Correlational Spectral Clustering

Matthew B. Blaschko Christoph H. Lampert
Max Planck Institute for Biological Cybernetics
72076 Tübingen, Germany
{blaschko,chl}@tuebingen.mpg.de

Abstract

We present a new method for spectral clustering with paired data based on kernel canonical correlation analysis, called correlational spectral clustering*. Paired data are common in real world data sources, such as images with text captions. Traditional spectral clustering algorithms either assume that data can be represented by a single similarity measure, or by co-occurrence matrices that are then used in biclustering. In contrast, the proposed method uses separate similarity measures for each data representation, and allows for projection of previously unseen data that are only observed in one representation (*e.g. *images but not text). We show that this algorithm generalizes traditional spectral clustering algorithms and show consistent empirical improvement over spectral clustering on a variety of datasets of images with associated text.*

1. Introduction

Image categorization is often approached in a supervised setting. The image categories are selected by hand *a priori* and typically involve tens to hundreds of classes [12, 14]. Other approaches involve many human participants labeling objects in images, requiring processing with a language model to identify labels with the same semantic meaning due to misspellings, polysemy, closely related topics, multiple languages, *etc*. [26, 31]. To truly scale with the range of semantic visual information experienced in a typical collection of images, unsupervised or weakly supervised methods are required to leverage information sources that do not require extra human effort to generate. In this work, we propose to make use of correlations between the visual content of images and other sources of paired information, such as image captions or associated spatiotemporal cues from video sequences, in order to find clusters that are more closely related to the underlying semantics of the content.

A paired dataset is one in which the data are simultaneously represented in two (or more) different spaces. A common latent aspect relates the representations, which can be thought of as embeddings of an underlying object into the respective feature spaces (Figure 1). Paired datasets are common in practice due to different methods of measurement, which may have different associated costs (*e.g*. infrared and visual imagery), or the use of different media such as images, text, and video. We assume here that one representation, images, are always available, but only some portion of these images will have associated media. We will use the images with associated media for training, and will learn representations that allow for the projection of previously unobserved images without associated media.

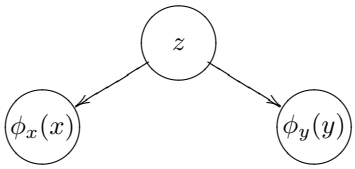


Figure 1. A paired dataset. A latent aspect z relates the observed values $\phi_x(x)$ and $\phi_y(y)$.

Specifically, we propose a generalization of spectral clustering based on kernel canonical correlation analysis that makes use of associated media at training time, but allows for projection of images without the associated media at test time. This is possible because kernel-CCA simultaneously learns linear projections from multiple spaces into a common latent space. In the kernelization of the algorithm, solutions are constrained to lie in the span of the projection of the training data, and projection is achieved by a linear combination of kernel evaluations between the training and test data.

Kernel-CCA generalizes Fisher linear discriminant analysis (LDA), which uses ground truth labels to find discriminant projections. Therefore, the additional modalities can be thought of as a weak form of labels. Because many sources of additional modalities are available, *e.g*. using text surrounding images on webpages, correlational spectral clustering allows for more accurate category learning without requiring expensive manual labels.

The rest of the paper is organized as follows: in Sec-

tion 2 we discuss related work in the area and its relation to the proposed method. In Section 3 we define correlational spectral clustering and give an overview of kernel canonical correlation analysis. We relate the proposed clustering algorithm to spectral clustering in Section 4.1 and show that the former is a generalization of the latter. In Section 4.2 we explore how canonical correlation analysis reduces the effect of noise, and in Section 5 we describe our experimental setup, datasets, and results. We analyze why the proposed method shows empirical improvement over spectral clustering in Section 6. Finally, we end with concluding remarks in Section 7.

Algorithm 1 Correlational Spectral Clustering

Require: $x_{train}, y_{train}, x_{test}, k_x(\cdot,\cdot), k_y(\cdot,\cdot)$
Ensure: c are the cluster ids assigned to the test data
Training:
$[K_x]_{i,j} = k_x(x_{train_i}, x_{train_j})$
$[K_y]_{i,j} = k_y(y_{train_i}, y_{train_j})$
α, β computed using KCCA on K_x and K_y
centroids $= k$-means$(K_x\alpha)$
Testing:
$c_j =$ the centroid nearest to $\sum_i \alpha_i k_x(x_{train_i}, x_{test_j})$

2. Related Work

A variety of methods have been proposed to model the relationship between images and text. Much of this work has been done in the context of finding associations between image content and individual words, noun phrases, or named entities [3, 7, 18, 19]. Blei and Jordan proposed *correspondence latent Dirichlet allocation* to model the joint distribution of images and text, and the conditional distribution of text given the image [8]. This has a natural application in automatic image annotation. Bekkerman and Jeon have recently proposed an image clustering algorithm based on a variation on combinatorial Markov random fields [5]. Additional modalities (*e.g*. text) are represented as nodes in the graph that are attached to the target modality (images), which is clustered using a local search to find an approximate solution to the combinatorial partitioning problem. Quattoni *et al*. devise a semi-supervised learning algorithm that exploits text captions to linearly constrain the visual representation to one that predicts well the presence or absence of individual words [1].

Another important set of approaches for clustering images with additional modalities belong to the family of spectral clustering algorithms [24, 28, 32]. Dhillon expressed the co-clustering problem in the framework of spectral clustering by considering bipartite graph structures where edge strengths are computed from co-occurrence matrices [11]. More recently, this has been extended from bipartite graphs to multipartite graphs in order to include additional modalities and has been applied to image and text data [13, 25]. Alternatively, one can build a matrix that combines similarities from both image and text representations [9, 22]. It is straightforward to then apply a standard spectral clustering technique [24, 28, 32]. Zhou and Burges combine the spectral clustering objectives for each of the modalities in order to trade off the costs of making a cut in each modality [33]. In contrast, our technique generalizes the family of spectral clustering algorithms to data with multiple modalities, but does not require any notion of co-occurrences between images and individual words, or similarities for both images and text in order to assign clusters to previously unseen data. Instead, correlational spectral clustering allows for the assignment of labels to unseen images that do not have associated text, and allows more flexibility in the representation used for the similarity matrices than is afforded by techniques built on co-occurrence matrices.

The proposed technique relies on kernel canonical correlation analysis to find projections of image representations that are correlated to the paired text. Kernel canonical correlation analysis has previously been employed with images and text in an image retrieval context [16], but has not been explored as a component of a clustering algorithm. Song *et al*. have considered the case of clustering with structured labels (*e.g*. hierarchical labels, ring structured data) by maximizing a norm of the cross-covariance operator between the projections of the input and the structure of the labels [29]. They have not, however, considered the case of multiple modalities or made use of the advantages of correlation rather than covariance.

3. Correlational Spectral Clustering

The clustering algorithm proposed in this paper, *correlational spectral clustering*, consists of kernel canonical correlation analysis computed with a training set followed by k-means in the projected space (Algorithm 1). At test time, the data are projected using linear combinations of kernel evaluations and assigned to the nearest cluster center. MatLab source code is available for download at `http://www.kyb.mpg.de/~blaschko`.

The subsequent sections give a brief introduction to kernel canonical correlation analysis and introduce notation.

3.1. Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) seeks to utilize paired datasets to simultaneously find projections from each feature space that maximize the correlation between the projected representations [17]. Given a sample from a paired dataset $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ we would like to simultaneously find directions w_x and w_y that maximize the correlation of the projections of x onto w_x with the projec-

tions of y onto w_y. This is expressed as

$$\max_{w_x,w_y} \frac{\hat{E}\left[\langle x, w_x\rangle\langle y, w_y\rangle\right]}{\sqrt{\hat{E}\left[\langle x, w_x\rangle^2\right]\hat{E}\left[\langle y, w_y\rangle^2\right]}}, \tag{1}$$

where $\hat{E}$ denotes the empirical expectation. We denote the covariance matrix of (x, y) by C and use the notation C_{xy} (C_{xx}) to denote the cross (auto) covariance matrices between x and y. Equation (1) is equivalent to

$$\max_{w_x,w_y} \frac{w_x^T C_{xy} w_y}{\sqrt{w_x^T C_{xx} w_x \; w_y^T C_{yy} w_y}}. \tag{2}$$

This Rayleigh quotient can be optimized as a generalized eigenvalue problem, or by decomposing the problem using the Schur complement as described in [16].

There is a natural extension of CCA in the event where there are more than two modalities. This can be written as a generalized eigenvector problem that subsumes two-way CCA as a special case

$$\begin{pmatrix} C_{11} & \dots & C_{1k} \\ \vdots & \ddots & \vdots \\ C_{k1} & \dots & C_{kk} \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ w_k \end{pmatrix} = \lambda \begin{pmatrix} C_{11} & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & C_{kk} \end{pmatrix} \begin{pmatrix} w_1 \\ \vdots \\ w_k \end{pmatrix}. \tag{3}$$

3.2. Kernel Canonical Correlation Analysis

We define kernels over x and y: $k_x(x_i, x_j) = \langle \phi_x(x_i), \phi_x(x_j)\rangle$ and $k_y(y_i, y_j) = \langle \phi_y(y_i), \phi_y(y_j)\rangle$. CCA is readily kernelized (KCCA) by searching for solutions that lie in the span of $\phi_x(x)$ and $\phi_y(y)$: $w_x = \sum_i \alpha_i \phi_x(x_i)$ and $w_y = \sum_i \beta_i \phi_y(y_i)$ [21]. Denoting the kernel matrices defined by our samples as K_x and K_y, we wish to optimize

$$\max_{\alpha,\beta} \frac{\alpha^T K_x K_y \beta}{\sqrt{\alpha^T K_x^2 \alpha \beta^T K_y^2 \beta}}. \tag{4}$$

As discussed in [16] this optimization leads to degenerate solutions in the case that either K_x or K_y is invertible so we maximize the following regularized expression

$$\frac{\alpha^T K_x K_y \beta}{\sqrt{\alpha^T \left((1-\tau_x)K_x^2 + \tau_x K_x\right)\alpha \beta^T \left((1-\tau_y)K_y^2 + \tau_y K_y\right)\beta}} \tag{5}$$

In the case that τ_x and τ_y are both set to 0, we have the same optimization as Equation (4). In the case that $\tau_x = \tau_y = 1$, this is equivalent to maximizing covariance instead of correlation.

The formulation of CCA in Equation (3) is also readily regularized and kernelized, and allows one to take advantage of additional modalities such as spatiotemporal features in video, higher resolution imagery, and other modalities that indirectly contain label information but are not necessarily available at test time.

4. Analysis of the Algorithm

4.1. Relation to Spectral Clustering

Spectral clustering algorithms make use of the spectrum of a similarity matrix to embed data into a vector space in a way that separates natural clusters in the data. After this projection, another technique such as k-means can be employed to determine the final labeling (see *e.g.* [32] for a tutorial introduction). Given a similarity matrix K, we define the unnormalized Laplacian $L \equiv D - K$ where D is a diagonal matrix that contains the row sums of K, and the normalized Laplacian $\mathcal{L} \equiv D^{-1}L$. The Laplacian eigenmap of $\mathcal{L}$ is defined as the embedding of the data that solves

$$\min_{\alpha,\alpha^T D\alpha=1} \alpha^T L \alpha = \min_{\alpha} \frac{\alpha^T L\alpha}{\alpha^T D\alpha} = \max_{\alpha} \frac{\alpha^T K\alpha}{\alpha^T D\alpha}. \tag{6}$$

This embedding, followed by k-means forms one of the main spectral clustering techniques [32]. It is also possible to recover this algorithm using kernel-PCA rather than the above generalized eigenvector problem with a kernel defined to be the negative commute distance on the graph defined by the similarity matrix K [15].

Kernel-PCA can be recovered with KCCA by setting $K_x = K_y$ and by setting τ_x and τ_y to 1. If K_x is set to the negative commute distance, we have recovered the above spectral clustering method. *Correlational spectral clustering therefore is a generalization of spectral clustering to the case of arbitrary kernels and paired data.*

4.2. A Latent Variable Interpretation

We can see why using paired data can be helpful in reducing the effects of noise by considering the covariance matrix of paired data with independent additive noise $\tilde{x} = x + \varepsilon$ and $\tilde{y} = y + \eta$. Their empirical covariance and cross-covariance matrices are

$$C_{\tilde{x}\tilde{x}} = C_{xx} + \underbrace{2C_{x\varepsilon} + C_{\varepsilon\varepsilon}}_{=:C_{xx}^{noise}}, \quad C_{\tilde{y}\tilde{y}} = C_{yy} + \underbrace{2C_{y\eta} + C_{\eta\eta}}_{=:C_{yy}^{noise}},$$
$$C_{\tilde{x}\tilde{y}} = C_{xy} + \underbrace{C_{x\eta} + C_{\varepsilon y} + C_{\varepsilon\eta}}_{=:C_{xy}^{noise}}. \tag{7}$$

In contrast to C_{xx}^{noise} and C_{yy}^{noise}, which contain the noise auto-covariances, C_{xy}^{noise} contains only cross-covariances of independent terms and will therefore be quite small. This shows that whenever there is paired data available, it makes sense to rely on the cross-covariance matrix, because this reduces the influence of noise in the data.

In the limit case of infinite data C_{xy}^{noise} will tend to zero. However, when dealing with finite sample sets, it can still have a spectrum that is large compared to that of C_{xy}. This is in particular the case for image data, where the noise consists not only of measurement errors, but also of varying

lighting conditions, changes in perspective *etc*. Text can contain irrelevant variances due to, *e.g*., misspellings and use of synonyms, or differences in morphology.

We can reduce this effect further by normalizing with the auto-covariance matrices. Making the noise contribution explicit in Equation (2), we obtain

$$\frac{w_x^T(C_{xy}+C_{xy}^{noise})w_y}{\sqrt{w_x^T(C_{xx}+C_{xx}^{noise})w_x\ w_y^T(C_{xx}+C_{xx}^{noise})w_y}}. \quad (8)$$

For projection directions w_x, w_y that are correlated only to the noise, the quotient will be dominated by $w_x^T C_{xy}^{noise} w_y / \sqrt{w_x^T C_{xx}^{noise} w_y w_x^T C_{yy}^{noise} w_y}$, which we know is close to 0 because C_{xy}^{noise} is much smaller than C_{xx}^{noise} and C_{yy}^{noise}. In contrast, in noise-free directions, the quotient becomes $w_x^T C_{xy} w_y / \sqrt{w_x^T C_{xx} w_y w_x^T C_{yy} w_y}$ which we can expect to be large for correlated signals x, y. This argument shows that the directions found by CCA are less influenced by noise than those found by maximizing cross-covariance.

Bach and Jordan have proposed a probabilistic interpretation of CCA that is analogous to a maximum likelihood interpretation of PCA [2]. We denote the dimensionalities of the vectors $\phi_x(x)$ and $\phi_y(y)$ as d_x and d_y, respectively, and interpret the diagram of a paired dataset in Figure 1 as a graphical model with parameters distributed

$$z \sim \mathcal{N}(0, I_d) \quad (9)$$

$$\phi_x(x)|z \sim \mathcal{N}(u_x z + \mu_x, \Psi_x) \quad (10)$$

$$\phi_y(y)|z \sim \mathcal{N}(u_y z + \mu_y, \Psi_y) \quad (11)$$

where $\min\{d_x, d_y\} \geq d \geq 1$ is the dimensionality of the projected output, $u_x \in \mathbb{R}^{d_x \times d}$ and $u_y \in \mathbb{R}^{d_y \times d}$ are parameters of the different modalities, and $\Psi_x \succeq \mathbf{0}$ and $\Psi_y \succeq \mathbf{0}$ are arbitrary noise covariance matrices. The maximum likelihood estimates of the parameters u_x and u_y are closely related to the first d canonical directions. Specifically, $\hat{u}_x = C_{xx} w_x \rho^{\frac{1}{2}} R$ and $\hat{u}_y = C_{yy} w_y \rho^{\frac{1}{2}} R$ where ρ is the diagonal matrix that contains the first d canonical correlations, and R is an arbitrary orthogonal matrix [2]. Because R is orthogonal, it does not affect the pairwise distances of the projection, and can be ignored. We see that the main difference between the canonical directions computed by CCA, w_x and w_y, and maximum likelihood estimators $\hat{u}_x$ and $\hat{u}_y$ is that the latter include the auto-covariance matrices C_{xx} and C_{yy}. We have argued above that the use of auto-covariance matrices is undesirable due to the potential effects of high noise variance that is not related to the underlying semantic problem. *Canonical correlation analysis computes directions that relate the two observations in a latent variable model that is derived from the generation of paired data, and that remove the influence of potentially irrelevant auto-covariance terms.*

5. Experimental Results

5.1. Evaluation Methodology

To evaluate the quality of the clustering, we have chosen paired datasets that contain images with associated text, as well as a human defined category label. We use the conditional entropy, $H(l|c)$, between the category labels, l, and the cluster ids, c, computed by the algorithm [10]. An important advantage of this evaluation is that we do not have to explicitly compute correspondences between class labels and cluster ids, which would involve searching through $\mathcal{O}(n!)$ possible assignments, where n is the number of classes. Conditional entropy is intimately related to mutual information

$$I(l;c) = H(l) - H(l|c). \quad (12)$$

Because $H(l)$ is fixed for a given dataset

$$\max_c I(l;c) = \min_c H(l|c) \quad (13)$$

and $H(l|c) \geq 0$ with equality only in the case that knowing the cluster id, c, allows one to compute the label, l, with certainty, i.e. the clusters are pure. Thus, for a fixed dataset and number of clusters, the clustering with the lowest conditional entropy score gives the clusters most related to the semantic grouping assigned by a human. Note, however, that conditional entropy scores are not comparable across different datasets.

We have used the following experimental protocol in all of the results reported here, unless explicitly indicated otherwise. The data are randomly split into equally sized train and test portions. The train portion is used to compute the projection and cluster-centroids using k-means, while the test portion is simply projected and assigned the cluster id of the nearest centroid in the projected space. In each training phase, k-means is trained 10 times with random initialization and the run with the smallest k-means objective is used. We compute the conditional entropy between the labels of the test set and the predicted cluster ids. The labels are never observed by the clustering algorithm, and the text annotations are only observed for the training portion of the dataset. The resulting conditional entropy scores are computed for 20 random splits of the data into train and test and visualized using a box plot [23].

5.2. Data

In order to demonstrate the broad applicability of correlational spectral clustering, we have done tests on a range of published datasets of images and text. We have used the Israeli-Images dataset described in [5] which consists of 1823 image-text pairs from 11 classes. We extracted SURF descriptors without rotation invariance and with the keypoint threshold set to 0 [4] and constructed a codebook

of 1000 visual words using k-means with 50000 sampled descriptors. Images were represented by a normalized histogram of these visual words. Additionally, we extracted HSV color histograms using 8 uniformly spaced bins for hue, 4 for saturation, and 2 for value, and represented each image by the normalized histogram. The histograms of visual words and of HSV colors were appended and the χ^2 kernel

$$k(x, x') = e^{-\frac{1}{2A}\sum_{i=1}^{d}\frac{(x_i - x'_i)^2}{x_i + x'_i}} \quad (14)$$

was used with normalization parameter A set to the mean of the χ^2 distances in the training set. Similarly, for text, we computed term frequency histograms, filtering special characters and stop words using the list from [30], and also used a χ^2 kernel.

Additionally, we have used the multimedia image-text web database used in [16, 20] which consists of samples from three classes: sports, aviation, and paintball, with 400 image-text pairs each. Images were represented using HSV color and Gabor textures as in [16, 20]. Text was represented using term frequencies. As in [16] we have used a Gaussian kernel for the image space, and a linear kernel for text.

Finally, we have used the three datasets included in the UIUC-ISD collection [22]. These consist of images collected from search engines using ambiguous search terms, "bass," "crane," and "squash," the webpages in which the images originally appeared, and an annotation of which sense of the word the image represents, *e.g.* fish vs. musical instrument. There are 2881 images in the Bass dataset which have been grouped into 6 categories, 2650 in the Crane dataset grouped into 9 categories, and 1948 images in the Squash dataset grouped into 6 categories. For all three datasets, we have represented images by 128 dimensional SURF features that have been vector quantized into 1000 bins using k-means on 50000 sampled features. For the text representation, we used word histograms extracted from the webpage title, removing special characters and stop words. Both image and text similarities were computed using a χ^2 kernel.

5.3. Parameter Selection

In our experiments we have used the implementation of KCCA described in [16], which makes use of Partial Gram-Schmidt Orthogonalization. As in [16] we fix the Gram-Schmidt precision parameter to 0.5 and have not optimized over this value. τ_x and τ_y are determined automatically by maximizing the ℓ_2 norm of the difference between the spectrum of correlations for randomized image and text associations, and the spectrum for the original unrandomized database (see [16] for details). The number of dimensions to project in KCCA has been set to the number of clusters, and the number of clusters has been set to the true number of classes. This last choice is chosen to avoid the comparison of algorithms that select different numbers of clusters; conditional entropy scores are not directly comparable in this case.

	PCA	CCA	KPCA	KCCA
Israeli	3.1318	3.0638	2.9722	**2.8046***
S.A.P.	0.9224	1.4699	0.8957	**0.8588**
Bass	2.2372	2.1880	2.1825	**2.1053***
Crane	2.6416	2.6297	2.5642	**2.5075***
Squash	2.3485	2.3452	2.2697	**2.2517**

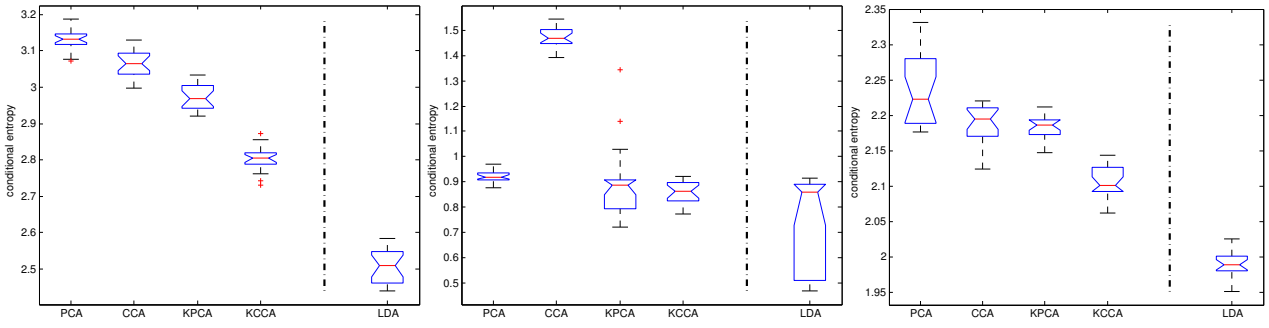
Table 1. Mean conditional entropy scores. Lower values indicate better clusters, and * indicates statistical significance. The proposed method, labeled KCCA, outperforms the other methods.

5.4. Results

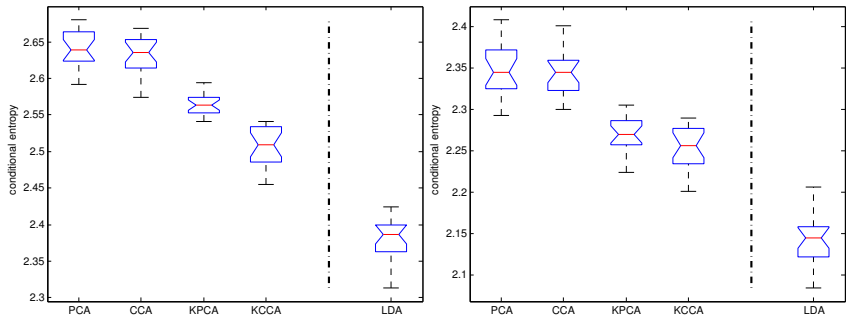
As baseline methods, we have selected linear PCA on image descriptors, kernel-PCA [27] on image descriptors, and CCA without kernelization. This gives an indication of the improvements that are gained by kernelization and by having text available at training time. Kernel-PCA can be viewed as a variant of spectral clustering that allows for the projection of unseen data, which allows us to compare in our experimental framework correlational spectral clustering to spectral clustering with only one modality [6]. Additionally, we have included results for KCCA experiments using the true labels for training. As discussed in [2] this is equivalent to Fisher linear discriminant analysis (LDA) in the case that $\tau_x = \tau_y = 0$. Using the labels at training time is not comparable to our previous results, but gives a form of upper bound on the improvement we could achieve using additional modalities. Figures 2(a)–2(e) give box plots of the conditional entropy scores for the five datasets described in Section 5.2, while Table 1 gives mean conditional entropy for the same experiments. We see that correlational spectral clustering (labeled KCCA) outperforms or is statistically tied with the previous methods for all datasets.

6. Discussion

Some clear patterns emerge from the plots in Figures 2(a)–2(e). Both applying linear CCA before clustering and kernelization of PCA tend to improve results over linear PCA, with the exception of the *Sports Aviation Paintball* dataset. In all datasets, correlational spectral clustering gave the best conditional entropy scores on average, with statistical significance in a majority of datasets. The LDA column of the figures indicates an upper bound of the improvement that is possible using correlational spectral clustering, since the second modality contains perfect information about the clustering task. We see that text provides a proxy for the labels; it informs the relevant directions with-



(a) Israeli Images (b) Sports, Aviation, Paintball (c) UIUC-ISD Bass



(d) UIUC-ISD Crane (e) UIUC-ISD Squash

Figure 2. Box plot results for each dataset. Conditional entropy scores are calculated across 20 runs of the various clustering algorithms. A lower score indicates better clusters. The proposed method, labeled KCCA, outperforms or is statistically tied with the previous methods for all datasets. The LDA column is shown separately because, unlike the other methods, it made use of the labels during training. See Section 5.4 for details.

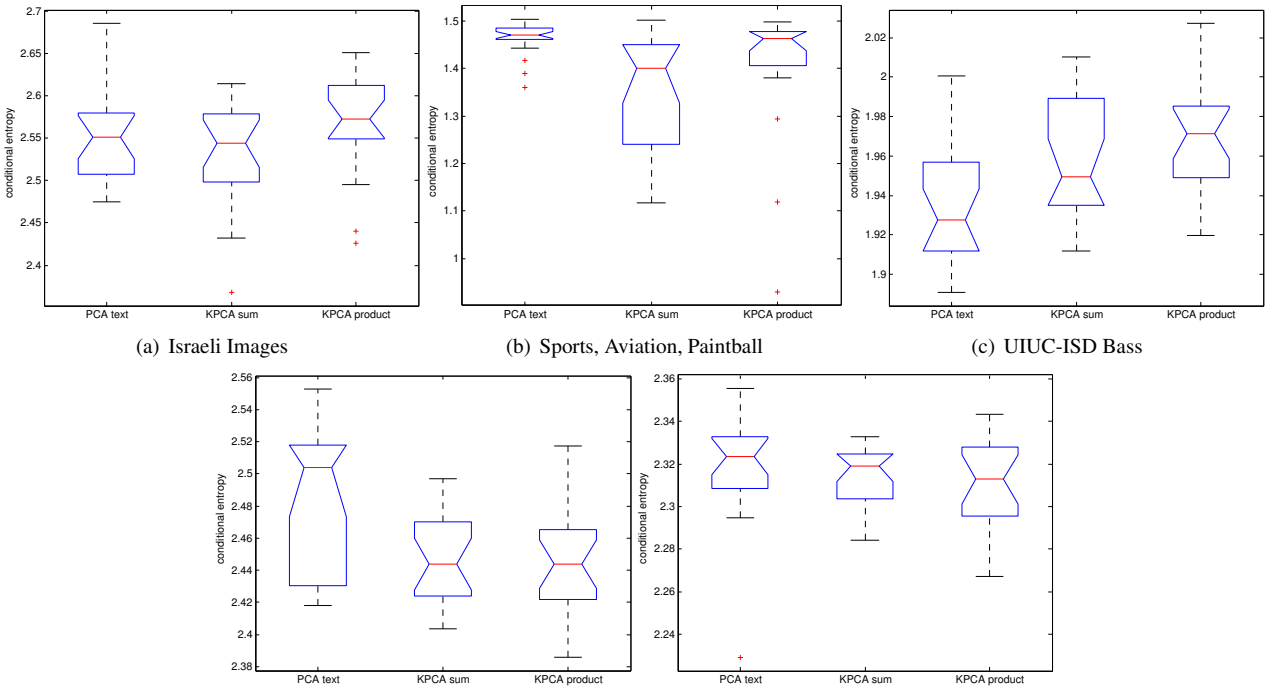
out having access to the labels directly. The improvement gained by having access to the labels at training time is, as expected, significantly better than that from text for the majority of datasets. This indicates that additional paired data could improve results further by using additional modalities as in Equation (3).

In the two datasets that did not show statistical significance, *Sports Aviation Paintball* and *UIUC-ISD Squash*, we also did not see an improvement with linear CCA. For the *Sports Aviation Paintball* dataset, we also did not see a statistically significant improvement of LDA over kernel PCA. It appears that for this dataset, the noise is low enough that the maximum variance directions in the image representation are already well suited to the clustering task, and there is no significant improvement to be had by searching for different directions.

To further understand the causes of the differences in performance between the different datasets, we have performed additional experiments to evaluate the amount of information present in the text component of the datasets. We have run experiments where the text was available not only at training time, but at test time as well. We have computed conditional entropy results for clustering after linear projections of the text using PCA, and for KPCA with kernels that combine the text and images using the sum of the two kernels $k_{sum}((x_i, y_i), (x_j, y_j)) = k_x(x_i, x_j) + k_y(y_i, y_j)$, and the product of the two kernels, $k_{product}((x_i, y_i), (x_j, y_j)) = k_x(x_i, x_j) \cdot k_y(y_i, y_j)$. Figure 3 shows box plots for the conditional entropy in this modified setting. We see that for the *Israeli Images*, *UIUC-ISD Bass*, and *UIUC-ISD Crane* datasets having text available at test time significantly improves performance over the setting where text is available only at training time (Figure 2). These are also the datasets where we have significant improvements from using correlational spectral clustering. Both the *Sports Aviation Paintball* and *UIUC-ISD Squash* datasets showed decreased performance when using the text representations, which indicates the text is not informative for the clustering task. Nevertheless, correlational spectral clustering was not adversely affected by the text as it ensures that the directions in the text are also correlated to a signal present in the images, which in these cases provided a more reliable cue.

7. Conclusions

We have proposed a new method, *correlational spectral clustering*, for clustering images given associated paired data, such as text or video information. This is achieved by finding non-linear projections of the images that are correlated with the associated data. Correlational spectral



(a) Israeli Images (b) Sports, Aviation, Paintball (c) UIUC-ISD Bass

(d) UIUC-ISD Crane (e) UIUC-ISD Squash

Figure 3. Box plot results for the text experiments. Conditional entropy scores are calculated for 20 runs of clustering using text data. The first column indicates projection with PCA only on the text representations. The second and third columns are for kernel PCA projections using the sum of the kernels for images and text, and the product of the kernels, respectively.

clustering generalizes spectral clustering to data with an arbitrary number of modalities. By examining the effect of using empirical covariance matrices on noise processes, and by employing a probabilistic interpretation of CCA, we have shown why correlational spectral clustering improves spectral clustering with one modality. We have shown statistically significant empirical improvement over traditional spectral clustering on a range of publicly available datasets.

Acknowledgements

The first author is supported by a Marie Curie fellowship under the EC funded project PerAct, EST 504321. This work is funded in part by the CLASS project, IST 027978. The authors would like to thank Ulrike von Luxburg and Arthur Gretton for helpful discussions.

References

[1] T. D. Ariadna Quattoni, Micheal Collins. Learning Visual Representations using Images with Captions. In *CVPR*, 2007.

[2] F. R. Bach and M. I. Jordan. A Probabilistic Interpretation of Canonical Correlation Analysis. Technical Report 688, Department of Statistics, University of California, Berkeley, 2005.

[3] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan. Matching Words and Pictures. *JMLR*, 3:1107–1135, 2003.

[4] H. Bay, T. Tuytelaars, and L. J. V. Gool. SURF: Speeded Up Robust Features. In *ECCV*, pages 404–417, 2006.

[5] R. Bekkerman and J. Jeon. Multi-modal Clustering for Multimedia Collections. In *CVPR*, 2007.

[6] Y. Bengio, O. Delalleau, N. Le Roux, J.-F. Paiement, P. Vincent, and M. Ouimet. Learning Eigenfunctions Links Spectral Embedding and Kernel PCA. *Neural Computation*, 16(10):2197–2219, 2004.

[7] T. L. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Y.-W. Teh, E. Learned-Miller, and D. A. Forsyth. Names and Faces in the News. In *CVPR*, pages 848–854, 2004.

[8] D. M. Blei and M. I. Jordan. Modeling Annotated Data. In *SIGIR*, pages 127–134, 2003.

[9] D. Cai, X. He, Z. Li, W.-Y. Ma, and J.-R. Wen. Hierarchical Clustering of WWW Image Search Results using Visual, Textual and Link Information. In *MULTIMEDIA*, pages 952–959, 2004.

[10] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, NY, USA, 1991.

[11] I. S. Dhillon. Co-clustering Documents and Words using Bipartite Spectral Graph Partitioning. In *KDD*, pages 269–274, 2001.

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object

Classes Challenge 2007 (VOC2007) . http://www.pascal-network.org/challenges/VOC/databases.html.

[13] B. Gao, T.-Y. Liu, T. Qin, X. Zheng, Q.-S. Cheng, and W.-Y. Ma. Web Image Clustering by Consistent Utilization of Visual Features and Surrounding Texts. In *MULTIMEDIA*, pages 112–121, 2005.

[14] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007.

[15] J. Ham, D. D. Lee, S. Mika, and B. Schölkopf. A Kernel View of the Dimensionality Reduction of Manifolds. In *ICML*, pages 369–376, 2004.

[16] D. R. Hardoon, S. Szedmák, and J. R. Shawe-Taylor. Canonical Correlation Analysis: An Overview with Application to Learning Methods. *Neural Computation*, 16(12):2639–2664, 2004.

[17] H. Hotelling. Relations Between Two Sets of Variates. *Biometrika*, 28:321–377, 1936.

[18] V. Jain, E. Learned-Miller, and A. McCallum. People-LDA: Anchoring Topics to People using Face Recognition. In *ICCV*, 2007.

[19] M. Jamieson, A. Fazly, S. Dickinson, S. Stevenson, and S. Wachsmuth. Learning Structured Appearance Models from Captioned Images of Cluttered Scenes. In *ICCV*, 2007.

[20] T. Kolenda, L. K. Hansen, J. Larsen, and O. Winther. Independent Component Analysis for Understanding Multimedia Content. In *IEEE Workshop on Neural Networks for Signal Processing*, pages 757–766, 2002.

[21] P. L. Lai and C. Fyfe. Kernel and Nonlinear Canonical Correlation Analysis. *IJNS*, 10(5):365–377, 2000.

[22] N. Loeff, C. O. Alm, and D. A. Forsyth. Discriminating Image Senses by Clustering with Multimodal Features. In *ACL*, 2006.

[23] R. McGill, J. W. Tukey, and W. A. Larsen. Variations of Boxplots. *The American Statistician*, 32:12–16, 1978.

[24] A. Y. Ng, M. I. Jordan, and Y. Weiss. On Spectral Clustering: Analysis and an Algorithm. In *NIPS*, pages 849–856, 2001.

[25] M. Rege, M. Dong, and J. Hua. Clustering Web Images with Multi-modal Features. In *MULTIMEDIA*, pages 317–320, 2007.

[26] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: A database and web-based tool for image annotation. Technical Report TR-2005-056, Massachusetts Institute of Technology, 2005.

[27] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. *Neural Computation*, 10(5):1299–1319, 1998.

[28] J. Shi and J. Malik. Normalized Cuts and Image Segmentation. *PAMI*, 22(8):888–905, 2000.

[29] L. Song, A. Smola, A. Gretton, and K. M. Borgwardt. A Dependence Maximization View of Clustering. In *ICML*, pages 815–822, 2007.

[30] C. J. van Rijsbergen. *Information Retrieval*. Butterworths, 1975.

[31] L. von Ahn and L. Dabbish. Labeling Images with a Computer Game. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 319–326, 2004.

[32] U. von Luxburg. A Tutorial on Spectral Clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[33] D. Zhou and C. J. C. Burges. Spectral Clustering and Transductive Learning with Multiple Views. In *ICML*, pages 1159–1166, 2007.