To make this modification readable the notation is introduced

$$\Sigma_{kk}^{-\frac{1}{2}} \Sigma_{kl} \Sigma_{ll}^{-\frac{1}{2}} = D_{kl},\ k, l = 1, \ldots, K, \tag{6.75}$$

where we exploit the symmetricity of the matrices $\Sigma_{kk}^{-\frac{1}{2}}$.

Thus the constrains get the form

$$y_i^{(k)'} y_j^{(k)} = \begin{cases} 1 & if\ i = j, \\ 0 & if\ i \neq j, \end{cases} \tag{6.76}$$

$$k = 1, \ldots, K,\ i, j = 1, \ldots, p, \tag{6.77}$$

$$y_i^{(k)'} D_{kl} y_j^{(l)} = 0, \tag{6.78}$$

$$k, l = 1, \ldots, K,\ k \neq l,\ i, j = 1, \ldots p,\ i \neq j, \tag{6.79}$$

for which we can recognise the singular decomposition problems of the matrices $\{D_{kl}\}$. If we consider the matrix D_{kl} for a fixed pair of the indeces k, l and apply the singular decomposition we have

$$D_{kl} = Y^{(k)} \Lambda_{kl} Y^{(l)'}, \tag{6.80}$$

the matrices $Y^{(k)}$ and $Y^{(l)}$ have columns being equal to the vectors $y_i^{(k)}$ and $y_i^{(l)}$ respectively, where $i = 1, \ldots, p$. The singular decomposition Λ_{kl} is a diagonal matrix and $Y^{(k)'} Y^{(k)} = I,\ Y^{(l)'} Y^{(l)} = I$. The constrains do not contain the items having indeces with the properties $k \neq l$ and $i = j$. They give the singular values of the matrix D_{kl}

$$y_i^{(k)'} D_{kl} y_i^{(l)} = \Lambda_{ii}. \tag{6.81}$$

The consequence of the singular decomposition form is that the set of the feasible solutions of the optimisation problem with constrains (6.76) are equal to the set of the singular vectors of the matrices $\{D_{kl}, k, l = 1 \ldots, K\}$.

To express the objective function of the optimisation problem (6.70) we use the notations

$$Q_k = H^{(k)} \Sigma_{kk}^{-\frac{1}{2}}, \tag{6.82}$$

$$D_{kl} = Q_k^T Q_l. \tag{6.83}$$

We can derive another statement about the optimal solution of the problem. Exploiting the definition of the Frobenius norm the objective function (6.70) can be rewritten as a sum of the Euclidean norm of the column vectors, where x_i denotes the ith column of the matrix X,

$$\frac{1}{K}\sum_{k=1}^{K}\left\|X - H^{(k)}A^{(k)}\right\|_F = \tag{6.84}$$

$$= \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{p}\left\|x_i - H^{(k)}a_i^{(k)}\right\|_2^2 = \tag{6.85}$$

$$= \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{p}\left\|x_i - Q_k y_i^{(k)}\right\|_2^2 = \tag{6.86}$$

$$= \frac{1}{K}\sum_{k=1}^{K}\sum_{i=1}^{p}\langle x_i - Q_k y_i^{(k)}, x_i - Q_k y_i^{(k)}\rangle. \tag{6.87}$$

The constrains are formulated in equation (6.76).

For the Lagrangian function of the optimisation problem we have:

$$L = \sum_{k=1}^{K}\sum_{i=1}^{p}\langle x_i - Q_k y_i^{(k)}, x_i - Q_k y_i^{(k)}\rangle + \tag{6.88}$$

$$+\sum_{k}^{K}\sum_{i}^{p}\lambda_{k,ii}\left(1 - y_i^{(k)\prime}y_i^{(k)}\right) + \tag{6.89}$$

$$+\sum_{k}^{K}\sum_{\substack{i,j\\ i\neq j}}^{p}\lambda_{k,ij}\left(-y_i^{(k)\prime}y_j^{(k)}\right) + \tag{6.90}$$

$$+\sum_{\substack{k,l\\ k\neq l}}^{K}\sum_{\substack{i,j\\ i\neq j}}^{p}\lambda_{kl,ij}\left(-y_i^{(k)\prime}D_{kl}y_j^{(l)}\right). \tag{6.91}$$

We disregard the constant $\frac{1}{K}$ from the objective function (6.70).

After computing the partial derivatives, where x_i signs the ith column of the matrix X, we get

$$\frac{\partial L}{\partial x_i} = \sum_{k=1}^{K}\left(2x_i - 2Q^k y_i^{(k)}\right) = 0,\ i = 1,\ldots,p, \tag{6.92}$$

$$\frac{\partial L}{\partial y_i^{(k)}} = 2D_{kk}y_i^{(k)} - 2Q^{k\prime}x_i - 2\lambda_{k,ij}\sum_{j}^{p}y_j^{(k)} - 2\sum_{\substack{l\\ l\neq k}}^{K}\sum_{\substack{j\\ j\neq i}}^{p}\lambda_{kl,ij}D_{kl}y_j^{(l)} = 0, \tag{6.93}$$

$$k = 1,\ldots,K,\ i = 1,\ldots,p. \tag{6.94}$$

We can express x_i for any $i = 1, \ldots, p$ from (6.92)

$$x_i = \frac{1}{K} \sum_{l=1}^{K} Q^l y_i^{(l)}, \ i = 1, \ldots, p. \tag{6.95}$$

Based on the proposition (3) we can replace the variable X in equation (6.70) by an expression of the other variables without changing the optimum value and the optimal solution. Thus we have the variance problem.

7 Conclusions

Through 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 Generalised Vector Space Model. We demonstrate that one can choose the regularisation 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 generalisation of the canonical correlation analysis we can see that the original problem can be transformed and reinterpreted as a total distance problem or variance minimisation problem. This special duality between the correlation and the distance requires more investigation to give more suitable description of the structure of some special spaces generated by different kernels.

These approaches can give 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. to do cluster analysis in the kernel space and to compute the coordinates of the cluster centres only.

Acknowledgments

We would like to acknowledge the financial support of EU Projects KerMIT, No. IST-2000-25341 and LAVA, No. IST-2001-34405.

1 Proof $\|K - G^i G^{i'}\| \leq \eta$

1.1 Some notation

Lemma 4. *Let A and B be an square matrices such that $Trace(A) = \sum_i^n a_{ii}$ then we have $Trace(AB) = Trace(BA)$*

Proof.

$$\begin{aligned} Trane(AB) &= \sum_{i}^{n}(ab)_{ii} \\ &= \sum_{i,j}^{n} a_{ij}b_{ji} \\ &= \sum_{j,i}^{n} b_{ji}a_{ij} \\ &= \sum_{j}^{n}(ba)_{jj} \\ &= Trace(BA) \end{aligned}$$

□

Lemma 5. *Let A be a symmetric matrix having eigenvalue decomposition equal to $A = V'\Lambda V$ (we are able to write $\Lambda = V'AV$) and using Lemma 4, then $Trace(\Lambda) = Trace(A)$.*

Proof.

$$\begin{aligned} Trace(\Lambda) &= Trace(V'AV) \\ &= Trace((V'A)V) \\ &= Trace(V(V'A)) \\ &= Trace(VV'A) \\ &= Trace(A) \end{aligned}$$

Hence we show that the following holds

$$\sum_{i}^{n} a_{ii} = \sum_{i}^{n} \lambda_i$$

□

Lemma 6. *If we have a symmetric matrix A, the Euclidian norm is equal with the maximum eigenvalue of A*

Proof.

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|}.$$

For any $c \in \mathbb{R}$ the scaling does not change

$$\frac{\|cAx\|}{\|cx\|} = \frac{c\|Ax\|}{c\|x\|} = \frac{\|Ax\|}{\|x\|}$$

Hence we obtain

$$\begin{aligned} \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} &= \max_{\|x\|=1} \|Ax\| \\ \|Ax\| &= \sqrt{(x'A'Ax)} \\ \|Ax\|^2 &= x'A'Ax \end{aligned}$$

Let UDU' be the eigenvalue decomposition of AA' such that D is a diagonal matrix containing square of the eigenvalues of A

$$\begin{aligned} A'A &= UDU' \\ \|Ax\|^2 &= x'UDUx \end{aligned}$$

Setting $w = U'x$ and as U is orthognoal we can rewrite $\|x\| = 1$ to $\|w\| = 1$

$$\begin{aligned} \|A\|^2 &= w'Dw \\ &= \sum \lambda_i^2 w_i^2 \end{aligned}$$

We can see that the following holds

$$\max_{(\sum w_i^2 = 1)} \sum \lambda_i^2 w_i^2 = \max_i \lambda_i^2$$

Hence we obtain

$$\|A\| = \max_i \lambda_i$$

□

1.2 Proof

Theorem 7. *If K is a positive definite matrix and GG' is its incomplete cholesky decomposition then the Euclidian norm of GG' subtracted from K is less than or equal to the trace of the uncalculated part of K. Let ΔK^i be the uncalculated part of K and let $\eta = Trace(\Delta K^i)$ then $\|K - G^i G^{i'}\| \leq \eta$.*

Proof. Let GG' be the being the complete cholesky decomposition $K = GG'$ where G is a lower triangular matrix were the upper triangular is zeros.

$$G = \begin{bmatrix} A & 0 \\ B & C \end{bmatrix}.$$

Let $G^i G^{i'}$ to be the incopmlete decomposition of K where i are the iterations of the Cholesky factorization procedure

$$G^i = G_{1:n,1:i} = \begin{bmatrix} A \\ B \end{bmatrix}$$

such that $G^i G^{i'} = \tilde{K^i}$, where $\tilde{K^i}$ is the approximation of K subject to a symmetric permutation of rows and columns. Assuming that the rows and columns

of K are ordered and no permutation is necessary (this is only for convenience of the proof). Let $\Delta K^i = K - \tilde{K^i}$.
Let $A \in G_{1:i,1:i}$, $B \in G_{i+1:n,1:i}$ and $C \in G_{1+i:n,1+i:n}$

$$K = GG' = \begin{bmatrix} AA' & AB' \\ BA' & BB' + CC' \end{bmatrix}$$
$$\tilde{K^i} = G^i G^{i'} = \begin{bmatrix} AA' & AB' \\ BA' & BB' \end{bmatrix}$$
$$\Delta K^i = \begin{bmatrix} 0 & 0 \\ 0 & CC' \end{bmatrix}$$

We show that CC' is positive semi-definite

$$\begin{aligned} CC' &= K_{i+1:n,i+1:n} - \tilde{K^i}_{i+1:n,i+1:n} \\ &= K_{i+1:n,i+1:n} - BB' \\ &= K_{i+1:n,i+1:n} - B \cdot A^{-1}A \cdot B' \\ &= K_{i+1:n,i+1:n} - B \cdot A^{-1} \cdot (AB') \\ &= K_{i+1:n,i+1:n} - G_{i+1:n,1:i} \cdot G^{-1}_{1:i,1:i} \cdot K_{1:i,i+1:n} \\ &= K_{i+1:n,i+1:n} - G^i_{i+1:n} \cdot {G^i_{1:i}}^{-1} \cdot K_{1:i,i+1:n} \end{aligned}$$

therefore

$$\begin{aligned} xCC'x &= < xC, (xC)' > \\ &\geq 0 \\ \lambda_c &\geq 0 \end{aligned}$$

CC' is a positive semi-definite matrix, hence ΔK^i is also a positive semi-definite matrix. Using Lemma 6 we are now able to show that

$$\begin{aligned} \|K - \tilde{K^i}\| &= \|\Delta K^i\| \\ \|K - G^i G^{i'}\| &= \|\Delta K^i\| \\ &= \sum_i^n \lambda_i w_i \\ &= \max_i \lambda_i \end{aligned}$$

As the maximum eigenvalue is less than or equal to the sum of all the eigenvalues, using Lemma 5, we are able to rewrite the expression as

$$\begin{aligned} \|K - G^i G^{i'}\| &\leq \sum_i^n \lambda_i \\ &\leq Trace(\Lambda) \\ &\leq Trace(\Delta K^i_{ii}). \end{aligned}$$

Therefore,

$$\|K - G^i G^{i'}\| \leq \eta.$$

□

Bibliography

[1] Shotaro Akaho. A kernel method for canonical correlation analysis. In *International Meeting of Psychometric Society*, Osaka, 2001.

[2] Francis Bach and Michael Jordan. Kernel independent component analysis. *Journal of Machine Leaning Research*, 3:1–48, 2002.

[3] Magnus Borga. *Learning Multidimensional Signal Processing*. PhD thesis, Linkping Studies in Science and Technology, 1998.

[4] Magnus Borga. Canonical correlation *a tutorial*, 1999.

[5] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press, 2000.

[6] Nello Cristianini, John Shawe-Taylor, and Huma Lodhi. Latent semantic kernels. In Caria Brodley and Andrea Danyluk, editors, *Proceedings of ICML-01, 18th International Conference on Machine Learning*, pages 66–73. Morgan Kaufmann Publishers, San Francisco, US, 2001.

[7] Colin Fyfe and Pei Ling Lai. Ica using kernel canonical correlation analysis.

[8] Colin Fyfe and Pei Ling Lai. Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems*, 2001.

[9] A. Gifi. *Nonlinear Multivariate Analysis*. Wiley, 1990.

[10] G. H. Golub and C. F. V. Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, MD, 1983.

[11] David R. Hardoon and John Shawe-Taylor. Kcca for different level precision in content-based image retrieval. In *Submitted to Third International Workshop on Content-Based Multimedia Indexing*, IRISA, Rennes, France, 2003.

[12] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:312–377, 1936.

[13] E. Isaacson and H. B. Keller. *Analysis of Numerical Methods*. John Wiley & Sons, Inc, 1966.

[14] J. R. Ketterling. Canonical analysis of several sets of variables. *Biometrika*, 58:433–451, 1971.

[15] T. Kolenda, L. K. Hansen, J. Larsen, and O. Winther. Independent component analysis for understanding multimedia content. In H. Bourlard, T. Adali, S. Bengio, J. Larsen, and S. Douglas, editors, *Proceedings of IEEE Workshop on Neural Networks for Signal Processing XII*, pages 757–766, Piscataway, New Jersey, 2002. IEEE Press. Martigny, Valais, Switzerland, Sept. 4-6, 2002.

[16] Malte Kuss and Thore Graepel. The geometry of kernel canonical correlation analysis. 2002.

[17] Yong Rui, Thomas S. Huang, and Shih-Fu Chang. Image retrieval: Current techniques, promising directions, and open issues. *Journal of Visual Communications and Image Representation*, 10:39–62, 1999.

[18] Alexei Vinokourov, David R. Hardoon, and John Shawe-Taylor. 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, 2003.

[19] Alexei Vinokourov, John Shawe-Taylor, and Nello Cristianini. Inferring a semantic representation of text via cross-language correlation analysis. In *Advances of Neural Information Processing Systems 15 (to appear)*, 2002.