Cluster Canonical Correlation Analysis

Nikhil Rasiwasia* **Dhruv Mahajan†** **Vijay Mahadevan*** **Gaurav Aggarwal***
*Yahoo Labs Bangalore(nikux,vmahadev,gaggarwa@yahoo-inc.com), †Microsoft Research(dhrumaha@microsoft.com)

Abstract

In this paper we present *cluster canonical correlation analysis* (cluster-CCA) for joint dimensionality reduction of two sets of data points. Unlike the standard pairwise correspondence between the data points, in our problem each set is partitioned into multiple clusters or classes, where the class labels define correspondences between the sets. Cluster-CCA is able to learn discriminant low dimensional representations that maximize the correlation between the two sets while segregating the different classes on the learned space. Furthermore, we present a kernel extension, *kernel cluster canonical correlation analysis* (cluster-KCCA) that extends cluster-CCA to account for non-linear relationships. Cluster-(K)CCA is shown to be computationally efficient, the complexity being similar to standard (K)CCA. By means of experimental evaluation on benchmark datasets, cluster-(K)CCA is shown to achieve state of the art performance for cross-modal retrieval tasks.

1 Introduction

Joint dimensionality reduction techniques such as Canonical Correlation Analysis (CCA) [12], Partial Least Squares (PLS) [23], Bilinear Model [26], Cross-modal Factor Analysis (CFA) [19] etc. have become quite popular in recent years. These approaches differ from the standard dimensionality reduction techniques such as principal component analysis (PCA) [5] or linear discriminant analysis (LDA) [5], as the dimensionality reduction is performed simultaneously across two (or more) modalities[1]. Given a dataset with two *paired* modalities — where each data point in the first modality is paired with a data point in the second modality — these approaches learn a common low dimensional feature space where representation-specific details are discarded to yield the common underlying structure. Of these, CCA is fast becoming the de facto standard [19, 21, 22, 24, 27]. CCA has been applied to several multimedia problems, such as cross-modal retrieval [10, 22, 24] — retrieval of data points from a given modality in response to a query from a different modality, image segmentation [21], cross-lingual retrieval [27], etc. CCA has also been successfully kernelized to enable learning of non-linear relationships in [4, 9].

However, CCA requires paired modalities and can not be directly applied when either *multiple clusters* of points in a

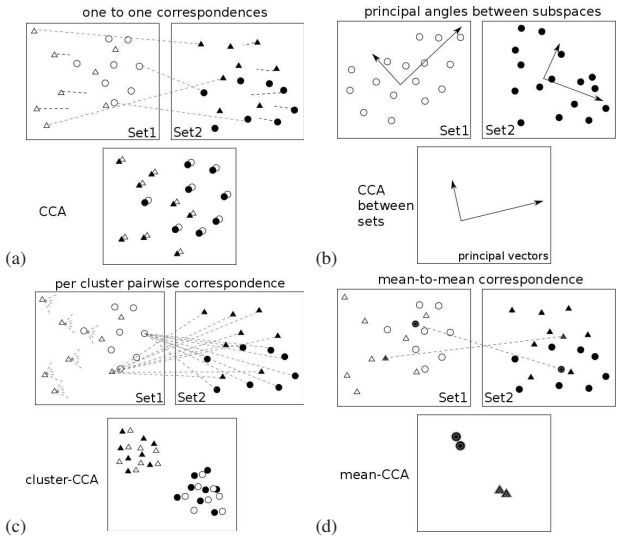


Figure 1: Representation of various methods to obtain correlated subspaces between sets. For each method, the two sets are shown at the top and the joint projected space at the bottom where △' and '◯' represent two clusters in each set. (a) CCA: uses pairwise correspondences between sets and cannot segregate the two clusters, (b) CCA for sets: computes principal angles between two subspaces and cannot handle multiple clusters, (c) cluster-CCA: uses all pairwise correspondences within a cluster across the two sets and results in cluster segregation and (d) mean-CCA: computes CCA between mean cluster vectors.

[1] It is common to refer to 'modalities' as 'sets' or 'views'.

Appearing in Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33.

given modality *correspond to multiple clusters* of points in another modality, or when the paired modalities are supplemented with *class labels*. Such a scenario arises when, for each class, several data points are available in two different modalities which may or may not be paired. For example, images and web-pages obtained using web search queries for different class labels. Note that for the case of paired modalities with class labels, CCA can still be applied ignoring the class labels, however, as shown in Fig. 1(a), CCA would be ineffective in achieving class discrimination.

In this work, we are interested in the above scenario, where each set consists of multiple clusters/classes. The correspondences between the sets are established by the class labels (the sets may or may not be paired). The aim is to learn a *discriminative common low dimensional* representation for the two sets. The contributions of this work are as follows:

- We propose a *very simple*, yet *effective* adaptation of CCA, referred to as *cluster canonical correlation analysis* (cluster-CCA). As shown in Fig. 1(c), in cluster-CCA *all* points from one modality within a class are paired with *all* points from the other modality in the same class and thereafter the projections are learned using the standard CCA framework.
- A *naive implementation* of cluster-CCA is computationally infeasible for large datasets, as the number of pairwise correspondences grows *quadratically with the number of data points per cluster*. We present a formulation of cluster-CCA that is computationally efficient and grows *linearly*.
- We also propose *mean-CCA*, a yet simpler adaptation of CCA for our task, where the mean vectors per cluster are used to learn the projections. We show that the fundamental difference between cluster-CCA and mean-CCA is in the estimation of the *within-set covariances*, which results in significant difference in their performance.
- Finally, we present a kernelized extension of cluster-CCA, referred to as *cluster kernel canonical correlation analysis* (cluster-KCCA) to extract non-linear relationships between modalities.

The efficacy of the proposed approaches is tested by measuring their performance in cross-modal retrieval tasks on benchmark datasets. The experiments show that cluster-(K)CCA is not only superior to (K)CCA but, *despite its simplicity*, outperforms other state-of-the art approaches that use class-labels to learn low dimensional projections. It is also shown that its performance can be improved by adding data to a single modality independently of the other modality, a benefit which is not shared by standard (K)CCA.

2 Related Work

Several extensions of CCA have been proposed in the literature [15–17, 21, 25, 28]. One class of modifications aims at using CCA for supervised dimensionality reduction [1, 11]. Given a set of samples with their class labels, CCA is used to learn a low dimensional representation. The data samples themselves serve as the first modality and the class labels as the second. Many variations in how the class labels are used have been proposed [17, 21, 25]. Nevertheless, the above approaches are targeted toward a single labeled modality, and cannot be directly applied for joint supervised dimensionality reduction of two labeled modalities.

CCA for matching two sets that are not paired, was proposed in [12]. Canonical vectors are obtained which minimize the principal angles — the minimal angles between vectors of two subspaces. Fig. 1(b) shows a simple schematic representation of 'CCA for sets'. CCA for sets has been applied to various problems [15, 16, 28], however it is only useful for the case where sets are unlabelled, e.g. to find canonical vectors for a given set of images and text where all the images and text belong to the same cluster. It cannot be directly applied to the case where there are multiple clusters in each set. CCA for sets was modified for classification of images into multiple classes in [16]. However, this approach too is applicable only to datasets consisting of a single labeled set.

Recently several approaches have been proposed for joint dimensionality reduction using class labels. *Semantic Matching* (SM) was proposed in [22], where two mappings are implemented using classifiers of the two modalities. Each modality is represented as vector of posterior probabilities with respect to the *same* class vocabulary, which serves as the common feature representation. Generalized Multivew Linear Discriminant Analysis (GMLDA) proposed in [24] formulates the problem of finding correlated subspaces as that of jointly optimizing covariance between sets and separating the classes in the respective feature spaces. The three objective functions are coupled linearly using suitable constants. Multi-view Discriminant Analysis (MvDA) proposed in [13] forgoes the free parameters by directly separating the classes in the joint feature space, but it is not clear how correlated the samples from different modalities are. Weakly-Paired Maximum Covariance Analysis (WMCA) proposed in [18], learns a correlated discriminative feature space without the need for pairwise correspondences. However, WMCA is based on maximum correlation analysis while the proposed work extends CCA for a similar problem setting.

3 Canonical Correlation Analysis (CCA)

In this section we briefly review CCA (for a more detailed introduction to CCA see [10]). Consider two multivariate

random variables $\mathbf{x} \in \Re^{D_x}$ and $\mathbf{y} \in \Re^{D_y}$ with zero mean. Let the sets $\mathcal{S}_x = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and $\mathcal{S}_y = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$, be paired. CCA aims at finding a new coordinate for $\mathbf{x}$ by choosing a direction $\mathbf{w} \in \Re^{D_x}$ and similarly for $\mathbf{y}$ by choosing a direction $\mathbf{v} \in \Re^{D_y}$, such that the correlation between the projection of $\mathcal{S}_x$ and $\mathcal{S}_y$ on $\mathbf{w}$ and $\mathbf{v}$ is maximized,

$$\rho = \max_{\mathbf{w},\mathbf{v}} \frac{\mathbf{w}' C_{xy} \mathbf{v}}{\sqrt{\mathbf{w}' C_{xx} \mathbf{w}} \sqrt{\mathbf{v}' C_{yy} \mathbf{v}}} \quad (1)$$

where ρ is the correlation, $C_{xx} = E[\mathbf{x}\mathbf{x}'] = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i{}'$ and $C_{yy} = E[\mathbf{y}\mathbf{y}'] = \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i \mathbf{y}_i{}'$ are the within-set covariance matrices and $C_{xy} = E[\mathbf{x}\mathbf{y}'] = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i \mathbf{y}_i{}'$ the between-set covariance matrix, E denoting empirical expectation. The problem can be reduced to a generalized eigenvalue problem [10], where $\mathbf{w}$ corresponds to the top eigenvector:

$$C_{xx}^{-1} C_{xy} C_{yy}^{-1} C_{yx} \mathbf{w} = \lambda^2 \mathbf{w} \quad (2)$$

The asymptotic time complexity for CCA is $O(nd^2) + O(d^3)$ where $d = max(D_x, D_y)$; $O(nd^2)$ for computing the covariance matrices and $O(d^3)$ for matrix multiplication, inverse and eigenvalue decomposition. Fig. 1(a) shows a simple schematic representation of CCA.

Kernel canonical correlation analysis (KCCA), reformulates CCA to extract non-linear relationships using the "kernel trick" [4, 9, 10],

$$\rho = \max_{\boldsymbol{\omega},\boldsymbol{\nu}} \frac{\boldsymbol{\omega}' K_x K_y \boldsymbol{\nu}}{\sqrt{\boldsymbol{\omega}' K_x^2 \boldsymbol{\omega}} \sqrt{\boldsymbol{\nu}' K_y^2 \boldsymbol{\nu}}} \quad (3)$$

where $K_{x(ij)} = k_x(\mathbf{x}_i, \mathbf{x}_j)$, $K_{y(ij)} = k_y(\mathbf{y}_i, \mathbf{y}_j)$ are the $n \times n$ kernel matrices, $k_x()$ and $k_y()$ the kernel functions, and $\boldsymbol{\omega} = (\omega_i, \ldots, \omega_n) \in \Re^n$, $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n) \in \Re^n$ the projection coefficients. The optimization problem of (3) can again be formulated as an eigenvalue problem. Note that both CCA and KCCA require paired modalities to learn the common low-dimensional representation.

4 Cluster Canonical Correlation Analysis

Consider two sets of data where each set is divided into C different but corresponding clusters/classes. Let $\mathcal{T}_x = \{\mathbf{X}_1, \ldots, \mathbf{X}_C\}$ and $\mathcal{T}_y = \{\mathbf{Y}_1, \ldots, \mathbf{Y}_C\}$, where $\mathbf{X}_c = \{\mathbf{x}_1^c, \ldots, \mathbf{x}_{|X_c|}^c\}$ and $\mathbf{Y}_c = \{\mathbf{y}_1^c, \ldots, \mathbf{y}_{|Y_c|}^c\}$ are the data points in the c^{th} cluster for the first and the second set respectively. Similar to CCA, the aim is to find a new coordinate for $\mathbf{x}$ by choosing a direction $\mathbf{w}$ and for $\mathbf{y}$ by choosing a direction $\mathbf{v}$, such that the correlation between the projections of $\mathcal{T}_x$ and $\mathcal{T}_y$ on $\mathbf{w}$ and $\mathbf{v}$ is maximized and simultaneously the clusters are well separated. However, unlike CCA, a direct correlation between these projections cannot be computed since the sets $\mathcal{T}_x$ and $\mathcal{T}_y$, and therefore their projections on $\mathbf{w}$ and $\mathbf{v}$ lack any direct correspondence. To address this, we propose two solutions viz. mean canonical correlation analysis (mean-CCA) and cluster canonical correlation analysis (cluster-CCA).

4.1 Mean Canonical Correlation Analysis

One simple solution is to establish correspondences between the mean cluster vectors of the two sets. This yields C vectors per set with one-to-one correspondences. Given the cluster means $\mu_x^c = \frac{1}{|X_c|}\sum_{j=1}^{|X_c|} \mathbf{x}_j^c$ and $\mu_y^c = \frac{1}{|Y_c|}\sum_{k=1}^{|Y_c|} \mathbf{y}_k^c$, mean canonical correlation analysis (mean-CCA) problem is formulated as,

$$\rho = \max_{\mathbf{w},\mathbf{v}} \frac{\mathbf{w}' V_{xy} \mathbf{v}}{\sqrt{\mathbf{w}' V_{xx} \mathbf{w}} \sqrt{\mathbf{v}' V_{yy} \mathbf{v}}} \quad (4)$$

where the covariance matrices V_{xy}, V_{xx} and V_{yy} are defined as:

$$V_{xy} = \frac{1}{C}\sum_{c=1}^{C} \mu_x^c \mu_y^{c\prime} \quad (5)$$

$$V_{xx} = \frac{1}{C}\sum_{c=1}^{C} \mu_x^c \mu_x^{c\prime} \quad (6)$$

$$V_{yy} = \frac{1}{C}\sum_{c=1}^{C} \mu_y^c \mu_y^{c\prime} \quad (7)$$

The asymptotic time complexity for mean-CCA is $O(Cd^2) + O(d^3)$. Fig. 1(d) shows a simple schematic representation of mean-CCA.

4.2 Cluster Canonical Correlation Analysis

In cluster canonical correlation analysis (cluster-CCA), instead of establishing correspondences between the cluster means, a one-to-one correspondence between all pairs of data points in a given cluster across the two sets is established and thereafter standard CCA is used to learn the projections. Fig. 1(c) shows a simple schematic representation of cluster-CCA. The cluster-CCA problem is formulated as,

$$\rho = \max_{\mathbf{w},\mathbf{v}} \frac{\mathbf{w}' \Sigma_{xy} \mathbf{v}}{\sqrt{\mathbf{w}' \Sigma_{xx} \mathbf{w}} \sqrt{\mathbf{v}' \Sigma_{yy} \mathbf{v}}} \quad (8)$$

where the covariance matrices Σ_{xy}, Σ_{xx} and Σ_{yy} are defined as:

$$\Sigma_{xy} = \frac{1}{M}\sum_{c=1}^{C}\sum_{j=1}^{|X_c|}\sum_{k=1}^{|Y_c|} \mathbf{x}_j^c \mathbf{y}_k^{c\prime} \quad (9)$$

$$\Sigma_{xx} = \frac{1}{M}\sum_{c=1}^{C}\sum_{j=1}^{|X_c|} |Y_c| \mathbf{x}_j^c \mathbf{x}_j^{c\prime} \quad (10)$$

$$\Sigma_{yy} = \frac{1}{M}\sum_{c=1}^{C}\sum_{k=1}^{|Y_c|} |X_c| \mathbf{y}_k^c \mathbf{y}_k^{c\prime} \quad (11)$$

where $M = \sum_{c=1}^{C} |X_c||Y_c|$, is the total number of pairwise correspondences. Similar to CCA, the optimization problem of (8) can be formulated as an eigenvalue problem. Note that, for both mean-CCA and cluster-CCA, we assume that covariance matrices are computed for zero mean random variables[2].

4.2.1 Computational Complexity

The asymptotic complexity of cluster-CCA is $O(Md^2) + O(d^3)$ where M is the number of pairwise correspondences and $d = \max(D_x, D_y)$; $O(Md^2)$ to compute the between set covariance matrix in (9) and $O(d^3)$ for matrix multiplication, inverse and eigenvalue decomposition. Since M grows quadratically with the number of examples per cluster (for example when $|X_c| = |Y_c| = L$, $M = CL^2$), cluster-CCA becomes computationally infeasible for large datasets. However (9) can be reformulated as the covariance matrix of the vector of cluster means,

$$\Sigma_{xy} = \frac{1}{M} \sum_{c=1}^{C} |X_c||Y_c| \mu_x^c \mu_y^{c\prime} \tag{12}$$

This reduces the computational complexity of cluster-CCA to $O(kd^2) + O(d^3)$ where $k = max(\sum_c |X_c|, \sum_c |Y_c|)$, thereby growing linearly with the number of data points. In summary, although a naive implementation of cluster-CCA can be computationally prohibitive for large datasets, the reformulation of (12) makes it efficient.

4.2.2 Cluster-CCA vs Mean-CCA

From the covariance matrices of various methods, two observations can be made. First, after the reformulation of (9) to (12), the between-set covariance matrix Σ_{xy} of cluster-CCA bears close resemblance to the between-set covariance matrix V_{xy} of mean-CCA (both being equal, modulo the normalization, when $|X_c|$ and $|Y_c|$ are class independent). Second, the within-set covariance matrices Σ_{xx} and Σ_{yy} of cluster-CCA in (10)-(11), bear close resemblance to the within-set covariance matrices C_{xx} and C_{yy} of CCA (again, both being equal, modulo the normalization, when $|X_c|$ and $|Y_c|$ are class independent).

This suggests that, the fundamental difference between mean-CCA and cluster-CCA is in the estimation of the *within set covariance matrices*. For mean-CCA they are estimated using the cluster means, effectively ignoring the rich information present in the data points themselves. For cluster-CCA, unlike mean-CCA, they they are estimated using all the data points as in CCA. As we shall see, this fundamental difference between mean-CCA and cluster-CCA, causes significant improvement in performance of cluster-CCA over mean-CCA.

[2]Thus in practice the sample means need to be adjusted accordingly.

4.3 Cluster Kernel Canonical Correlation Analysis

Similar to KCCA, cluster-CCA can also be extended to discover non-linear relationships between the two sets using non-linear projections of the data onto high dimensional spaces. Using the kernels functions $k_x()$ and $k_y()$, cluster kernel canonical correlation analysis (cluster-KCCA) can be formulated as,

$$\rho = \max_{\boldsymbol{\omega},\boldsymbol{\nu}} \frac{\boldsymbol{\omega}' R_{xy} \boldsymbol{\nu}}{\sqrt{\boldsymbol{\omega}' R_{xx} \boldsymbol{\omega}} \sqrt{\boldsymbol{\nu}' R_{yy} \boldsymbol{\nu}}} \tag{13}$$

$$R_{xy} = \frac{1}{M} \sum_{c=1}^{C} (\sum_{j=1}^{|X_c|} K_{xj}^c)(\sum_{k=1}^{|Y_c|} K_{yk}^c)' \tag{14}$$

$$R_{xx} = \frac{1}{M} \sum_{c=1}^{C} \sum_{j=1}^{|X_c|} |Y_c| K_{xj}^c K_{xj}^{c\prime} \tag{15}$$

$$R_{yy} = \frac{1}{M} \sum_{c=1}^{C} \sum_{k=1}^{|Y_c|} |X_c| K_{yk}^c K_{yk}^{c\prime} \tag{16}$$

where $K_{xj}^c = [\ \ldots,\ k_x(\mathbf{x}_j^c, \mathbf{x}_i),\ \ldots\ ]'$ and $K_{yk}^c = [\ \ldots,\ k_y(\mathbf{y}_k^c, \mathbf{y}_i),\ \ldots\ ]'$, where i indexes the data points in $\mathcal{T}_x$ and $\mathcal{T}_y$ for K_{xj}^c and K_{yk}^c respectively. Similar to KCCA, the optimization problem of (13) can be reformulated as an eigenvalue problem. The computational complexity of both KCCA and cluster-KCCA is $O(N^3)$.

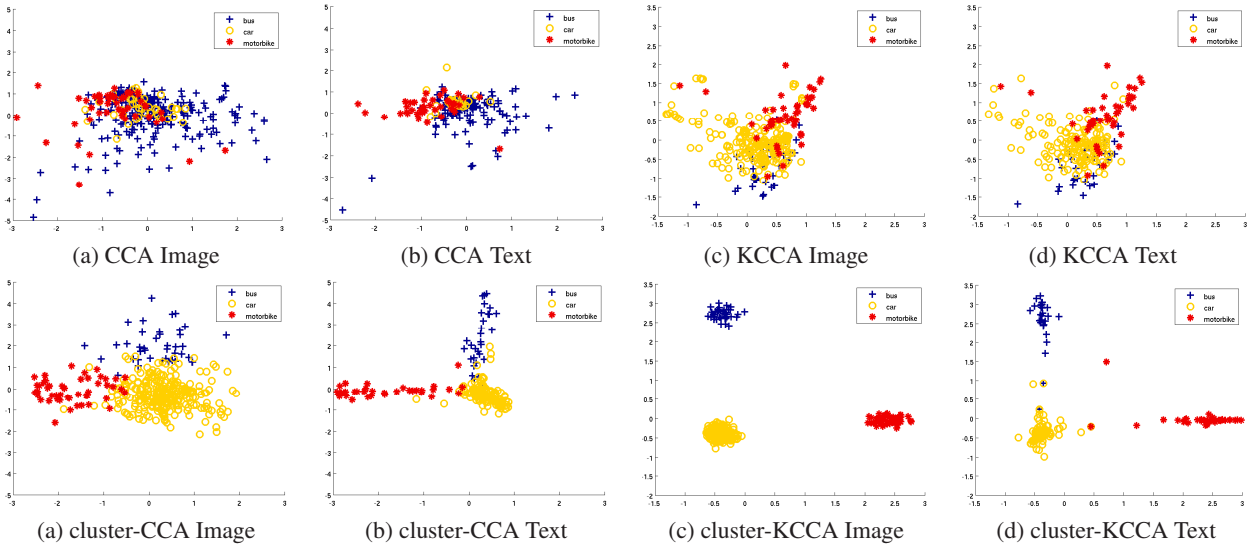
5 Experiments

In this section we present experimental evaluation for cluster-(K)CCA using cross-modal retrieval tasks. Precision-recall (PR) curves and mean average precision (MAP) scores are used for evaluation (except for the HFB dataset where rank-1 recognition is the standard). A retrieved item is considered to be correct if it belongs to the same class (cluster) as that of the query.

5.1 Datasets

The evaluation is conducted on five publicly available datasets, viz. Pascal VOC 2007 [6], TVGraz [14], Wiki Text-Image Dataset [22], Heterogeneous Face Biometrics (HFB) [20] and Materials Dataset [18].

Pascal dataset consists of $5011/4952$ train/test images and their annotations divided into 20 classes. The images were provided by the Pascal challenge [6] and the text annotations were collected in [8]. The image annotation serves as the text modality and is defined over a vocabulary of 804 keywords. We restrict our experiments to images and annotations which belong to a single class, reducing the train/test set to $2954/3192$. Of these, some of the annotations are empty, i.e., contain no keywords, thus we form two different datasets viz. *VOC* and *VOCfull*. In VOC, we remove all the images with empty annotations to maintain a



(a) CCA Image (b) CCA Text (c) KCCA Image (d) KCCA Text

(a) cluster-CCA Image (b) cluster-CCA Text (c) cluster-KCCA Image (d) cluster-KCCA Text

Figure 2: Low-dimensional mapping of images and text from 'bus', 'car' and 'motorbike' classes of Pascal VOC dataset. The top row shows the mappings obtained using CCA (left) and KCCA (right). The bottom rows shows the mapping obtained using the proposed cluster-CCA (left) and cluster-KCCA (right) for both images and text (left and right respectively for each scenario). Notice that cluster-(K)CCA is able to segregate different classes while mapping the two modalities to have high correlation.

balance between the number of image objects and text object (which also helps in comparing to other works). This yields a total of $1905/2032$ train/test images and annotations. In VOCfull we retain the annotation-less images for training. The test set in VOC and VOCfull is same, enabling us to evaluate the performance of cluster-CCA with respect to adding more data independently to one of the modalities.

TVGraz dataset was first compiled by Khan et. al. [14], where web pages were retrieved for ten classes of the Caltech-256 [7] dataset. Due to copyright issues, the TVGraz dataset is stored as a list of URLs and must be recompiled by each new user. We collected 2058 images and text from webpages (out of 2592 URLs), since some URLs were defunct. This set was randomly divided into $1558/500$ train/test.

Wiki dataset was compiled by Rasiwasia et. al. [22] using the featured articles from the Wikipedia website. It consists of $2173/693$ train/test images and text articles from 10 different classes.

HFB dataset consists of four near infrared (NIR) and four visual (VIS) images for each of the 100 subjects (classes) without any natural pairing. Note that both modalities are images but procured from different sensors. We follow Protocol II [13, 20] where images from 70 subjects are used to learn the projections and rest 30 subjects serve as the test set.

Materials consists of images as well as audio signatures from 17 different materials [18]. We present comparison with the published result for the classification task and also for cross-modal retrieval task.

In all datasets retrieval is performed on the test set, where each test set example from one modality is used to rank the test set examples of the other modality.

5.2 Features

For the VOC dataset we use the publicly available features of [8] (dense SIFT bag-of-words (BOW) for images and the raw BOW for text). For Wiki again we use the publicly available features of [22] (dense SIFT BOW for images 10-topic Latent Dirichlet Model (LDA) [3] for text). This helps us to compare directly with existing results. We also present results on a richer set of 200-topic LDA and 4096-codebook SIFT BOW, henceforth referred to as *WikiRich*. This enables us to present results for high-dimensional feature spaces. For TVGraz, a feature extraction procedure similar to Wiki was adopted (400-topic LDA for text and 4096-codebook SIFT BOW for images). For HFB, raw pixel data from the 32×32 cropped images is used without any further processing resulting in a 1024 dimensional feature space. For Materials dataset, we use the publicly available features of [18].

5.3 Experimental Protocol

We used regularized versions of (K)CCA for all our experiments, where the regularization constant is obtained using cross-validation. For kernelized approaches radial χ^2 kernel [2] is used for both text and images, where the normalization parameter is set to the mean of the χ^2 distances in the training set. Finally, normalized correlation score is used to compute similarities between the low-dimensional projected vectors. All experiments have been repeated ten times using random test-train splits.

6 Results

In this section we present the results for cluster-(K)CCA.

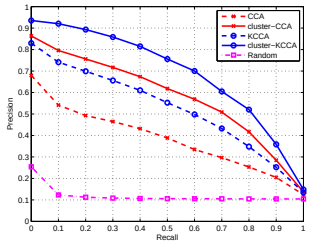
6.1 cluster-(K)CCA

In this section we show that cluster-(K)CCA results in a discriminative low-dimensional representation with high correlation between the projected sets. To demonstrate this, a toy dataset is constructed using 'bus', 'car' and 'motorbike' classes of the Pascal dataset. As shown in Fig. 2(left), a two dimensional mapping is computed by projecting the data points on the 1st and 2nd most correlated CCA (top) and cluster-CCA (bottom) components, for both text and images. It is clear from the figures that CCA, although yielding mappings with high correlation between text and images (0.79 and 0.72 for the first and second dimensions respectively), is not able to achieve class discrimination. Cluster-CCA achieves high correlation between text and images (0.67 and 0.62 for first two dimensions[3]) and simultaneously separates different classes into different regions of the space, supporting class discrimination. Fig. 2(right) shows similar mappings obtained using KCCA (top) and cluster-KCCA(bottom) where cluster-KCCA is again able to yield a better discriminative low-dimensional feature space.

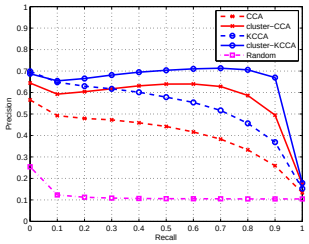
6.2 Cross-Modal Retrieval

In this section we presents the results of cluster-(K)CCA for cross-modal retrieval and compare it with (K)CCA. Table 1 presents the MAP performance of (K)CCA and cluster-(K)CCA on VOC, TVGraz, WikiRich, Wiki and HFB datasets. We also present results obtained using mean-(K)CCA and random chance performance. (K)CCA uses only pairwise correspondences and, mean-(K)CCA/cluster-(K)CCA use only the class labels. First, notice that mean-CCA is able to outperform CCA. Given that the number of pairwise correspondences in mean-CCA

[3]Correlation values for CCA and cluster-CCA are not directly comparable as for CCA the correlation is computed between pairs of text and images and for cluster-CCA , between all same class pairs of text and images.



(c) TVGraz-Text



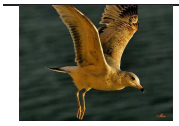
(d) TVGraz-Image

Figure 3: 11-point PR curves for (K)CCA and cluster-(K)CCA for TVGraz. The cluster versions outperform their respective standard algorithms at all levels of recall with cluster-KCCA achieving the best retrieval performance.

Table 1: Cross-modal Retrieval Performance.

Experiment	Image Query (MAP)	Text Query (MAP)	Dataset
random	0.0882	0.0882	VOC
CCA	0.263 ± 0.007	0.225 ± 0.005	
mean-CCA	0.290 ± 0.007	0.241 ± 0.009	
clusterCCA	**0.377 ± 0.007**	**0.335 ± 0.007**	
KCCA	0.326 ± 0.007	0.301 ± 0.007	
mean-KCCA	0.116 ± 0.005	0.186 ± 0.005	
clusterKCCA	**0.445 ± 0.006**	**0.429 ± 0.008**	
random	0.1191	0.1191	TVGraz
CCA	0.395 ± 0.013	0.359 ± 0.013	
mean-CCA	0.438 ± 0.021	0.411 ± 0.024	
clusterCCA	**0.554 ± 0.011**	**0.545 ± 0.014**	
KCCA	0.530 ± 0.014	0.511 ± 0.016	
mean-KCCA	0.194 ± 0.003	0.242 ± 0.006	
clusterKCCA	**0.612 ± 0.012**	**0.607 ± 0.011**	
random	0.1149	0.1149	WikiRich
CCA	0.281 ± 0.011	0.223 ± 0.010	
mean-CCA	0.303 ± 0.009	0.220 ± 0.009	
clusterCCA	**0.334 ± 0.011**	**0.250 ± 0.009**	
KCCA	0.263 ± 0.008	0.226 ± 0.008	
mean-KCCA	0.166 ± 0.013	0.186 ± 0.006	
clusterKCCA	**0.365 ± 0.008**	**0.288 ± 0.010**	
random	0.1191	0.1191	Wiki
CCA	0.252 ± 0.010	0.202 ± 0.008	
mean-CCA	0.246 ± 0.005	0.194 ± 0.005	
clusterCCA	**0.273 ± 0.008**	**0.218 ± 0.005**	
KCCA	0.269 ± 0.009	0.221 ± 0.009	
mean-KCCA	0.163 ± 0.010	0.164 ± 0.005	
clusterKCCA	**0.318 ± 0.010**	**0.249 ± 0.009**	
Experiment	NIR Query (Rank-1%)	VIS Query (Rank-1%)	Dataset
random	0.0333	0.0333	HFB
CCA	0.564 ± 0.085	0.583 ± 0.041	
mean-CCA	0.468 ± 0.044	0.456 ± 0.052	
clusterCCA	**0.627 ± 0.076**	**0.628 ± 0.089**	
KCCA	0.596 ± 0.084	0.597 ± 0.065	
mean-KCCA	0.497 ± 0.061	0.487 ± 0.058	
clusterKCCA	**0.632 ± 0.092**	**0.638 ± 0.070**	

{'350d, diesel, locomotive, railway, train'}, {'china, railroad, railway, steam, train'}, {'locomotive, rail, railroad, railway, train'}



{'bird, birds, impressedbeauty, specanimal'}, {'abigfave, birds, bokeh, florida, nature'}, {'bird, florida, pier'}

{'animal, cat, love, nature, pets, puppy'} {'cats, dogs'}, {'cat, cute, pet'}

{'island, sea, summer, yacht'}

Grace, of a family of elephants that researchers call the Virtues, touches the ailing Eleanor, the matriarch of the First Ladies family, who has fallen in Kenya's Samburu National Reserve on October 10, 2003. Grace will soon push Eleanor back to her feet, though the ailing elephant's resurgence will be short-lived. Elephants show compassionate behavior to others in distress, even to elephants not closely related to them, according to the researchers who produced these photos and an accompanying report published in the July 2006 issue of the journal Applied Animal Behaviour Science. Before this picture was taken, Eleanor, a new mother, had been found with a swollen trunk, abrasions to an ear and a leg, and a broken tusk probably from a previous fall. About two minutes after Eleanor had fallen, Grace rapidly approached. Her tail was raised and her temporal glands located on either side of the head between the eye and ear were excreting fluid. "The raised tail and the streaming temporal gland are typical signs of alarm and stress," said zoologist Iain Douglas-Hamilton, lead author of the study and founder of the nonprofit Save the Elephants.

On 31 January, the effort to retake the city began anew. The attack was launched at 08:30 hours, and was met by inaccurate Iraqi fire which knocked-out two Saudi V-150 wheeled vehicles. Stanton, claims that two vehicles were destroyed, while Westermeyer, claims that three were knocked-out. The 8th battalion of the Saudi brigade was ordered to deploy to the city by 10:00 hours, while 5th Battalion to the north engaged another column of Iraqi tanks attempting to reach the city. The latter engagement led to the destruction of around 13 Iraqi tanks and armored personnel carriers, and the capture of 6 more vehicles and 116 Iraqi soldiers, costing the Saudi battalion two dead and two wounded. The 8th Battalion engaged the city from the northeast, linking up with 7th Battalion. These units cleared the southern portion of the city, until 7th Battalion withdrew south to rest and rearm at 18:30 hours, while the 8th remained in Al-Khafji.Stanton, The 8th continued clearing buildings and by the time the 7th had withdrawn to the south, the Saudis had lost approximately 18 dead and 50 wounded, as well as seven V-150 vehicles. Coalition aircraft continued to provide heavy support throughout the day and night.Westermeyer, A veteran of the Iran-Iraq War later mentioned that Coalition airpower "imposed more damage on his brigade in half an hour than it had sustained in eight years of fighting against the Iranians."

Figure 4: Some examples of cross-modal retrieval. Top three rows show examples of image-to-text retrieval for images from the VOC dataset. The query image is shown in the first column and the retrieved documents (collection of tags) in the rest. The bottom three rows show examples of text-to-image retrieval, one each from the VOC, TVGraz and Wiki dataset.

is quite low (equal to the number of classes), the higher performance of mean-CCA over CCA, suggests that using class labels to learn the low-dimensional subspaces is beneficial for cross modal retrieval tasks. On the other hand, mean-KCCA is not able to achieve satisfactory performance. One reason for this could be that kernel based algorithms define the hyperplane as a linear combination of the data points, which being quite low for mean-KCCA, compromises its ability to learn good projections.

On comparing (K)CCA with cluster-(K)CCA, it is clear from the table that cluster-(K)CCA significantly outperforms (K)CCA. Cluster-CCA is able to achieve a significant performance gain of 45.90%, 45.88%, 15.87%, 7.93% and 9.23% over CCA on VOC, TVGraz, WikiRich, Wiki and HFB datasets respectively. The corresponding gains for cluster-KCCA over KCCA stand at 39.17%, 17.11%, 33.47%, 15.51% and 6.36%. Cluster-KCCA achieves the best retrieval performance, an average MAP score (over image and text query) of 0.4370, 0.6090, 0.3270, 0.2830 for VOC, TVGraz, WikiRich, Wiki and Rank-1 score of 0.635 for HFB dataset. Also note that the performance of cluster-(K)CCA is significantly superior to mean-(K)CCA. This highlights the importance of robust estimation of the covariance structure of the data, which is the fundamental difference between the two approaches.

Fig. 3 shows the precision recall curves for (K)CCA and cluster-(K)CCA for TVGraz. Cluster-(K)CCA outperforms (K)CCA at all levels of recall with cluster-KCCA achieving the best retrieval performance. Finally, Fig. 4 shows some examples of retrieval using cluster-KCCA. The top three rows shows examples of image to text retrieval, where images from the VOC dataset are used to retrieve annotation sets. The bottom three rows shows examples of text to image retrieval where a text document each from the three text-based datasets is used to retrieve images. As is evident from the figure, the retrieved examples belong to the same general class as that of the query.

6.3 Comparison with Existing Work

In this section we compare cluster-(K)CCA with existing works, where all approaches use class labels to learn the common low-dimensional space. In particular we present a comparison with SM of [22] — where two isomorphic spaces are learned using image/text classifiers, GMLDA of [24] — where a joint objective function is defined for maximizing correlation and class discrimination [4], MvDA of [13] — where LDA is performed on the low dimensional feature space and WMCA [18] — where maximum covariance analysis is modified for cluster-wise correspondences. To make the comparison fair, we also reimplemented the SM using kernel support vector machines as the classifier

[4]Published results of GMLDA [24] on the VOC dataset are on different set of features as used in this work.

with radial χ^2 kernel instead of logistic regression, referred to as SM(χ^2).

Table 2 presents the retrieval performance obtained with different approaches on various datasets. *Despite its simplicity*, cluster-(K)CCA outperforms other more involved approaches to achieve state-of-art performance on *all* datasets, indicating that cluster-(K)CCA can reliably learn discriminative common low-dimensional representations in diverse settings. Table 2 also presents comparisons to WMCA [18] on the Materials dataset. To compare we conduct two experiments. First, classification using nearest neighbors as evaluated in [18]. Cluster-CCA improves the classification error from 5.3% to 4.2%. Second, using the cross-modal retrieval task. Using WMCA features yields an MAP of 0.567/0.791 for the two reciprocal cross-modal tasks. The corresponding numbers using cluster-CCA features are 0.827/0.881. Thus, for both these experiments, cluster-CCA outperforms WMCA. The performance is significantly better for the cross-modal tasks (relative gain of 45%/11%).

Table 2: Comparison with Existing Works.

Experiment	Image (MAP)	Text (MAP)	Dataset
GMLDA [24]	0.427	0.339	VOC
SM(χ^2)	0.426 $\pm$ 0.009	0.403 $\pm$ 0.009	
clusterKCCA	**0.445 $\pm$ 0.006**	**0.429 $\pm$ 0.008**	
SM(χ^2)	**0.651 $\pm$ 0.018**	**0.647 $\pm$ 0.018**	TVGraz
clusterKCCA	0.612 $\pm$ 0.012	0.607 $\pm$ 0.011	
SM(χ^2)	0.358 $\pm$ 0.010	0.278 $\pm$ 0.009	WikiRich
clusterKCCA	**0.365 $\pm$ 0.008**	**0.288 $\pm$ 0.010**	
SM [22]	0.225	0.223	Wiki
GMLDA [24]	0.272	0.232	
SM(χ^2)	0.294 $\pm$ 0.007	0.233 $\pm$ 0.009	
clusterKCCA	**0.318 $\pm$ 0.010**	**0.249 $\pm$ 0.009**	
	NIR (Rank-1%)	VIS (Rank-1%)	HFB
SM(χ^2)	0.339 $\pm$ 0.063	0.368 $\pm$ 0.050	
MvDA [13]	50.0	53.3	
clusterKCCA	**0.632 $\pm$ 0.092**	**0.638 $\pm$ 0.070**	
	Audio (MAP)	Image (MAP)	Materials
WMCA [18]	0.567$\pm$0.020	0.791$\pm$0.037	
clusterCCA	**0.827$\pm$0.023**	**0.881$\pm$0.040**	
	Classification Error %	-	
WMCA [18]	5.3$\pm$2.3	-	
clusterCCA	**4.2$\pm$2.4**	-	

6.4 Effect of Additional Uni-modal Data

As discussed in 5.1, the Pascal VOC dataset contains some images which are not annotated. In CCA these images cannot be used for learning the low-dimensional subspaces as the corresponding text data is absent. However, cluster-(K)CCA does not require pair-wise correspondences and any additional data in one modality can potentially help to improve the retrieval performance. To test this, we conduct experiments using the VOCfull dataset (where all images, with or without annotations, are used to learn the low-dimensional subspaces). Table 3 presents the retrieval results and a comparison to the VOC dataset. Note that the test set is the same for both the datasets, thus any gain in performance is due to the (1049) additional images of the VOCfull dataset. The table shows that using additional images, without any additional text, improves the retrieval performance. Similar behavior is observed in other datasets where on increasing (decreasing) data for any single modality increases (decreases) the retrieval performance.

Table 3: Effect of Additional Unimodal Data (MAP).

Experiment	Image Query	Text Query	Average	Dataset
cluster-CCA	0.3672	0.3346	0.3509	VOC
cluster-CCA	**0.3802**	**0.3442**	**0.3622**	VOCfull
cluster-KCCA	0.4287	0.4162	0.4245	VOC
cluster-KCCA	**0.4446**	**0.4355**	**0.4401**	VOCfull

7 Conclusion

In this work, we proposed *cluster-CCA* for joint dimensionality reduction across two sets of variables, where each set is divided into multiple clusters. Further, we proposed a kernelized version of cluster-CCA, *cluster-KCCA* that is able to incorporate non-linear relationships between the two sets. Cluster-(K)CCA works on the principle of establishing correspondences between all pairs of data point in a given cluster across the two sets. It was shown that by doing so, cluster-(K)CCA is simultaneously able to achieve cluster segregation and high correlation between the sets. Cluster-(K)CCA was also shown to be computationally efficient, having the same computational complexity as that of (K)CCA. Cluster-(K)CCA, *despite its simplicity*, achieves superior state of the art performance in cross-modal retrieval tasks on benchmark datasets. Finally it was shown that its performance can be improved by adding data independently to either of the sets, a benefit that is exclusive to cluster-(K)CCA and is not shared by (K)CCA.

References

[1] M. Bartlett. Further aspects of the theory of multiple regression. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 34, pages 33–40. Cambridge Univ Press, 1938. 2

[2] M. Blaschko and C. Lampert. Correlational spectral clustering. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. 6

[3] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003. 5

[4] N. Cristianini and J. Shawe-Taylor. *An introduction to support Vector Machines: and other kernel-based learning methods*. Cambridge Univ Pr, 2000. 1, 3

[5] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley and Sons, 2001. 1

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html. 4

[7] G. Griffin, A. Holub, and P. Perona. The caltech-256. Technical report, Caltech, 2006. 5

[8] M. Guillaumin, J. Verbeek, and C. Schmid. Multimodal semi-supervised learning for image classification. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 902 – 909, jun 2010. 4, 5

[9] D. Hardoon and J. Shawe-Taylor. Kcca for different level precision in content-based image retrieval. In *Proceedings 3rd International Workshop on Content-Based Multimedia Indexing*. Citeseer, 2003. 1, 3

[10] D. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: an overview with application to learningmethods. *Neural Computation*, 16(12):2639–2664, 2004. 1, 2, 3

[11] T. Hastie, A. Buja, and R. Tibshirani. Penalized discriminant analysis. *The Annals of Statistics*, 23(1):73–102, 1995. 2

[12] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936. 1, 2

[13] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen. Multi-view discriminant analysis. In *Computer Vision–ECCV 2012*, pages 808–821. Springer, 2012. 2, 5, 7, 8

[14] I. Khan, A. Saffari, and H. Bischof. Tvgraz: Multi-modal learning of object categories by combining textualand visual features. In *Proceedings 33rd Workshop of the Austrian Association for PatternRecognition*, 2009. 4, 5

[15] T. Kim, O. Arandjelovic, and R. Cipolla. Boosted manifold principal angles for image set-based recognition. *Pattern Recognition*, 40(9):2475–2484, 2007. 2

[16] T. Kim, J. Kittler, and R. Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):1005–1018, 2007. 2

[17] O. Kursun, E. Alpaydin, and O. Favorov. Canonical correlation analysis using within-class coupling. *Pattern Recognition Letters*, 32(2):134–144, 2011. 2

[18] C. H. Lampert and O. Krmer. Weakly-paired maximum covariance analysis for multimodal dimensionality reduction and transfer learning. In *Proceedings of the 11th European conference on Computer vision*, pages 566–579, 2010. 2, 4, 5, 7, 8

[19] D. Li, N. Dimitrova, M. Li, and I. Sethi. Multimedia content processing through cross-modal association. In *Proceedings of the eleventh ACM international conference on Multimedia*, pages 604–611. ACM, 2003. 1

[20] S. Z. Li, Z. Lei, and M. Ao. The HFB face database for heterogeneous face biometrics research. In *IEEE Conf. on Computer Vision and Pattern Recognition Workshops, 2009*, pages 1–8. 4, 5

[21] M. Loog, B. van Ginneken, and R. Duin. Dimensionality reduction of image features using the canonical contextual correlation projection. *Pattern recognition*, 38(12):2409–2418, 2005. 1, 2

[22] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G.and Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings 18th ACM International Conference on Multimedia*, 2010. 1, 2, 4, 5, 7, 8

[23] R. Rosipal and N. Krämer. Overview and recent advances in partial least squares. *Subspace, Latent Structure and Feature Selection*, pages 34–51, 2006. 1

[24] A. Sharma, A. Kumar, H. Daume, and D. Jacobs. Generalized multiview analysis: A discriminative latent space. *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, Providence, Rhode Island, 2012. 1, 2, 7, 8

[25] T. Sun and S. Chen. Class label versus sample label-based cca. *Applied Mathematics and computation*, 185(1):272–283, 2007. 2

[26] J. Tenenbaum and W. Freeman. Separating style and content with bilinear models. *Neural computation*, 12(6):1247–1283, 2000. 1

[27] A. Vinokourov, D. Hardoon, and J. Shawe-Taylor. Learning the semantics of multimedia content with application toweb image retrieval and classification. In *4th International Symposium on Independent ComponentAnalysis and Blind Source Separation*, 2003. 1

[28] O. Yamaguchi, K. Fukui, and K. Maeda. Face recognition using temporal image sequence. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 318–323. IEEE, 1998. 2