Discriminative extended canonical correlation analysis for pattern set matching

Ognjen Arandjelović

Abstract In this paper we address the problem of matching sets of vectors embedded in the same input space. We propose an approach which is motivated by *canonical correlation analysis* (CCA), a statistical technique which has proven successful in a wide variety of pattern recognition problems. Like CCA when applied to the matching of sets, our *extended canonical correlation analysis* (E-CCA) aims to extract the most similar modes of variability within two sets. Our first major contribution is the formulation of a principled framework for robust inference of such modes from data in the presence of uncertainty associated with noise and sampling randomness. E-CCA retains the efficiency and closed form computability of CCA, but unlike it, does not possess free parameters which cannot be inferred directly from data (inherent data dimensionality, and the number of canonical correlations used for set similarity computation). Our second major contribution is to show that in contrast to CCA, E-CCA is readily adapted to match sets in a discriminative learning scheme which we call *discriminative extended canonical correlation analysis* (DE-CCA). Theoretical contributions of this paper are followed by an empirical evaluation of its premises on the task of face recognition from sets of rasterized appearance images. The results demonstrate that our approach, E-CCA, already outperforms both CCA and its quasi-discriminative counterpart constrained CCA (C-CCA), *for all values of their free parameters*. An even greater improvement is achieved with the discriminative variant, DE-CCA.

Keywords Set · Matching · Vectors · Principal · Angles

1 Introduction

Central to any applied problem of pattern recognition is the issue of how the entities of interest should be represented. A numerical description based on readily measurable quantities is sought, one which (as much as possible) minimizes variability due to confounding factors

O. Arandjelović (✉)
Centre for Pattern Recognition and Data Analytics, Deakin University, Geelong, VIC 3216, Australia
e-mail: ognjen.arandjelovic@gmail.com

and maximizes that due to differing class memberships. Clearly, this is a highly domain specific task. In computer vision, for example, photometric or geometric models may be used to normalize for illumination and viewpoint changes, or to recover the three-dimensional structure of the scene. After this explicit separation of relevant and confounding variables is performed, the problem ultimately becomes that of inferring class boundaries by matching patterns. In this paper we are specifically interested in set-to-set matching, that is, the case when multiple examples from each class are available both for training, and querying using unlabelled input.

Assembly approaches   The preferred approach to set comparison is inherently governed by the nature of the particular task to which it is applied. Thus, a large number of different set similarity measures have been proposed and successfully used in different recognition problems. Many of these are non-parametric methods, based on comparisons of individual members of sets. The simplest examples included the *minimum minimorum* (Satoh 2000) and *maximum minimorum* (Vivek and Sudha 2007) (or Hausdorff) distances which reduce inter-set distance to the distance between only a pair of their elements. Others aggregate member similarities over entire sets, or chosen representative subsets (Fan and Yeung 2006).

Probability density approaches   Stronger assumptions are made by methods which assume that different data sets corresponding to the same class are drawn from related probability distributions. These may be estimated using non-parametric, semi-parametric (Shakhnarovich et al. 2002) or parametric models, and the closeness between them quantified using, amongst many others, the Bhattacharyya (1943), Chernoff (1952) and resistor-average (Sinanović and Johnson 2007; Arandjelović and Cipolla 2006) distances, or an asymmetric similarity measure such as the Kullback-Leibler divergence (Kullback and Leibler 1951). A major shortcoming of probability density-based set matching is the underlying premise that statistical properties of a novel, unlabelled set and the corresponding training set are in some sense alike (Arandjelović and Cipolla 2013). Implicit in this is the assumption that novel and training data are acquired in similar conditions (such as the viewpoint and illumination) or very robustly normalized—both are conditions which are difficult to ensure in nearly all cases of interest.

Manifold approaches   While the conditions in which they are acquired can change the observed distribution of data samples, this variation is nonetheless constrained by the intrinsic properties of that class—often to a manifold, as illustrated in Fig. 1. This can be exploited by learning the structure of this manifold while disregarding higher order statistics along it (Lee et al. 2003). A wide range of manifold learning approaches has been described in the literature including multidimensional scaling (Borg and Groenen 2005), local topology preserving embedding (Roweis and Saul 2001), eigenspace (Gunturk et al. 2003) piece-wise linear approximation (Lee et al. 2003; Kim et al. 2007) and nonlinear unfolding using a Mercer kernel (Bach and Jordan 2002; Wolf and Shashua 2003; Melzer et al. 2003; Yang 2002; Fukumizu et al. 2007).

The other major issue is that of devising a suitable inter-manifold metric (or rather pseudo-metric). Of particular interest to us are approaches employing *canonical correlation analysis* (CCA), a statistical technique that has been applied with much success to a wide variety of problems, including 3D reconstruction (Reiter et al. 2006), infrared to visual image conversion (Dou et al. 2007), recognition of texture (Saisan et al. 2001), objects (Wolf and Shashua 2003; Melzer et al. 2003) and speech (Choukri and Chollet 1986). Broadly speaking, the key idea behind CCA is that a meaningful measure of similarity between two linear
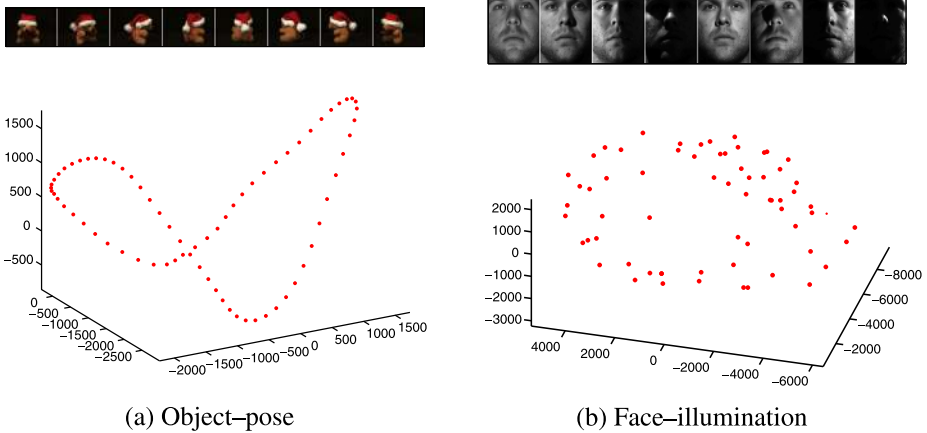
(a) Object–pose   (b) Face–illumination

Fig. 1 To facilitate learning and be of discriminative use, the representation of patterns must possess a certain structure. Often, this structure is exhibited by confining the locus of members of each class to a manifold. While their higher order statistics can greatly vary (both in the original space, as well as the corresponding class manifold) depending on the manner in which data are acquired, the manifold structure can be regarded as containing only class-specific information

(or linearized, as touched upon previously) manifolds can be derived from the most correlated modes of variation between them. Empirical evidence suggests that this indeed is the case in a broad spectrum of problems. What makes CCA additionally attractive in practice is that canonical correlations between linear subspaces can be computed efficiently and in a numerically stable manner (Björck and Golub 1973). In this paper we propose a framework which inherits these appealing properties of CCA, whilst at the same time differing from CCA in that when applied to set matching it does not have any parameters which need to be manually tuned and is readily extended to a discriminative framework. Our approach to computing a distance between two sets can be considered manifold-based, but implicitly so, as the distance is computed without explicit manifold fitting. Rather, the distance is robustly inferred by employing second order statistics to account for the confidence that a particular observed intra-set variation corresponds to a phenomenon of interest and not noise. The proposed discriminative framework which follows the main method is based on a similar idea.

These issues and technical details pertaining to CCA and its application to set matching are addressed next, in Sect. 2, followed by Sects. 3 and 4 introducing respectively extended CCA and discriminative extended CCA, empirical evaluation in Sect. 5 and a conclusion with a summary of contributions in Sect. 5.3.

2 Set matching using canonical correlation analysis

Consider two finite sets of vectors, $\mathcal{X} \subset \mathbb{R}^D$ and $\mathcal{Y} \subset \mathbb{R}^D$:

$$\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\} \tag{1}$$

$$\mathcal{Y} = \{\mathbf{y}_1, \dots, \mathbf{y}_M\}. \tag{2}$$

Canonical correlation analysis seeks to find a pair of latent variables or *canonical vectors* (Hotelling 1936), $\mathbf{u}_1$ and $\mathbf{v}_1$, such that:

$$\exists \mathbf{a}_1 \in \mathbb{R}^N: \quad \mathbf{u}_1 = [\mathbf{x}_1 \mid \ldots \mid \mathbf{x}_N]\mathbf{a}_1 = \mathbf{X}\mathbf{a}_1 \tag{3}$$

$$\exists \mathbf{b}_1 \in \mathbb{R}^M: \quad \mathbf{v}_1 = [\mathbf{y}_1 \mid \ldots \mid \mathbf{y}_M]\mathbf{b}_1 = \mathbf{Y}\mathbf{b}_1, \tag{4}$$

which maximizes the *canonical correlation coefficient* $\rho_1 \in [0, 1]$ defined as

$$\rho_1 = \max_{\mathbf{u}_1, \mathbf{v}_1} \frac{\langle \mathbf{u}_1, \mathbf{v}_1 \rangle}{\|\mathbf{u}_1\| \|\mathbf{v}_1\|} = \max_{\mathbf{u}_1, \mathbf{v}_1} \frac{\mathbf{u}_1^T \mathbf{v}_1}{\|\mathbf{u}_1\| \|\mathbf{v}_1\|}. \tag{5}$$

Canonical vectors and correlation coefficients of higher orders, up to $\min(N, M)$, can be defined recursively under the constraint of mutual orthogonality between all $\mathbf{u}_i$, as well as all $\mathbf{v}_i$:

$$\forall i, j.\ i \neq j: \ \langle \mathbf{u}_i, \mathbf{u}_j \rangle = 0 \quad \text{and} \quad \langle \mathbf{v}_i, \mathbf{v}_j \rangle = 0 \tag{6}$$

By construction, it holds that:

$$1 \geq \rho_1 \geq \rho_2 \geq \ldots \rho_{\min(N,M)} \geq 0 \tag{7}$$

2.1 Set matching using CCA

In most cases the application of CCA to set matching considers sets of vectors over the same type of features (Hua and Peib 2005; Hotta 2012; Arandjelović 2012). For example, each vector may be a rasterized representation of an image as in Sect. 5 and Kim et al. (2007), Arandjelović (2012) for example. The usual manner of applying CCA to set matching consists of three steps. (i) First, an orthonormal basis set $\mathbf{B}_X$ of the subspace characterizing variations within a set is estimated using principal component analysis. Specifically, then $\mathbf{B}_X \in \mathbb{R}^{D \times d_X}$, a matrix with columns consisting of orthonormal basis vectors spanning the d_X-dimensional linear subspace embedded in a D-dimensional image space, can be computed from the corresponding non-centred covariance matrix $\mathbf{C}_X$:

$$\mathbf{C}_X = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^T, \tag{8}$$

as the row and column space basis of the best rank-D approximation to $\mathbf{C}_X$:

$$\mathbf{B}_X = \arg \min_{\substack{\mathbf{B}_X \in \mathbb{R}^{D \times d_X} \\ \mathbf{B}_X^T \mathbf{B}_X = I}} \min_{\substack{\Lambda \in \mathbb{R}^{d_X \times d_X} \\ \Lambda_{ij} = 0, i \neq j}} \left\| \mathbf{C}_X - \mathbf{B}_X \Lambda \mathbf{B}_X^T \right\|_F^2, \tag{9}$$

where $\|.\|_F$ is the Frobenius norm of a matrix.

The dimensionality d_X of this subspace may be preset, it may be inferred from the distribution of data energy across eigenvector directions or indeed left equal to N—the number of data points. (ii) Then, the canonical correlation coefficients ρ_k between two subspaces $\mathbf{X}$ and $\mathbf{Y}$ can be computed using singular value decomposition (SVD) as the singular values of the matrix $\mathbf{B}_X^T \mathbf{B}_Y$ (Björck and Golub 1973). (iii) Finally, a similarity measure is computed as a function of canonical correlation coefficients. This is often done by averaging the first (i.e. the largest) few (Maki and Fukui 2004), although more complex learning schemes have been proposed (Kim et al. 2007).

Fig. 2 Canonical correlations are invariant to isotropic scale changes, such as those caused by illumination or exposure differences. This effectively constrains the learning of appearance variation to a hypersphere centred at the origin
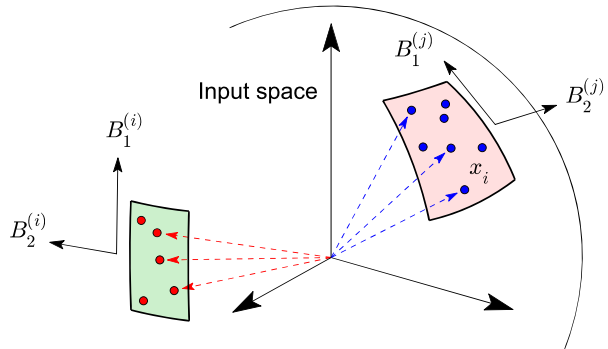
2.1.1 Motivation and advantages

In intuitive terms, canonical vectors extract the most similar modes of variation between two sets, while the corresponding correlation coefficients quantify the degree to which these modes actually match. This focus on that which is common is desirable because it makes the similarity score insensitive to the presence of *differing* modes of variation. These may be present in the data because of different acquisition conditions, or they may indeed correspond to corrupted samples. In contrast, probability density-based methods, such as those using the Bhattacharyya distance or the Kullback-Leibler divergence, do not exhibit such robustness.

By modelling variations within a set by a subspace, canonical correlations are also inherently unaffected by uniform scaling (or, equivalently, the contrast) of individual patterns. This makes CCA-based methods particularly suitable for various computer vision tasks, where such variation may be introduced by the changes in illumination or the duration of exposure of the photosensitive medium. Learning is effectively performed on a hypersphere, as illustrated in Fig. 2.

Finally, in many cases data variations within a single set are low dimensional making CCA-based matching practically appealing due to computational efficiency and low storage requirements.

2.1.2 Limitations of CCA

In modelling class variation, the application of canonical correlation analysis inherently requires the partitioning of the input space into two disjoint subspaces: the class subspace $\mathbf{B}$ of observed variability, and its complementary subspace null($\mathbf{B}^T$). This hard division causes several undesirable consequences.

Optimal parameter choice and performance sensitivity In practice the presence of noise in data is unavoidable and the optimal choice of the dimensionality of the class subspace can seldom be guaranteed (Gou and Fyfe 2009). Most commonly, this parameter is determined empirically, using a training and a validation set (Kim et al. 2007). Not only through ineffective use of data, this approach is also unattractive due to its amplification of noise in the non-dominant directions of the class subspace and the loss of information in the discarded orthogonal directions.
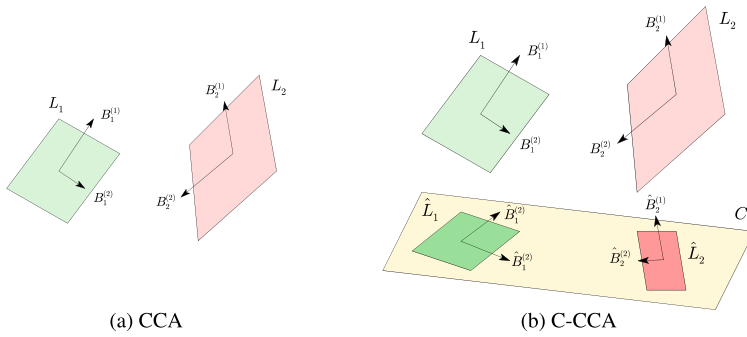
Fig. 3 A conceptual illustration of the principles underlying set matching using a similarity measure based on (**a**) canonical correlation analysis (CCA) and (**b**) its discriminative extension, constrained canonical correlation analysis (C-CCA). Both methods effectively compute the angles between subspaces corresponding to vector sets, preceded by a quasi-discriminative projection in the case of C-CCA

Second order statistics   To make matters worse, any principal direction is either included at an equal footing with all others in the class subspace, or entirely discarded. As we shall demonstrate, the performance of CCA-based matching is very sensitive to the choice of this parameter. This is particularly pronounced in cases suffering from limited sample size, when the number of data points needed for robust estimation for as few as two canonical correlations, is 40 to 70 times the dimensionality of the class subspace (Barcikowski and Stevens 1975).

Discriminative learning   The described CCA-based set distance is not discriminative in nature—sets are compared in an independent, pair-wise manner, without regard for inter-class and intra-class variability. That this cannot be optimal can be seen easily by noting that depending on the application, the same two sets can be regarded as belonging to either the same or different classes: two sets of face appearance images of two different individuals correspond to the same class if the problem is that of face detection (classification to "face" and "non-face"), and different classes if it is face recognition (classification by the identity).

The first discriminative extension of CCA was proposed by Oja (1983). It consists of a linear projection of data onto a subspace which orthogonalizes basis vectors of different classes. The main shortcoming of this approach lies in its lack of robustness, with the orthogonal discriminative criterion leading to overfitting. A different linear subspace approach, which we will refer to as *constrained canonical correlation analysis* (C-CCA) was described by Fukui and Yamaguchi (2003). They introduce a discriminative, *constraint* subspace, defined by the principal components corresponding to the *smallest* eigenvalues of the mean projection matrix across all classes:

$$\bar{\mathbf{P}} = \sum_i \mathbf{P}_i = \sum_i \left(\mathbf{B}_i \mathbf{B}_i{}^T \right) \tag{10}$$

The computation of canonical correlations is then preceded by a linear projection to the constraint subspace, as illustrated conceptually in Fig. 3. For the optimal size of its dimensionality, C-CCA generally outperforms Oja's orthogonal subspace method (Kim et al. 2007), as well as non-discriminative CCA (Nishiyama et al. 2005; Arandjelović and Cipolla 2006). However, ensuring that the optimal value is chosen is difficult. In addition, the construction of the described constraint subspace is *ad hoc* in nature—it does not maximize any

meaningful discriminative function and it does not take into account inter-class variation, relying purely on intra-class variability. A solution to this problem was proposed by Kim et al. (2007), in the form of an iterative method which incrementally adjusts the optimal projection subspace to maximize expected inter-class to intra-class canonical correlations. This is achieved at a great computational cost, the loss of closed-form computability and with the restriction of discrimination between only two classes. Also, just like the previous two approaches, this method suffers from an increased number of free parameters and the "all or nothing" modelling of class distributions.

Nonlinearity   Finally, for completeness, we briefly mention that simple CCA assumes linear intra-class variability. Effective solutions to this problem have been described in the literature, involving either piece-wise linearization of class manifolds (Kim et al. 2007) or their "unfolding" (Fukui et al. 2006; Roweis and Saul 2001). All of these approaches eventually reduce to the computation of CCA in its original form and can thus without modification be applied with the methods we propose in this paper.

3 Extended canonical component analysis (E-CCA)

Motivated by the standard canonical component analysis, in this section we too seek to find the most correlated modes of variation within two sets. However, unlike before, we wish to do so without discarding any data i.e. without the partitioning of the input space into principal (relevant) and complementary (noise) subspaces. As explained in Sect. 2.1.2 this partitioning is a source of major problems when CCA is applied to set matching. For illustration, consider the problem of matching sets of images (see Sect. 5 for a practical example). Generally, each of the image pixels is affected by noise which contains a non-correlated component across different pixels. This means that if sufficient data is available, each of the image sets will exhibit variation in all directions of the input image space. Consequently, all of the canonical correlations between any two sets would be 1.0 if subspace projection described in Sect. 2.1 was not applied (i.e. if the input space was not partitioned). To avoid the need for this projection and the potential for information loss thus effected, we wish to derive a CCA inspired set similarity measure which takes into account the confidence that the observed mode of variation is indeed due to data variability and not noise. We achieve this by incorporating second order statistics into the similarity measure.

Our approach is motivated by observing that for the pair of canonical vectors $\mathbf{u}_1$ and $\mathbf{v}_1$ in Eqs. (3)–(4) there exists a direction $\mathbf{w}_1$ for which:

$$\mathbf{u}_1 = \mathbf{B}_X \mathbf{B}_X{}^T \mathbf{w}_1 \tag{11}$$

$$\mathbf{v}_1 = \mathbf{B}_Y \mathbf{B}_Y{}^T \mathbf{w}_1 \tag{12}$$

The first multiplication, by $\mathbf{B}_X{}^T$ or $\mathbf{B}_Y{}^T$, has the effect of removing any variation not spanned by the columns of $\mathbf{B}_X$ and $\mathbf{B}_Y$ respectively (the two principal subspaces), while the second multiplication by $\mathbf{B}_X$ or $\mathbf{B}_Y$, re-embeds the vectors in the original input space.

The first canonical correlation coefficient can then be written as:

$$\rho_1 = \frac{\mathbf{w}_1{}^T \mathbf{B}_X \mathbf{B}_X{}^T \mathbf{B}_Y \mathbf{B}_Y{}^T \mathbf{w}_1}{\|\mathbf{B}_X{}^T \mathbf{w}_1\| \|\mathbf{B}_Y{}^T \mathbf{w}_1\|} = \frac{\mathbf{w}_1{}^T \mathcal{W}(\hat{\boldsymbol{\Sigma}}_X){}^T \mathcal{W}(\hat{\boldsymbol{\Sigma}}_Y) \mathbf{w}_1}{\|\mathbf{B}_X{}^T \mathbf{w}_1\| \|\mathbf{B}_Y{}^T \mathbf{w}_1\|} \tag{13}$$
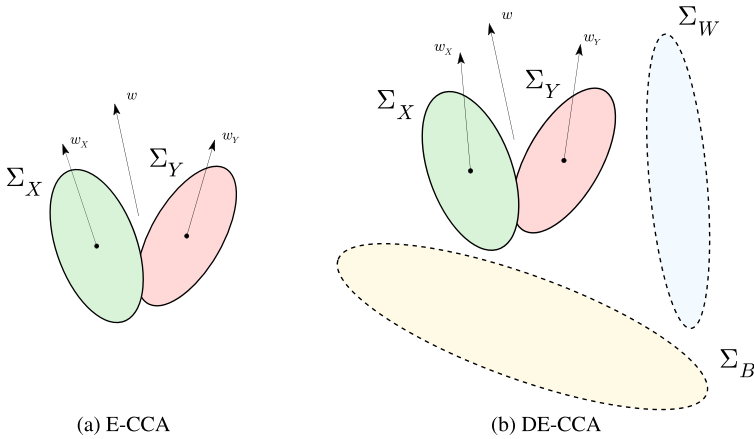
Fig. 4 A conceptual illustration of the principles underlying set matching using a similarity measure based on the proposed (**a**) *extended* canonical correlation analysis (E-CCA) and (**b**) its discriminative variant, *discriminative extended* canonical correlation analysis (DE-CCA). Set similarity is computed by considering projective space distortions corresponding to the sets' covariances and, in the case of DC-CCA, discriminative inter-class and intra-class covariances inferred from training data

where $\mathcal{W}(\ldots)$ is the covariance matrix whitening function (Duda et al. 2000), and:

$$\hat{\boldsymbol{\Sigma}}_X = \mathbf{B}_X \hat{\boldsymbol{\Lambda}}_X \mathbf{B}_X{}^T \tag{14}$$

$$\hat{\boldsymbol{\Sigma}}_Y = \mathbf{B}_Y \hat{\boldsymbol{\Lambda}}_Y \mathbf{B}_Y{}^T \tag{15}$$

Instead of a normalized projection of the vector $\mathbf{w}_1$ onto a subspace, say $\mathbf{B}_X$, consider its transformation effected by the corresponding "deviation" matrix $\boldsymbol{\Upsilon}_X$ (also positive semi-definite and symmetric) which we define as:

$$\boldsymbol{\Upsilon}_X = (\boldsymbol{\Sigma}_X)^{1/2} = \mathbf{V}_X \begin{bmatrix} \sqrt{\lambda_X^{(1)}} & 0 & 0 & 0 \\ 0 & \sqrt{\lambda_X^{(2)}} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \sqrt{\lambda_X^{(D)}} \end{bmatrix} \mathbf{V}_X{}^T = \mathbf{V}_X (\boldsymbol{\Lambda}_X)^{1/2} \mathbf{V}_X{}^T \tag{16}$$

where:

$$\boldsymbol{\Sigma}_X = \mathbf{X}\mathbf{X}^T = \mathbf{V}_X \boldsymbol{\Lambda}_X \mathbf{V}_X{}^T \tag{17}$$

is the full data covariance matrix. The said transformation anisotropically scales its input, amplifying it in the directions in which $\mathcal{X}$ exhibits significant variability (large $\lambda_X^{(i)}$) and attenuating in those with little (or indeed no) variability (small $\lambda_X^{(i)}$), as illustrated in Fig. 4. This effect can be considered a generalization of the projection described in Eq. (15)—while in the application of standard CCA all variation in the complementary subspace is discarded and the variation in the principal subspace whitened, here the transformation has the effect of smoothly emphasizing or de-emphasizing different directions of variability according to its extent (specifically, proportionally to the standard deviation in the corresponding direction).

In the special case of data which exhibits isotropic variability constrained to a subspace the result is exactly the same as in Eq. (15) (up to scale).

Motivated by this intuition and extending the analogy to Eq. (13), we seek such unit vector $\hat{\mathbf{w}}_1$ whose projections by $\mathbf{\Upsilon}_X$ and $\mathbf{\Upsilon}_Y$ have the highest degree of correlation. Formally, we define the first extended canonical correlation coefficient ψ_1 between $\mathcal{X}$ and $\mathcal{Y}$ as:

$$\psi_1 = \max_{\hat{\mathbf{w}}_1}\left\{(\mathbf{\Upsilon}_X\hat{\mathbf{w}}_1)^T(\mathbf{\Upsilon}_Y\hat{\mathbf{w}}_1)\right\} = \max_{\hat{\mathbf{w}}_1}\left\{\hat{\mathbf{w}}_1^T\mathbf{\Upsilon}_X^T\mathbf{\Upsilon}_Y\hat{\mathbf{w}}_1\right\} \tag{18}$$

$$= \max_{\hat{\mathbf{w}}_1}\left\{\hat{\mathbf{w}}_1^T\mathbf{\Upsilon}_X^T\mathbf{\Upsilon}_Y\hat{\mathbf{w}}_1\right\} = \max_{\hat{\mathbf{w}}_1}\left\{\hat{\mathbf{w}}_1^T\mathbf{\Phi}_{XY}\hat{\mathbf{w}}_1\right\}, \tag{19}$$

under the constraint:

$$\|\hat{\mathbf{w}}_1\| = 1. \tag{20}$$

The key idea here is that this measure will favour those directions of the space in which both $\mathcal{X}$ and $\mathcal{Y}$ have significant variability (it is "amplified" both by $\mathbf{\Upsilon}_X$ and $\mathbf{\Upsilon}_Y$). Similarly, a direction in which $\mathbf{\Upsilon}_X$ (say) exhibits significant variability, but $\mathbf{\Upsilon}_Y$ does not, will contribute to $\hat{\mathbf{w}}_1$ less, while a direction in which neither of the sets exhibit significant variability will be greatly de-emphasized and have little effect on $\hat{\mathbf{w}}_1$.

Note that although matrices $\mathbf{\Upsilon}_X$ and $\mathbf{\Upsilon}_Y$ are symmetric, their product is not. Nonetheless, $\mathbf{\Phi}_{XY}$ is positive semi-definite. This is because linear transformations effected by $\mathbf{\Upsilon}_X$ and $\mathbf{\Upsilon}_Y$ involve no rotation, reflection or shearing, i.e. they can be expressed purely as a combination of orthogonal projection, and (generally anisotropic) scaling. Thus ψ_1 is maximized when $\hat{\mathbf{w}}$ is in the direction of the eigenvector of $\mathbf{\Phi}_{XY}$ corresponding to its largest eigenvalue, for which:

$$\psi_1 = \lambda_\Phi^{(1)}, \tag{21}$$

where $\lambda_\Phi^{(1)} \geq \lambda_\Phi^{(2)} \geq \ldots \lambda_\Phi^{(D)} \geq 0$ are the eigenvalues of $\mathbf{\Phi}_{XY}$.

When classical canonical correlation analysis is applied to pattern recognition in practice, only the first few correlation coefficients are computed. This is largely a consequence of the trade-off between the accuracy of matching and its speed. Our approach does not suffer from the same weakness. Extending the analysis to higher order extended correlation coefficients, we can quantify the degree of agreement between variations observed in sets $\mathcal{X}$ and $\mathcal{Y}$ as:

$$\mu_{XY} = \frac{\sum_{i=1}^{D} \psi_i}{\sum_{i=1}^{D} \sqrt{\lambda_X^{(i)}\lambda_Y^{(i)}}} \tag{22}$$

More elaborate combinations of ψ_i are possible, e.g. as described in Kim et al. (2007), but here we adopt a simple normalized summation because it lends itself to particularly efficient computation, as we will show shortly.

Note that μ_{XY} reaches its *maximum maximorum* when $\mathbf{\Upsilon}_X$ and $\mathbf{\Upsilon}_Y$ share the same eigenspace and when their eigenvectors of the same rank (with respect to the magnitude of the corresponding eigenvalue) are aligned. However, in general:

$$0 \leq \mu_{XY} \leq 1. \tag{23}$$

The class similarity μ_{XY} in Eq. (22) can be rapidly computed by noticing that:

$$\sum_{i=1}^{D} \psi_i = \text{Tr}[\mathbf{\Phi}_{XY}], \tag{24}$$

while the values of $\lambda_X^{(1)}, \ldots, \lambda_X^{(D)}$ and $\lambda_Y^{(1)}, \ldots, \lambda_Y^{(D)}$ are estimated in the same manner as in the case of classical CCA.

It is important to observe that in Eq. (19) there is no concern of an ill-defined result because of a vanishing denominator (unlike in the case of the method described in Arandjelović and Cipolla (2006) for example). Specifically, since $\lambda_X^{(i)}$ and $\lambda_Y^{(i)}$ are ordered in magnitude, the product $\lambda_X^{(1)} \lambda_Y^{(1)}$ cannot be 0 as both sets $\mathcal{X}$ and $\mathcal{Y}$ are assumed to contain at least some variability. Indeed, this is a necessary condition both for classical CCA and the proposed method to be meaningful in this context.

4 Discriminative extended canonical component analysis (DE-CCA)

In the previous section we derived a principled extension of canonical correlation analysis suitable for matching sets of patterns constrained to linear subspaces. We addressed the inherent limitations of CCA when applied on this problem: the inevitable "hard" partitioning of the input space, as well the practical difficulty of parameter estimation. Unlike classical CCA, we now show that our method readily lends itself to a discriminative learning framework.

In Sect. 3 we considered space transformation by means of projection using the square root of the class covariance matrix:

$$\mathbf{w}_1 \longrightarrow \Upsilon_X \mathbf{w}_1 = (\Sigma_X)^{1/2} \mathbf{w}_1. \tag{25}$$

Its effect is the amplification of modes of variation common to the input vector $\mathbf{w}_1$ and the set $\mathcal{X}$. However, this is achieved without any knowledge of variability both between and within different classes. Here we capture these through two covariance matrices, the *normalized inter-class scatter matrix* Σ_B and the *mean intra-class scatter matrix* Σ_W. Using the notation $\mathbf{Z}_i$ for different training set data matrices $i = 1, \ldots, N_C$ (each corresponding to a different class) and denoting their members by $\{\mathbf{z}_{i1}, \mathbf{z}_{i2}, \ldots\}$, we define the normalized inter-class scatter matrix as follows:

$$\Sigma_B = \frac{1}{N_C} \sum_{i=1}^{N_C} \left(\frac{E[\mathbf{Z}_i]}{\|E[\mathbf{Z}_i]\|} - \mathbf{m} \right) \tag{26}$$

where:

$$\mathbf{m} = \frac{1}{N_C} \sum_{i=1}^{N_C} \frac{E[\mathbf{Z}_i]}{\|E[\mathbf{Z}_i]\|}. \tag{27}$$

Note that the explicit normalization of class data means $E[\mathbf{Z}_i]$ in Eq. (26) is necessary here (see Sect. 2.1.1). The mean intra-class scatter matrix is simply:

$$\Sigma_W = \frac{1}{N_C} \sum_{i=1}^{N_C} \mathbf{Z}_i{}^T \mathbf{Z}_i = \frac{1}{N_C} \sum_{i=1}^{N_C} \Sigma_{Z_i} \tag{28}$$

Our definitions of inter-class and intra-class matrices are similar to those used in Fisher's discriminant analysis (Duda et al. 2000).

The two scatter matrices are then used to further transform an input vector, first by accentuating its components in the direction of common intra-class variations and then by attenuating those corresponding to inter-class variability:

$$\mathbf{w}_1 \longrightarrow (\boldsymbol{\Sigma}_B)^{-1/2}(\boldsymbol{\Sigma}_W)^{1/2}\boldsymbol{\Upsilon}_X\mathbf{w}_1. \tag{29}$$

The intuition behind this transformation is exactly the same as that for the non-discriminative version in Sect. 3. Instead of partitioning the space into discriminative and non-discriminative subspaces and projecting the data onto the former, like in CCA for example, our transformation of data is smoother in nature. While a projection onto the discriminative subspace entirely aligns the data with the subspace, our transformation instead realigns the data smoothly first by emphasizing discriminative directions (according to the inter-class scatter matrix) and then by further de-emphasizing the non-discriminative ones (according to the intra-class scatter matrix). The idea is illustrated conceptually in Fig. 4(b). Note that the aforementioned directions are not explicitly determined. Rather, inter- and intra-class variances are automatically combined into the optimal weighting matrix $(\boldsymbol{\Sigma}_B{}^{-1}\boldsymbol{\Sigma}_W)^{1/2}$. Formulating a criterion similar to that in Eq. (19) now leads to eigenvalue decomposition of the matrix $\hat{\boldsymbol{\Phi}}_{XY}$:

$$\hat{\boldsymbol{\Phi}}_{XY} = \boldsymbol{\Upsilon}_X\mathbf{P}\boldsymbol{\Upsilon}_Y \tag{30}$$

where:

$$\mathbf{P} = (\boldsymbol{\Sigma}_W)^{1/2}(\boldsymbol{\Sigma}_B)^{-1}(\boldsymbol{\Sigma}_W)^{1/2} \tag{31}$$

Just as $\boldsymbol{\Phi}_{XY}$ in the non-discriminative case and using the same argument as on Sect. 3, $\hat{\boldsymbol{\Phi}}_{XY}$ can be recognized as a non-symmetric but nonetheless positive semi-definite matrix. Thus, the first discriminative extended canonical correlation coefficient is equal to its largest eigenvalue and, as in Eq. (22), the overall similarity of sets $\mathcal{X}$ and $\mathcal{Y}$ becomes:

$$\hat{\mu}_{XY} = \frac{\sum_{i=1}^{D}\lambda_\phi^{(i)}}{\sum_{i=1}^{D}\sqrt{\lambda_X^{(i)}\lambda_Y^{(i)}}} = \frac{\mathrm{Tr}\,[\hat{\boldsymbol{\Phi}}_{XY}]}{\sum_{i=1}^{D}\sqrt{\lambda_X^{(i)}\lambda_Y^{(i)}}} \tag{32}$$

5 Experimental evaluation

We empirically examined the validity of theoretical arguments put forward in the preceding sections on the problem of matching sets of images of faces. To make our experiments as directly comparable as possible to those in the published previous work most closely related to ours, we evaluated the proposed methods on a database already widely used for this purpose (e.g. see Arandjelović and Cipolla 2006; Kim et al. 2007) and described in detail in Arandjelović (2012).

This database contains video sequences of face motion for 100 individuals of varying ages and ethnicities. For each person in the database there are 7 sequences of the person performing loosely constrained, pseudo-random motion (significant translation, yaw and pitch, negligible roll) for 10 s, as shown on an example in Fig. 5(a). Each sequence was acquired in a different illumination setting as illustrated in Fig. 5(b). Sequences were acquired at 10 fps
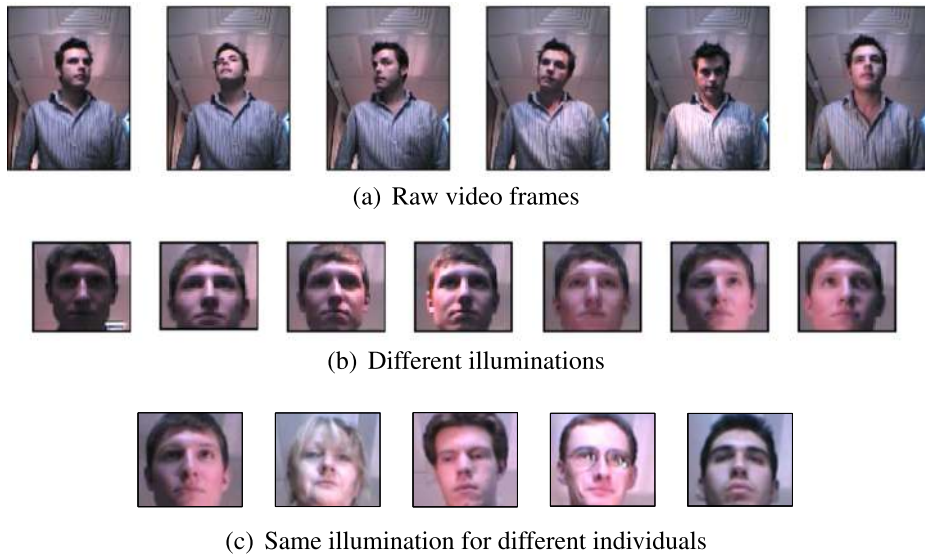
(a) Raw video frames

(b) Different illuminations

(c) Same illumination for different individuals

Fig. 5 (**a**) Frames from typical video sequences in the database used for evaluation. (**b**) Different illumination conditions used to acquire data, illustrated on manually selected frontal faces. (**c**) Five different individuals in the illumination setting number 6. In spite of the same spatial arrangement of light sources, their effect on the appearance of faces changes significantly due to variations in people's heights, the *ad lib* chosen position relative to the camera etc.

and in 320×240 pixel resolution (face size ≈ 60 pixels).[1] The users were asked to approach the camera while performing arbitrary head motion. Although the illumination was kept constant throughout each sequence, there is some variation in the manner in which faces were lit due to the change in the relative position of the user with respect to the lighting sources, as shown in Fig. 5(c).

Faces were detected automatically using the cascaded detector of Viola and Jones (2004) are rescaled to the uniform resolution of 50×50 pixels which is approximately the average size of a detected face (see Arandjelović 2010 for a related discussion). Face image patches were then converted into vectors by column-wise rasterization, each video sequence thus producing a set of vectors in the $\mathbb{R}^{2500}$ space. Different distance measures between sets were evaluated in the context of one-to-many matching. In other words, image sets extracted from video sequences of different individuals in a particular setting were used as training data while querying was performed using sets corresponding to a different illumination. Each query set was associated with the best matching training set. In this manner, we investigated:

– the sensitivity of the classical CCA to the number of correlation coefficients,
– the sensitivity of the C-CCA to the dimensionality of the constraining subspace,
– the performance of the proposed E-CCA distance measure, and
– the performance of the proposed DE-CCA distance measure.

---

[1] A thorough description of the University of Cambridge face database with examples of video sequences is available at http://mi.eng.cam.ac.uk/~oa214/.

Method	CCA			C-CCA (3 cc)			
Parameters	1 cc	2 cc	3 cc	−15d	−25d	−50d	−100d
Mean correct recognition rate	42.2	67.0	75.1	78.7	83.6	83.2	74.4
Recognition rate deviation	10.6	13.9	13.5	11.1	8.3	9.0	15.4

Method	E-CCA	**DE-CCA**
Mean correct recognition rate	83.4	**88.8**
Recognition rate deviation	9.6	**7.8**

Fig. 6 A summary of experimental results. Shown are the average correct recognition rate across different illuminations used for training and querying, and its standard deviation. Even for the optimal choice of parameters, which is impossible to ensure in practice, classical canonical correlation analysis (CCA: a non-discriminative approach) and its constrained extension (C-CCA: discriminative) were outperformed by the proposed methods which are parameter-free

5.1 Results and discussion

A summary of the key results is shown in Fig. 6. As expected in the case of data with such complex variability as exhibited by face appearance images, the performance of the classical CCA-based matching is greatly affected by the number of canonical correlation coefficients used to compute the set-to-set similarity measure. When their number is increased from one to two, the average performance is improved by 59 %, and further by 78 % for three coefficients. Additional sensitivity of the method to free parameter selection can be observed in Fig. 7(a), which illustrates the importance of the dimensionality of class subspaces.

It is revealing to notice that while an increase in the number of computed canonical correlations improves the average rate of correct identification, the confidence in the recognition decision (quantified by the deviation of the correct identification rate across different test and query illumination conditions) actually worsens. This phenomenon can be explained by observing that increasing the number of coefficients does not improve matching decisions in the most challenging cases (corresponding to test and query illumination combinations that result in the lowest correct identification rates). In other words, in these cases, the variations present in different appearance sets corresponding to the same person are indeed very unlike one another, and the only way of producing an improvement is by incorporating discriminative constraints.

The results obtained using constrained CCA highlight further limitations of classical canonical correlation analysis in set matching. In addition to the sensitivity of the method to the free parameters shared with CCA—the number of canonical correlations used to compute set similarities and the dimensionality of class subspaces, illustrated for C-CCA in Fig. 7(b)—an additional free parameter is introduced in the form of the dimensionality of the constraint subspace. Discriminative projection of data was found to improve performance only for constraint subspace dimensionalities greater than $(D - 95)$, which is a narrow range of just ∼3.8 % of the input space dimensionality $D = 2500$. In this range, the mean recognition rate is increased and its deviation decreased. Peak correct recognition rate of 83.6 % (33 % reduction in error rate compared to simple CCA) is achieved when three canonical correlation coefficients are used and the dimensionality of the constraint subspace set equal to $(D - 25)$. However, a suboptimal choice of the constraint subspace dimensionality can rapidly lead to significant worsening of performance, as shown in Fig. 7(c). This is a major limitation of C-CCA, as there is no fundamental theoretical basis which could facilitate the inference of this parameter.

Fig. 7 The measured variation of the correct recognition rate averaged over different illumination conditions used for training and querying, as a function of the dimensionality of the constraint subspace and the number of canonical correlation coefficients used for matching (*green* = 1, *red* = 2, *blue* = 3). Optimal performance is achieved using three canonical correlations and $(D - 25)$-dimensional constraint subspace (Color figure online)
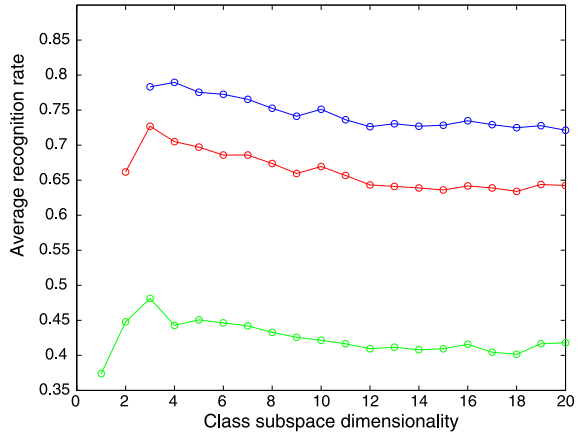
(a) Canonical correlation analysis
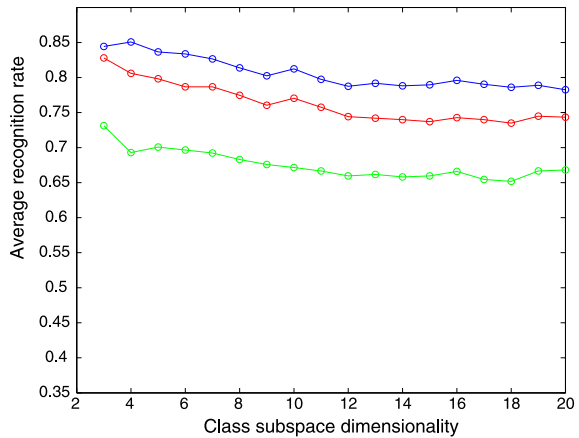
(b) Extended canonical correlation analysis

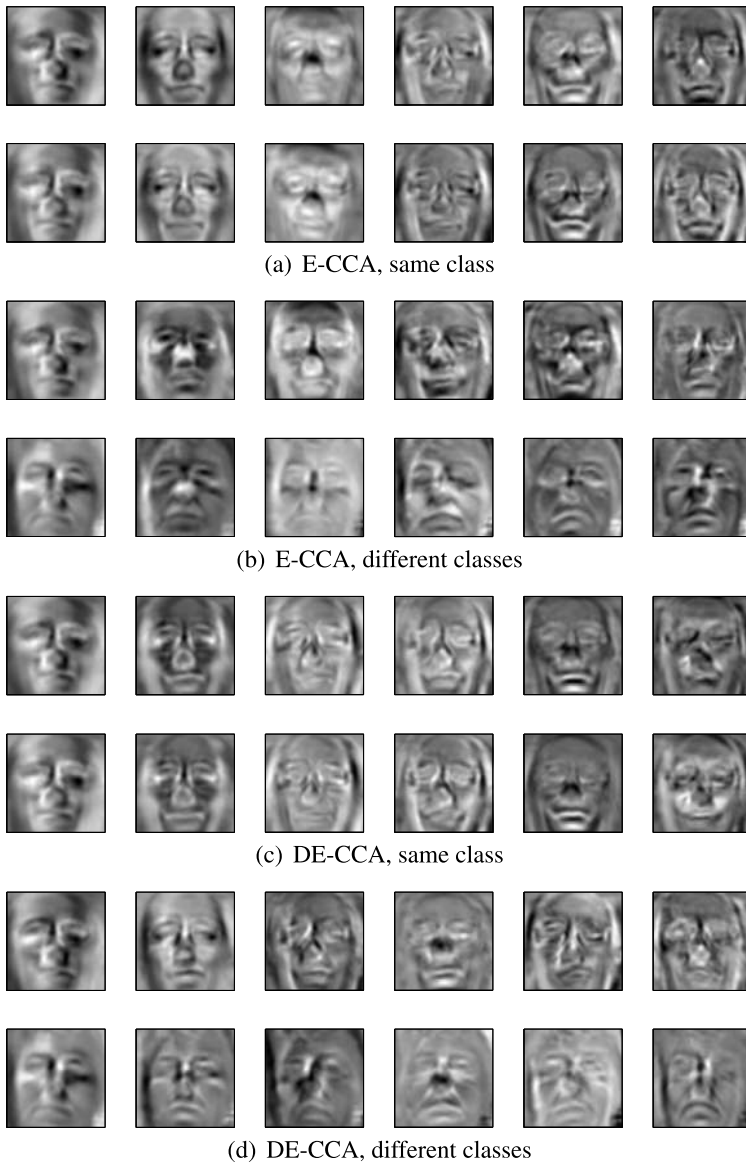(c) Discriminative extended canonical correlation analysis

(a) E-CCA, same class

(b) E-CCA, different classes

(c) DE-CCA, same class
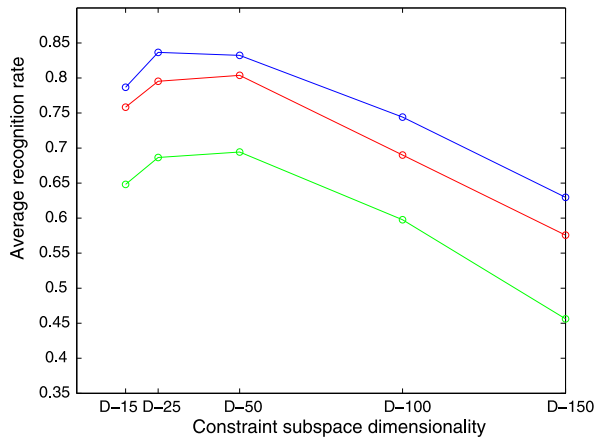
(d) DE-CCA, different classes

Fig. 8 The most similar 6 modes of variation extracted using (**a**, **b**) E-CCA and (**c**, **d**) DE-CCA using sets corresponding to the same person in different illuminations (**a**, **c**) and different persons. In the case of same class matching, both methods successfully extract common directions of appearance change between two sets. The benefit of DE-CCA is more pronounced when sets of images of different persons are matched—E-CCA erroneously manages to find rather more alike modes of variation by effectively matching on common illumination, while the discriminative algorithm correctly learns and de-emphasizes such confounding factors

As this paper argued in detail, the problem of free parameter choice is made entirely obsolete by our use of second order statistics. The proposed E-CCA was found to significantly outperform not only CCA but C-CCA as well, and remarkably for *all values of their free*

parameters, as can be seen in Fig. 6. Practically, this is a major advantage of our approach, due to the difficulty of ensuring optimal parameter selection for C-CCA. Theoretically, this confirms the premise set forth in Sect. 1 which emphasizes the loss of discriminative information in discarding higher order statistics of class variability. This is an important result given that unlike E-CCA, C-CCA is a discriminative approach. With that in mind, it is not surprising that our discriminative method, DE-CCA, improves performance even further, decreasing the error rate by 32 % from C-CCA—see Fig. 6.

5.2 Qualitative insight

It is insightful to visually examine the modes of the most probable common variability between sets of the same and different classes, extracted by the proposed methods. These can be found as the dominant eigenvectors of the matrix $\boldsymbol{\Phi}_{XY}$ in the case of E-CCA (see Sect. 3), and $\hat{\boldsymbol{\Phi}}_{XY}$ (see Sect. 4) in the case of DE-CCA. Examples, visualized as images, are shown in Fig. 8. Notice that E-CCA manages to find meaningful common modes even when the two sets correspond to different people, just as it does when they show the same person—it effectively matches similar appearing illumination effects. This is a consequence of geometric and textural similarity of human faces, which makes face recognition so difficult. By learning a discriminative criterion, matching using DE-CCA has the consequence of de-emphasizing confounding inter-personal variability and amplifying intra-personal information.

5.3 Summary and conclusions

In this paper we proposed a novel framework for matching vector sets. Our approach is based on inference of the most similar modes of variability within two sets. This led to a comparison with increasingly popular canonical correlation analysis-based methods. These were discussed in detail and it was shown, both theoretically and empirically, that they have significant practical limitations of which the most important are: (i) the presence of free parameters which cannot be inferred from the data, (ii) non-robust partitioning of the space into class and non-class subspaces, and (iii) intractability of discriminative learning. In contrast, the proposed *extended canonical correlation analysis*-based method inherently accounts for uncertainty in data, inferring *the most likely* common modes of variability. What is more, it is shown that this can be achieved without the loss of attractive computational efficiency of canonical correlation analysis, whereby set similarity is effectively reduced to the computation of the trace of a matrix. The proposed framework was then extended into a discriminative learning scheme which, unlike in the case of classical canonical correlation analysis, follows naturally. Finally, our theoretical arguments were empirically verified on the task of set-based face recognition. The proposed methods were shown superior even for the optimal choice of parameters of the classical canonical correlation analysis-based methods, which is impossible to ensure in practice.

References

Arandjelović, O. (2010). Recognition from appearance subspaces across image sets of variable scale. In *Proc. British machine vision conference (BMVC)*. doi:10.5244/C.24.79.

Arandjelović, O. (2012). Colour invariants under a non-linear photometric camera model and their application to face recognition from video. *Pattern Recognition*, *45*(7), 2499–2509.

Arandjelović, O. (2012). Computationally efficient application of the generic shape-illumination invariant to face recognition from video. *Pattern Recognition*, *45*(1), 92–103.

Arandjelović, O., & Cipolla, R. (2006). An information-theoretic approach to face recognition from face motion manifolds. *Image and Vision Computing*, *24*(6), 639–647. Special issue on Face Processing in Video

Arandjelović, O., & Cipolla, R. (2006). A new look at filtering techniques for illumination invariance in automatic face recognition. In *Proc. IEEE international conference on automatic face and gesture recognition (FG)* (pp. 449–454).

Arandjelović, O., & Cipolla, R. (2006). Face set classification using maximally probable mutual modes. In *Proc. IEEE international conference on pattern recognition (ICPR)* (pp. 511–514).

Arandjelović, O., & Cipolla, R. (2013). Achieving robust face recognition from video by combining a weak photometric model and a learnt generic face invariant. *Pattern Recognition*, *46*(1), 9–23.

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

Barcikowski, R., & Stevens, J. P. (1975). A Monte Carlo study of the stability of canonical correlations, canonical weights, and canonical variate-variable correlations. *Multivariate Behavioral Research*, *10*, 353–364.

Bhattacharyya, A. (1943). On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, *35*, 99–109.

Björck, Å., & Golub, G. H. (1973). Numerical methods for computing angles between linear subspaces. *Mathematics of Computation*, *27*(123), 579–594.

Borg, I., & Groenen, P. (2005). *Modern multidimensional scaling: theory and applications* (2nd ed.). New York: Springer.

Chernoff, H. (1952). A measure of asymptotic efficiency of tests for a hypothesis based on a sum of observation. *The Annals of Mathematical Statistics*, *23*, 493–507.

Choukri, K., & Chollet, G. (1986). Adaptation of automatic speech recognizers to new speakers using canonical correlation analysis techniques. *Computer Speech & Language*, *1*(2), 95–107.

Dou, M., Zhang, C., Hao, P., & Li, J. (2007). Converting thermal infrared face images into normal gray-level images. *Lecture Notes in Computer Science*, *4844*, 722–732.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000). *Pattern classification* (2nd ed.). New York: Wiley

Fan, W., & Yeung, D.-Y. (2006). Face recognition with image sets using hierarchically extracted exemplars from appearance manifolds. In *Proc. IEEE international conference on automatic face and gesture recognition (FG)* (pp. 177–182).

Fukui, K., & Yamaguchi, O. (2003). Face recognition using multi-viewpoint patterns for robot vision. In *International symposium of robotics research*.

Fukui, K., Stenger, B., & Yamaguchi, O. (2006). A framework for 3D object recognition using the kernel constrained mutual subspace method. In *Proc. Asian conference on computer vision (ACCV)* (pp. 315–324).

Fukumizu, K., Bach, F., & Gretton, A. (2007). Consistency of kernel canonical correlation analysis. *Journal of Machine Learning Research*, *8*, 361–383.

Gou, Z., & Fyfe, C. (2009). Generalised canonical correlation analysis. In *Proc. international conference on intelligent data engineering and automated learning, data mining, financial engineering, and intelligent agents (IDEAL)* (pp. 164–173).

Gunturk, B. K., Batur, Y., Altunbasak, A. U., Hayes, M. H., & Mersereau, R. M. (2003). Eigenface-domain super-resolution for face recognition. *IEEE Transactions on Image Processing*, *12*(5), 597–606.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, *28*, 321–372.

Hotta, K. (2012). Local co-occurrence features in subspace obtained by kpca of local blob visual words for scene classification. *Pattern Recognition*, *45*(10), 321–372.

Hua, M., & Peib, J. (2005). Clustering in applications with multiple data sources—a mutual subspace clustering approach. *Neurocomputing*, *92*, 133–144.

Kim, T.-K., Arandjelović, O., & Cipolla, R. (2007). Boosted manifold principal angles for image set-based recognition. *Pattern Recognition*, *40*(9), 2475–2484.

Kim, T.-K., Kittler, J. V., & Cipolla, R. (2007). Discriminative learning and recognition of image set classes using canonical correlations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(6), 1005–1018.

Kullback, R. A., & Leibler, S. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86.

Lee, K., Ho, J., & Kriegman, D. (2003). Nine points of light: acquiring subspaces for face recognition under variable lighting. In *Proc. IEEE conference on computer vision and pattern recognition (CVPR)* (Vol. 1, pp. 519–526).

Maki, A., & Fukui, K. (2004). Ship identification in sequential ISAR imagery. *Machine Vision and Applications*, *15*(3).

Melzer, T., Reiter, M., & Bischofm, H. (2003). Appearance models based on kernel canonical correlation analysis. *Pattern Recognition*, *36*(9), 1961–1971.

Nishiyama, M., Yamaguchi, O., & Fukui, K. (2005). Face recognition with the multiple constrained mutual subspace method. In *Proc. audio and video-based biometric person authentication* (pp. 71–80).

Oja, E. (1983). *Subspace methods of pattern recognition*. Letchworth/New York: Research Studies Press/Wiley.

Reiter, M., Donner, R., Langs, G., & Bischof, H. (2006). 3D and infrared face reconstruction from RGB data using canonical correlation analysis. In *Proc. IEEE international conference on pattern recognition (ICPR)* (Vol. 1, pp. 425–428).

Roweis, S., & Saul, L. K. (2001). Nonlinear dimensional reduction by locally linear embedding. *Science*, *290*.

Saisan, P., Doretto, G., Wu, Y. N., & Soatto, S. (2001). Dynamic texture recognition. In *Proc. IEEE conference on computer vision and pattern recognition (CVPR)* (Vol. 2, pp. 58–63).

Satoh, S. (2000). Comparative evaluation of face sequence matching for content-based video access. In *Proc. IEEE international conference on automatic face and gesture recognition (FG)* (pp. 163–168).

Shakhnarovich, G., Fisher, J. W., & Darrel, T. (2002). Face recognition from long-term observations. In *Proc. European conference on computer vision (ECCV)* (Vol. 3, pp. 851–868).

Sinanović, S., & Johnson, D. H. (2007). Toward a theory of information processing. *Signal Processing*, *87*, 1326–1344.

Viola, P., & Jones, M. (2004). Robust real-time face detection. *International Journal of Computer Vision*, *57*(2), 137–154.

Vivek, E. P., & Sudha, N. (2007). Robust Hausdorff distance measure for face recognition. *Pattern Recognition*, *40*(2), 431–442.

Wolf, L., & Shashua, A. (2003). Learning over sets using kernel principal angles. *Journal of Machine Learning Research*, *4*(10), 913–931.

Yang, M.-H. (2002). Kernel eigenfaces vs. kernel fisherfaces: face recognition using kernel methods. In *Proc. IEEE international conference on automatic face and gesture recognition (FG)* (pp. 215–220).