Weakly-Paired Maximum Covariance Analysis for Multimodal Dimensionality Reduction and Transfer Learning

Christoph H. Lampert[1] and Oliver Krömer[2]

[1] Institute of Science and Technology Austria, Klosterneuburg, Austria
[2] Max Planck Institute for Biological Cybernetics, Tübingen, Germany

Abstract. We study the problem of multimodal dimensionality reduction assuming that data samples can be missing at training time, and not all data modalities may be present at application time. *Maximum covariance analysis*, as a generalization of PCA, has many desirable properties, but its application to practical problems is limited by its need for perfectly paired data. We overcome this limitation by a latent variable approach that allows working with weakly paired data and is still able to efficiently process large datasets using standard numerical routines. The resulting *weakly paired maximum covariance analysis* often finds better representations than alternative methods, as we show in two exemplary tasks: texture discrimination and transfer learning.

1 Introduction

With the increasing availability of cheaper sensors, multimodal data has become nearly ubiquitous in practical computer vision tasks: images on the web have text captions, videos have audio tracks, and modern mobile phones can even record acceleration data in addition to their audio and visual recording capabilities. However, the field of multimodal data processing so far plays only a minor role in current computer vision research, where most algorithms are only able to process one data domain at a time. Those multimodal algorithms that do exist typically make restrictive assumptions, such as a priori known pairings between all data samples. They also commonly require that all sensor information is available reliably at all times, which is not always the case in practical problems because the use of multiple sensors increases the risk of subsystems failing.

In this paper, we introduce a dimensionality reduction method that can handle weakly paired data, and that is robust again the risk of partially missing data. Furthermore it incorporates two further advantages, which are of great importance for practical applications: it is simple, and it is efficient. By simplicity we mean that the method is based on elementary principles, in our case derived from statistics, which can be easily implemented and understood by a outsider of the field. An efficient method can be applied to data sets of realistic size, i.e. at least several thousand data vectors with thousands of dimensions.

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part II, LNCS 6312, pp. 566–579, 2010.

2 Multimodal Dimensionality Reduction

We assume that we are given related data samples in two or more data modalities of potentially very high dimension. The general goal of *multimodal dimensionality reduction* is to compute new representations for these data samples that lie in lower-dimensional feature spaces. In comparison to normal, unimodal, dimensionality reduction, we expect the availability of multiple data representations to give a better indication of what the true signal in the data is, that we want to retain, and what parts are noise that can be suppressed. As motivated in the introduction, we are interested in robust techniques that can handle missing examples in the original data. Additionally, once good dimensionality reduction mappings have been found, we want to be able to process each modality separately, in order to handle situations wherein some modalities are not always accessible. We formalize these intuitions in the following definitions.

Definition 1 (Inductive Dimensionality Reduction). *Let* $X = (x_1, \ldots, x_n) \subset \mathbb{R}^{d \times n}$ *be a set of data vectors. We call a procedure* ***inductive dimensionality reduction*** *if, given the input* X*, it outputs a functional mapping* $f : \mathbb{R}^d \to \mathbb{R}^q$ *with* $q < d$*. The image of* X *under* f *we call a* ***lower-dimensional representation*** *of* X *and denote it by* $\hat{X} = (\hat{x}_1, \ldots, \hat{x}_n)$*, i.e.* $\hat{x}_i = f(x_i)$*.*

In the rest of this paper, we will only consider inductive methods, which include *PCA* [25], *kernelPCA* [28] and *autoencoder networks* [10]. Non-inductive methods, e.g. *probabilistic latent semantic analysis (pLSA)* [11], and *Isomap* [30], also compute a lower-dimensional representation $\hat{X}$ from X, but do not provide a function f that could be applied to future data.

The two main families of inductive dimensionality reduction techniques, discriminative and generative, differ in the applications they are suitable for: discriminative techniques, such as *linear discriminant analysis (LDA)* [6] and *canonical correlation analysis (CCA)* [2,12], identify lower-dimensional representations that are suitable for a specific task that has to be known at the time of data processing, e.g. classification into a known set of classes. By discarding all signal dimensions that are not relevant for the specified task, discriminative techniques can often achieve a large reduction in dimensionality without loss of accuracy. Their drawback is that the representations found might not be well suited to tasks different from the specified one. In this work we concentrate on generative dimensionality reduction instead, where the goal is to find lower-dimensional data representations that are suited for various subsequent tasks, not just for a specific one. Intuitively, generative dimensionality reduction techniques can be seen as data compression methods, because it is often possible to recover the original data from the reduced representation with usually only a small reconstruction error.

Definition 2 (Multimodal Dimensionality Reduction) *Let* $X^{(1)} = (x_1^{(1)}, \ldots, x_{n_1}^{(1)}) \subset \mathbb{R}^{d_1 \times n_1}, \ldots, X^{(m)} = (x_1^{(m)}, \ldots, x_{n_m}^{(m)}) \subset \mathbb{R}^{d_m \times n_m}$ *be several data sets from potentially different spaces. We call an inductive dimensionality reduction technique* ***multimodal*** *if, given the inputs* $X^{(1)}, \ldots, X^{(m)}$*, it outputs functions* $f_1 : \mathbb{R}^{d_1} \to \mathbb{R}^q, \ldots, f_m : \mathbb{R}^{d_m} \to \mathbb{R}^q$ *for all data domains.*

Clearly, every inductive dimensionality reduction technique can in principle be used in a multimodal framework by just processing each data domain independently. However, since in the multimodal setup the functions f_i can depend on all data sets and not just on $X^{(i)}$ itself, one would expect multimodal techniques to use this information to find better representations than those of unimodal methods. The canonical way to construct multimodal algorithms is by making use of dependencies between the data samples that are induced by *pairings*:

Definition 3 (Weakly Paired Multimodal Data). *We call a collection of data sets $X^{(1)}, \ldots, X^{(m)}$ **weakly paired**, if each $X^{(i)}$ is split into k groups as*

$$X^{(i)} = (x^{(i)}_{1,1}, \ldots, x^{(i)}_{1,n^i_1}, \ldots, x^{(i)}_{k,1}, \ldots, x^{(i)}_{k,n^i_k}) \in \mathbb{R}^{d_i \times n_i} \tag{1}$$

*with $n_i = \sum_{l=1}^{k} n^i_l$. The special cases where $n^i_l = 1$ for all $i = 1, \ldots, m$ and $l = 1, \ldots, k$ we call **fully paired**. The other extremal case is $k = 1$, which we call the **unpaired** situation.*

Weakly paired data is common in multimodal data processing. For example, in video processing the groups could correspond to separate scenes for which we have data in the modalities: visual content, audio soundtrack, and textual subtitles. Unfortunately, existing techniques require fully paired data, which can introduce artificially overconstrained systems. In the above video example, one could pair each frame with the audio and subtitle content shown simultaneously with it. However, many of the correspondences introduced this way will be incorrect, as the synchronization between visual and other content is typically on a time scale much larger than the individual frame label.

3 Weakly Paired Maximum Covariance Analysis

In this section we derive a method for inductive multimodal dimensionality reduction with weakly paired data that we call *weakly paired maximum covariance analysis (WMCA)*. It can handle weakly paired and even unpaired data, because it infers suitable pairings directly from the data instead of requiring them a priori. This makes WMCA robust against missing data and enables it to process datasets where the domains have different numbers of samples, whereas previous techniques only worked if $n_1 = \cdots = n_m$ and the data was fully paired.

3.1 Linear Weakly Paired Covariance Maximization

We first study linear multimodal dimensionality reduction, and in order to simplify the notation we restrict the discussion to two modalities $X \in \mathbb{R}^{d \times n}$ and $X' \in \mathbb{R}^{d' \times n'}$. We will discuss the non-linear case in Section 3.2, and the extension to more than two modalities in Section 3.3.

In linear dimensionality reduction the dimensionality reduction functions can be written as $f(x) = W^t x$ for a matrix $W \in \mathbb{R}^{d \times q}$, and $f'(x') = W'^t x'$ for a matrix $W' \in \mathbb{R}^{d' \times q'}$. The lower dimensional representations are thus $\hat{X} = W^t X$

and $\hat{X}' = W'^t X'$. Typically, W and W' are assumed orthogonal matrices, so they contain the basis vectors of the linear subspaces of $\mathbb{R}^d$ and $\mathbb{R}^{d'}$ to be retained.

The most popular technique for generative linear dimensionality reduction is *principal component analysis (PCA)*. PCA finds a lower-dimensional representation that retains as much of the original signal's variance as possible. PCA can also be used to process fully paired multimodal data (by stacking the data vectors), but this does not qualify as a multimodal technique in the sense of Definition 2, since the construction requires that all modalities are also present in future data. The truly multimodal counterpart to PCA is *maximum covariance analysis (MCA)* [31], which would be ideal for our purposes, except that it also requires fully paired data.

Definition 4 (Maximum Covariance Analysis). *Let X and X' be fully paired datasets, i.e. for $X = (x_1, \ldots, x_n)$ and $X' = (x'_1, \ldots, x'_n)$ there is a pairing between each x_i and x'_i. Let X and X' be* centered*, i.e. $\frac{1}{n}\sum_{i=1}^n x_i = 0$ and $\frac{1}{n}\sum_{i=1}^n x'_i = 0$. **Maximum covariance analysis (MCA)** performs multimodal dimensionality reduction with projection matrices W, W' that solve*

$$\max_{W,W'} \operatorname{tr}\left[W^t X X'^t W'\right] \tag{2}$$

where the maximization runs over all orthogonal $d \times q$ and $d' \times q$ matrices.

Note that the condition of centered data is not severe, as we can center every dataset by subtracting the data mean from all samples.

MCA gets its name from the fact that the objective function (2) measures the total covariance between the individual dimensions of $\hat{X} = W^t X$ and $\hat{X}' = W'^t X'$, as one can see from rewriting $\operatorname{tr}[W^t X X'^t W'] = \sum_{p=1}^q [W^t X]_p^t [W'^t X']_p$ where $[.]_p$ indicates the p-th column.

Even though MCA is a strong method for multimodal dimensionality reduction, it has found relatively little application in computer vision contexts. We believe that the main reason for this is that MCA requires fully paired data, which realistic computer vision tasks often do not provide. In the rest of this section, we show how MCA can be extended to the weakly paired situation, calling the result *weakly paired maximum covariance analysis (WMCA)*.

Definition 5 (Weakly Paired Maximum Covariance Analysis). *Let X and X' be centered data sets that are weakly paired as specified in Definition 3. **Weakly paired maximum covariance analysis (WMCA)** performs multimodal dimensionality reduction with projection matrices W and W' that solve*

$$\max_{W,W',\Pi} \operatorname{tr}\left[W^t X \Pi X'^t W'\right], \tag{3}$$

where W and W' run over all orthogonal $d \times q$ matrices and $d' \times q$ matrices, respectively. Π runs over all $n \times n'$ pairing matrices *that respect the group structure of X and X', i.e. $\Pi = \operatorname{diag}(\Pi^1, \ldots, \Pi^k)$, where for $l = 1, \ldots, k$ we have $\Pi^l \in \{0,1\}^{n_l \times n'_l}$ such that $\sum_{i=1}^{n_l} \Pi^l_{i,j} \leq 1$ for all $j = 1, \ldots, n'_l$ and $\sum_{j=1}^{n'_l} \Pi^l_{i,j} \leq 1$ for all $i = 1, \ldots, n_l$.*

There is no single closed form solution to the optimization (3), as it requires both continuous optimization for W and W', and combinatoric optimization for Π. Furthermore, it is a high-dimensional non-convex problem, such that finding the global optimum in a numeric procedure is typically not possible. We can, however, efficiently find a locally optimal solution by *alternating maximization*:

- For known Π, solve

$$W, W' = \text{argmax}_{W,W'} \ \text{tr}\left[W^t X \Pi X'^t W'\right] \tag{4}$$

  Because Π is assumed to be known, the structure of this maximization is the same as when performing MCA with fully paired data. We obtain the basis vectors that form W and W' by computing the SVD of the matrix $X\Pi X'^t \in \mathbb{R}^{d \times d'}$, and keeping the q components in both domains with the largest singular values. When q is much smaller than d and d' (which is the typical case), we can use techniques for accelerated SVD computation, e.g. based on random projections [24]. This allows the efficient solution of Equation (4) even when d and d' are in the range of thousands or larger.
- For known W and W', solve

$$\Pi = \text{argmax}_{\Pi} \ \text{tr}\left[W^t X \Pi X'^t W'\right]. \tag{5}$$

  Given that $\text{tr}\left[W^t X \Pi X'^t W'\right] = \text{tr}\left[X'^t W' W^t X \Pi\right]$ and Π's special properties, the optimization (5) corresponds to a *linear assignment problem* with cost matrix $[X'^t W' W^t X]^t \in \mathbb{R}^{n \times n'}$. Furthermore, because of the diagonal block structure of Π, we can solve k separate problems of size $n_k \times n'_k$ instead of one big one of size $n \times n'$. Consequently, Equation (5) remains solvable in an efficient way even for large sample sizes, e.g. using the Hungarian algorithm [14] or LAPJV [13].

In both steps of the algorithm we maximize the same objective function. Therefore its value will increase monotonically over the iterations, which provides us with a natural stop criterion; we have reached a local maximum if the objective value does not increase any further.

To obtain a complete algorithm, we need a start value for Π. Unless some reasonable pairing is known a priori, we use $\Pi = \text{diag}(\Pi^1, \ldots, \Pi^k)$ with $\Pi^k \equiv \frac{1}{n_k n'_k}$. This is not a pairing matrix in the sense defined above, but it ensures that all data samples have influence on the initial choice of W and W'. The pairing property of Π will be established during the first solution of the maximization (5). As the alternating optimization algorithm is only locally convergent, it could also be run multiple times from different, e.g. random, start configurations. In our experiments, this did not lead to noticeable improvement, indicating that the above choice of Π is already a good heuristic.

3.2 Nonlinear Weakly Paired Covariance Maximization

Nonlinear dimensionality reduction techniques are often more powerful than linear ones, because they have more flexibility in the dimensionality reduction function that they output. MCA and WMCA can be made into non-linear techniques

by *kernelization*. As the necessary steps are very similar to, e.g., the derivation of kernelPCA from PCA we only outline them here, and refer the reader to [28] for a more detailed description of kernelization.

For kernelization, we require positive definite and symmetric similarity measures between samples, called kernel functions, that we denote by $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and $k' : \mathbb{R}^{d'} \times \mathbb{R}^{d'} \to \mathbb{R}$. Arguments from functional analysis show that any such kernel function corresponds to an inner product in a latent Hilbert space, and that it induces a latent feature map from the original data domain to this space [28]. Kernelized WMCA now consists of mapping the input data into the latent Hilbert spaces and performing linear WMCA on the resulting data sets. In the kernelized form, the optimization problem (3) becomes

$$\max_{A,A',\Pi} \ \mathrm{tr}\left[A\bar{K}\Pi\bar{K}'A'^t\right], \tag{6}$$

where $\bar{K}$ and $\bar{K}'$ are the centered kernel matrices. $\bar{K}$ is computed by forming the kernel matrix $K \in \mathbb{R}^{n\times n}$ as $[K]_{ij} = k(x_i, x_j)$ and then centering it using the formula $\bar{K} = K - \frac{1}{n}\mathbf{1}_n K - \frac{1}{n}K\mathbf{1}_n + \frac{1}{n^2}\mathbf{1}_n K\mathbf{1}_n$, where $\mathbf{1}_n$ denotes the $n \times n$ matrix in which all elements are 1. $\bar{K}'$ is computed from k' in the analogous way. Centering the kernels ensures that the implicitly defined feature vectors have zero mean in the latent feature space.

We solve the optimization problem (6) with the same alternating optimization scheme described previously with two differences:

- In contrast to W, W', the matrices $A \in \mathbb{R}^{n\times q}$ and $A' \in \mathbb{R}^{n'\times q}$ are not orthogonal. Instead they have to fulfill conditions $A^tKA = \mathrm{Id}$ and $A'^tK'A' = \mathrm{Id}$, which expresses orthogonality in the latent feature space. We obtain the rows of A and A' from a generalized eigenvalue problem:

$$\begin{pmatrix} 0 & K\Pi K' \\ K'\Pi^t K & 0 \end{pmatrix} \begin{pmatrix} a \\ a' \end{pmatrix} = \lambda \begin{pmatrix} K & 0 \\ 0 & K' \end{pmatrix} \begin{pmatrix} a \\ a' \end{pmatrix}. \tag{7}$$

  Solving Equation (7) is computationally more costly than solving (4). However, because we are interested only in the q eigenvectors of highest eigenvalue, we can still solve it efficiently using, e.g., the *power method* [7].
- When solving for A and A' in this way, the matrix $K\Pi K$ is of size $n \times n'$ instead of $d \times d'$. In the case where the number of data samples is smaller than the number of original data dimensions, it can be advantageous to use the kernelized formulation (6) also for the linear case. For this, one uses linear kernels $k(x, \tilde{x}) = x^t\tilde{x}$ and $k'(x', \tilde{x}') = x'^t\tilde{x}'$ and obtains the solutions of the problem (4) as $W = A^tX$ and $W' = A'^tX'$.

Kernelized WMCA provides reduction functions $f : \mathbb{R}^d \to \mathbb{R}^q$ and $f' : \mathbb{R}^{d'} \to \mathbb{R}^q$ by setting $f(x) = A^tK(x)$ with $K(x) = (k(x, x_1), \ldots, k(x, x_n))^t$ and $f'(x') = A'^tK'(x')$ with $K'(x') = (k'(x', x'_1), \ldots, k'(x', x_{n'}))^t$. Thus it is an inductive multimodal dimensionality reduction technique. Besides its flexibility to learn nonlinear projection functions, kernelization has another advantage. It allows us to process data sources that are provided in a different form than as vectors, e.g. text documents or graphs. In such scenarios, only a similarity measure, with the properties of a kernel function, needs to be defined to create Equation (6).

3.3 WMCA for More than Two Modalities

So far, we described WMCA for two data sources. An extension to more than two modalities is straightforward by reformulating the objective function as the sum of all pairwise covariances between all modalities. Thus, Equation (3) becomes

$$\max_{\substack{W^{(1)},\dots,W^{(m)} \\ \Pi^{(1,2)},\dots,\Pi^{(m-1,m)}}} \operatorname{tr}\Big[\sum_{i,j=1}^{m} W^{(i)} X^{(i)t} \Pi^{(i,j)} X^{(j)} W^{(j)t}\Big], \tag{8}$$

with the convention that $\Pi^{(i,i)} = 0$ and $\Pi^{(i,j)} = \Pi^{(j,i)t}$, and Equation (6) into

$$\max_{\substack{A^{(1)},\dots,A^{(m)} \\ \Pi^{(1,2)},\dots,\Pi^{(m-1,m)}}} \operatorname{tr}\Big[\sum_{i,j=1}^{m} A^{(i)} \bar{K}^{(i)t} \Pi^{(i,j)} \bar{K}^{(j)} A^{(j)t}\Big]. \tag{9}$$

Both systems can be solved by alternating maximization, where the step of finding the projection directions is solvable as an eigenvalue problem (generalized for the kernelized case), and finding the sample pairings requires solving $\frac{1}{2}m(m-1)$ linear assignment problems. Note that this quadratic scaling in the number of modalities does not pose a practical problems, since the majority of multimodal datasets utilize only a small number of modalities.

4 Related Work

As a classical dimensionality reduction technique, MCA comes from the same family of standard statistical methods as PCA, LDA and CCA. It also forms the basis for *partial least squares (PLS)* regression (PLS) [33]. Over the last 10 years, all of these techniques have been kernelized into non-linear versions [3,27,28]. The kernelization approach we take in Section 3.2 is similar to these, and the resulting expressions resemble the ones for *kernel canonical correlation analysis (kernelCCA)* [9]. KernelCCA also acts on multimodal data, but it would not have been a suitable basis for our purposes, as it is not generative. Furthermore, kernelCCA requires a priori setting of a regularization parameter for each modality, whereas, except for the number of output dimensions, MCA and WMCA are parameter-free. Nevertheless, CCA and kernelCCA are probably the most common methods for multimodal dimensionality reduction, typically in situations with a single fixed target application, e.g. *fMRI analysis* [8], *image clustering* [5], *speaker identification* [18], or *shape recovery* [16]. Alternative approaches include *multimodal pLSA* [17] or *Hilbert-Schmidt dependence maximization* [4], but these require a more careful experimental setup and are computationally more demanding. In contrast, the classical methods, and also WMCA, can be implemented with off-the-shelf components, typically just matrix operations.

To our knowledge, WMCA is the first multimodal dimensionality reduction technique that can efficiently handle weakly-paired data in the sense of Definition 3. The idea of treating unknown correspondences as latent variables and

optimizing over them, however, has been used in previous applications, including the classical k-means [20] algorithm, where one alternates between the centroid computation and the cluster assignment. An optimization similar to ours occurs in [4], which also alternates between a search for projection directions and for assignments. However in both cases the assignments are between sample and clusters, not between samples in different data modalities. WMCA's aspect of identifying relevant elements in groups of samples is somewhat related to witness approaches in *multiple instance learning* [1]. However, the criterion by which the elements are identified and the overall problem framework are very different.

5 Experimental Evaluation

In this section we show that due to its use of multimodal information, WMCA is often able to find low dimensional representations that reflect the information content of a data source better than a unimodal treatment of the same data. For this, we perform experiments on two realistic datasets: one for texture discrimination and one for transfer learning.

5.1 Texture Discrimination

As described in the introduction, generative dimensionality reduction aims at finding data representations that are suitable for different subsequent tasks. In this section we study this by performing texture discrimination both as an unsupervised and as a supervised learning problem. Note that both scenarios occur in real world scenarios. For example, in robot navigation it is important to classify surfaces into a set of known classes, such as *road* or *quick sand* (supervised). However, in order to collect probes in a new environment, the robot also needs to be capable of handling previously unobserved surface types, e.g. by grouping them based on their material properties (unsupervised).

To perform experiments on both setups we use a multimodal *Materials* dataset[1] that consists of images as well as audio signatures for 17 different materials (e.g. *bricks*, *styrofoam*, *wallpaper*, and *woven carpet*), see Figure 1. In contrast to available datasets with artificially constructed perfect pairings, the situation for this data is closer to the real problems that occur in multimodal data acquisition. The audio signal is recorded by dragging a small audio probe over the textured surfaces multiple times, and measuring the induced characteristic vibrations with a microphone. The images are captured using an ordinary digital camera. It is a priori unknown how a meaningful pairing should be constructed between the audio signals, which reflect a trajectory over the surface, and the rectangular regions depicted in the images. Also the conditions under which both modalities can be obtained differ: to capture images, one needs acceptable viewing conditions (e.g. no dust or fog). However, once this situation is established, each image contains a large amount of information from different physical locations. Audio recording in the described setup works by physical contact to

[1] The data set and source code are available at `http://www.ist.ac.at/~chl`

the material. The sensor can be shielded from environmental influences, but the information obtained is only very local.

We demonstrate how multimodal dimensionality reduction can be beneficial under such conditions by adopting an asymmetric multimodal setup: we use image and audio data to compute dimensionality reduction function, but we assume that only audio information is available at the time of application.

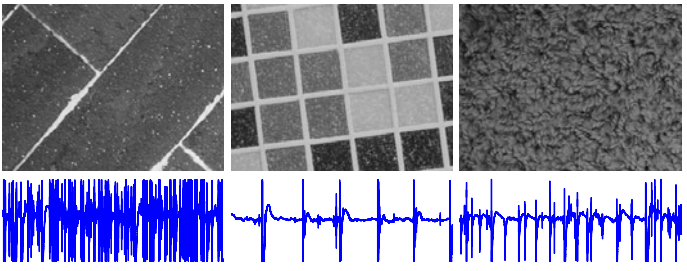

Fig. 1. Example images and audio signals from the multimodal *Materials* dataset

Data. The multimodal *Materials* dataset contains data from 26 textured plates made from 17 different material types. From each plate we recorded five audio signal with 44.1 kHz sampling frequency and segmented them into 450 overlapping sections of 50 ms, which we represented by phase and amplitude invariant *cepstral features* [19]. We clustered the resulting 58 500 feature vectors into an auditory codebook using k-means and represented each recording by a 1000-bin histogram, like in a bag-of-words representation. For the image data, we took high resolution photos with different in-plane rotations for a total of four to eight grayscale images per material. We computed *local binary patterns* over 8-neighborhoods considering only *uniform patterns* [21] such that any image region can be represented by a 58-dimensional histogram. Note that we intentionally chose a setup that is simple and easy to reproduce instead of a more powerful texture representation because our goal is not to improve the state of the art in texture classification but to examine the properties of multimodal feature extraction. To match the one-dimensional nature of the audio domain, we extracted single-pixel image strips with 16 pixel offset between them, resulting in a total of 32 histograms per image. For both, audio and visual data, we normalized each feature dimension to have zero mean and unit variance in order to reduce the influence of some histogram bins being more populated than others.

Experimental Setup. Our experiments reflect the situation where image and audio are present during dimensionality reduction itself, but only audio in the later application to new data. For this we split the data into two equally sized parts, called *context* and *task* data. We use WMCA to compute projection directions from the context data. As no perfect pairing between images and audio samples is available, we rely on the weak pairing information provided by the knowledge of which audio

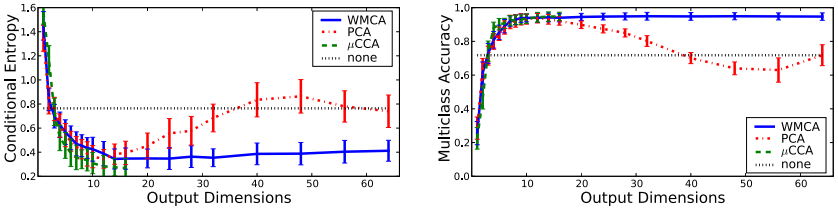

Fig. 2. Dimensionality reduction for unsupervised and supervised texture discrimination. The plots depict the conditional entropy (left, lower is better) and multi-class accuracy (right, higher is better) for different numbers of output dimensions.

signal was recording from which surface. In this linear bimodal case, each iteration of the WMCA algorithm takes only seconds. Convergence takes 2 to 50 iterations, depending on the output dimensionality.

We use the resulting dimensionality reduction functions to project the audio part of the task data to a new representation, and we measure the resulting clustering and classification performance. The unsupervised setup consists of applying *k-means* and measuring the quality of the resulting clusters by computing the *conditional entropy measure* [26,32] with respect to the ground truth. To simplify the setup we assume that the correct number of clusters is known a priori. In practical application, this number would have to be estimated from data. For the supervised setup, we measure the classification accuracy of a leave-one-out classifier; that is, for every point in the task set we determine its nearest neighbor and compute how often the labels of both samples coincide. For comparison we report the results of two baseline methods: unimodal dimensionality reduction with PCA that we apply separately to each modality, and fully-paired CCA, that is applicable when we use the data means of each weakly-paired group as input instead of the original samples (denoted μCCA). In addition we report the results without applying any dimensionality reduction.

Results. Figure 2 shows the results of the described procedure as mean and standard deviation over 100 random stratified splits of the data into context and task sets. We observe the same effect in both setups: all techniques identify the relevant output dimensions first and cause better results than when no dimensionality reduction is applied. However, when the number of output dimensions is increased, PCA starts to recover noise dimensions which decreases the performance, whereas WMCA's performance remains stable. Because μCCA uses the group means as inputs, it has only as many input samples as there are groups and therefore it cannot recover more than 17 output dimensions in this setup. In conclusion, the results of this section show that the main positive effect of using the multimodal dimensionality reduction in this case is improved noise suppression, which results in higher robustness in the choice of the number of output dimensions.

5.2 Transfer Learning

The previous experiments showed that WMCA is able to use multimodal data to infer which data dimensions are relevant and which are not. In this section we show how a similar effect can be used for *transfer learning* with attribute representations. Transfer learning consists of solving a learning task by making use of another, related, learning task, see [23] for a general overview and [22] for the specific case of transfer learning by dimensionality reduction. In our case, we want to improve the accuracy of an image classification system by making use of the data from another image classification task despite the fact that this has a disjoint set of classes and examples.

Data. For our experiments we use the *Animals with Attributes (AwA)*[2] dataset that has recently been introduced as a benchmark for attribute-based classification [15]. It consists of approximately 30,000 images of 50 animals classes as well as descriptions of the classes in terms of 85 binary semantic attributes, see Figure 3. The images are represented by the feature vectors that come with the dataset (based on SIFT, SURF, colorSIFT, local self similarity and color histogram features). We concatenate these into 10688-dimensional feature vectors and we remove the effect of inhomogeneous feature scaling by normalizing each dimension to zero mean and unit variance. The transformations necessary for this are saved in order to apply them to the task data later.

Experimental Setup. In our experiment largely follow the protocol of [15]. We split the set of classes into a context part consisting of forty classes and a task part consisting of ten classes. From the context data we chose 100 images per class, except for the *mole* category which has only 92 images that we use all, and we apply WMCA with the attribute representation as a second modality that is not available at test time. By assuming only a weak pairing between the domains, WMCA in particular is able to ignore outliers in the training set, whose actual image contents do not coincide well with the attribute vector. The quality of the resulting representation is determined by measuring the accuracy of a classifier for the task data. As baselines we again compute projection directions using PCA and CCA of the group means (μCCA). Because we assume that the context part has label information, we are able to also use LDA as a baseline. Additionally, we also include the case of not doing dimensionality reduction.

On the task set, we perform image classification in a Caltech-like setup. We randomly select a small number of training images per class, and classify a disjoint set of 30 randomly chosen test images using the nearest neighbor decision rule in the feature space induced by the projection directions found during the context stage. As in the case of texture discrimination our experimental setup is motivated by easy reproducibility. In particular we avoid free parameters that require model selection.

Results. Figure 4 shows the results for different numbers of training images and output dimensions as mean accuracy and standard error over 100 train/test

[2] Available for download at `http://attributes.kyb.tuebingen.mpg.de`

Fig. 3. Example images and attributes from the *Animals with Attributes* dataset

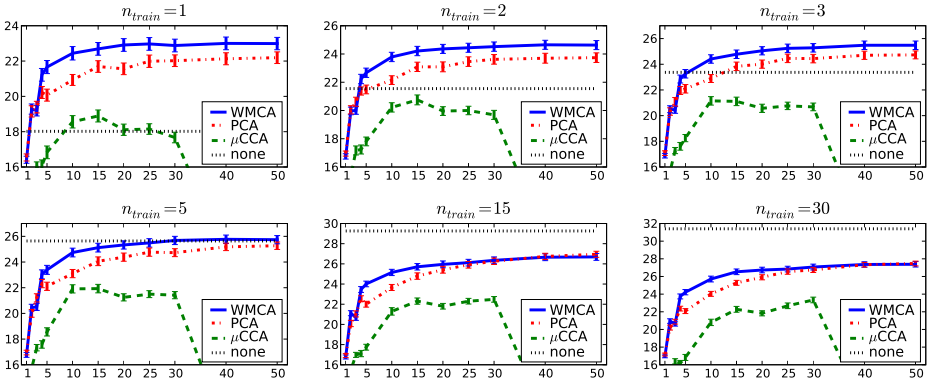

Fig. 4. Results of attribute-based transfer learning. The plots show the multi-class accuracy (y-axis) with n_{train} training images for different number of output dimensions (x-axis).

splits. When few training examples are available (top row), the representation found by WMCA leads to significantly higher classification accuracy than the representations obtained by PCA and also those by not using dimensionality reduction. When the number of training examples is increased WMCA is still superior to PCA when few output dimensions are wanted, but both are not able to exceed classification accuracy without dimensionality reduction anymore. This is consistent with the general observation that transfer learning works best in the regime when few training examples are available. However, dimensionality reduction can still be beneficial if runtime is an issue, as it makes the nearest neighbor lookup considerably faster than when the full features vectors are used.

μCCA leads to lower classification accuracy than both generative methods. Also, the performance does not improve any further when the number of output dimensions exceeds 10, which we interpret this as an overfitting effect. Because the data means provide only 40 data points, highly correlated directions can occur just due to noise effects. The plots in Figure 4 do not contain LDA, which never achieved classification accuracies that where significantly better than the chance level. The reason for this is LDA's discriminative objective. When applied to the context data it identifies projection directions that best encode the context class structure, but these do not reflect the class structure in the task set.

Overall, the results we achieve are comparable with previous work on the AwA dataset, which is known to be a difficult one. The most similar setup to ours is [29], where linear distance learning resulted in 23.7% accuracy in a one-shot setup, and a logistic representation in 27.2%. In [15], accuracies of 27.8% and 40.5% are reported, but based on a different test situation that made use of the attribute description at test time.

6 Conclusions

We have introduced weakly-paired maximum covariance analysis (WMCA) for multimodal dimensionality reduction. It overcomes the main limitation of MCA, from which it is derived, as it does not require fully paired data. Instead it treats missing pairings as latent variables which are inferred jointly with the projection directions. We showed how WMCA can be kernelized to perform non-linear dimensionality reduction. However, from a practical point of view, the most satisfactory setup is the linear two-modality case, where solving WMCA requires only two very efficient standard procedures: solving linear assignment problems and singular value decompositions.

In our experiments we illustrated two applications where multimodal dimensionality reduction was beneficial. In texture discrimination, WMCA produced more robust representations than the baselines. In transfer learning, when few training examples are available, WMCA was able to improve classification accuracy by transferring information from a context set to the main task.

Our initial experience with WMCA opens several directions for future work. Apart from practical application in robotics and video retrieval, we plan to derive more efficient techniques for applying kernelized WMCA at test time, e.g. based on reduced set methods and sparsification.

References

1. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: NIPS (2003)
2. Bach, F.R., Jordan, M.I.: A probabilistic interpretation of canonical correlation analysis. Technical Report 688, Department of Statistics, University of California, Berkeley (2005)
3. Baudat, G., Anouar, F.: Generalized discriminant analysis using a kernel approach. Neural Computation 12(10) (2000)
4. Blaschko, M., Gretton, A.: Learning taxonomies by dependence maximization. In: NIPS (2009)
5. Blaschko, M., Lampert, C.H.: Correlational spectral clustering. In: CVPR (2008)
6. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Annals of Eugenics 7 (1936)
7. Golub, G.H., Van Loan, C.F.: Matrix computations. Johns Hopkins Univ. Press, Baltinmore (1996)
8. Hardoon, D., Mourao-Miranda, J., Brammer, M., Shawe-Taylor, J.: Unsupervised analysis of fMRI data using kernel canonical correlation. NeuroImage 37(4) (2007)

9. Hardoon, D., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: an overview with application to learning methods. Neural Computation 16(12) (2004)
10. Hinton, G., Salakhutdinov, R.: Reducing the dimensionality of data with neural networks. Science 313(5786) (2006)
11. Hofmann, T.: Probabilistic latent semantic indexing. In: ACM SIGIR (1999)
12. Hotelling, H.: Relation between two sets of variates. Biometrika 28 (1936)
13. Jonker, R., Volgenant, A.: A shortest augmenting path algorithm for dense and sparse linear assignment problems. Computing 38(4) (1987)
14. Kuhn, H.W.: The hungarian method for the assignment problem. Naval Research Logistics Quarterly 2 (1955)
15. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: CVPR (2009)
16. Lei, Z., Bai, Q., He, R., Li, S.Z.: Face shape recovery from a single image using CCA mapping between tensor spaces. In: CVPR (2008)
17. Lienhart, R., Romberg, S., Hörster, E.: Multilayer pLSA for multimodal image retrieval. In: CIVR (2009)
18. Livescu, K., Stoehr, M.: Multi-view learning of acoustic features for speaker recognition. In: Automatic Speech Recognition and Understanding (2009)
19. Logan, B.: Mel frequency cepstral coefficients for music modeling. In: International Symposium on Music Information Retrieval (2000)
20. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: 5th Berkeley Symposium on Mathematics Statistics and Probability (1967)
21. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. PAMI 24(7) (2002)
22. Pan, S.J., Kwok, J.T., Yang, Q.: Transfer learning via dimensionality reduction. In: AAAI (2008)
23. Pan, S.J., Yang, Q.: A survey on transfer learning. Knowledge and Data Engineering (2009)
24. Papadimitriou, C.H., Raghavan, P., Tamaki, H., Vempala, S.: Latent semantic indexing: A probabilistic analysis. Computer and System Sciences 61(2) (2000)
25. Pearson, K.: On lines and planes of closest fit to systems of points in space. Philosophical Magazine Series 6. 2(11) (1901)
26. Rosenberg, A., Hirschberg, J.: V-measure: A conditional entropy-based external cluster evaluation measure. In: EMNLP-CoNLL (2007)
27. Rosipal, R., Trejo, L.J.: Kernel partial least squares regression in reproducing kernel Hilbert space. JMLR 2 (2002)
28. Schölkopf, B., Smola, A.J.: Learning with kernels. MIT Press, Cambridge (2002)
29. Tang, K., Tappen, M., Sukthankar, R., Lampert, C.H.: Optimizing one-shot recognition with micro-set learning. In: CVPR (2010)
30. Tenenbaum, J.B., Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science 290(5500) (2000)
31. Tucker, L.R.: An inter-battery method of factor analysis. Psychometrika 23 (1958)
32. Tuytelaars, T., Lampert, C.H., Blaschko, M.B., Buntine, W.: Unsupervised object discovery: A comparison. IJCV 88(2) (2010)
33. Wold, H.: Estimation of principal components and related models by iterative least squares. Multivariate Analysis 1 (1966)