Dimensionality Reduction by Canonical Contextual Correlation Projections

Marco Loog[1], Bram van Ginneken[1], and Robert P.W. Duin[2]

[1] Image Sciences Institute, University Medical Center Utrecht
Utrecht, The Netherlands
{marco,bram}@isi.uu.nl

[2] Information and Communication Theory Group, Delft University of Technology
Delft, The Netherlands
r.p.w.duin@ewi.tudelft.nl

Abstract. A linear, discriminative, supervised technique for reducing feature vectors extracted from image data to a lower-dimensional representation is proposed. It is derived from classical Fisher linear discriminant analysis (LDA) and useful, for example, in supervised segmentation tasks in which high-dimensional feature vector describes the local structure of the image. In general, the main idea of the technique is applicable in discriminative and statistical modelling that involves contextual data.

LDA is a basic, well-known and useful technique in many applications. Our contribution is that we extend the use of LDA to cases where there is dependency between the output variables, i.e., the class labels, and not only between the input variables. The latter can be dealt with in standard LDA.

The principal idea is that where standard LDA merely takes into account a single class label for every feature vector, the new technique incorporates class labels of its neighborhood in its analysis as well. In this way, the spatial class label configuration in the vicinity of every feature vector is accounted for, resulting in a technique suitable for e.g. image data. This spatial LDA is derived from a formulation of standard LDA in terms of canonical correlation analysis. The linearly dimension reduction transformation thus obtained is called the canonical contextual correlation projection.

An additional drawback of LDA is that it cannot extract more features than the number of classes minus one. In the two-class case this means that only a reduction to one dimension is possible. Our contextual LDA approach can avoid such extreme deterioration of the classification space and retain more than one dimension.

The technique is exemplified on a pixel-based segmentation problem. An illustrative experiment on a medical image segmentation task shows the performance improvements possible employing the canonical contextual correlation projection.

1 Introduction

This paper describes a supervised technique for linearly reducing the dimensionality of image feature vectors (e.g. observations in images describing the local gray level structure at certain positions) taking contextual label information into account (e.g.

T. Pajdla and J. Matas (Eds.): ECCV 2004, LNCS 3021, pp. 562–573, 2004.

the local class label configuration in a segmentation task). The technique is based on canonical correlation analysis and called the canonical contextual correlation projection (CCCP).

In general, the main goal of reducing the dimensionality of feature data is to prevent the subsequently used model from over-fitting in the training phase [9,12]. An important additional effect in, for example, pattern classifiers is often the decreased amount of time and memory required to perform the necessary operations. Consequently image segmentation, object classification, object detection, etc. may benefit from the technique, and also other discriminative methods may gain from it.

The problem this paper is concerned with is of great practical importance within real-world, discriminative and statistical modelling tasks, because in many of these image analysis and computer vision tasks the dimensionality, say n, of the feature data can be relatively large. For example, because it is not clear a priori what image information is needed for a good performance in a pixels classification task, many features per pixel may be included, which results in a high-dimensional feature vector. This already happens in 2-dimensional image processing, but when processing large hyper-spectral images, medical 3-dimensional volumes, or 4-dimensional space/time image data, it may even be less clear what features to take and consequently more features are added. However, high-dimensional data often leads to inferior results due to the curse of dimensionality [4, 12] even if all relevant information for accurate classification is contained in the feature vector. Hence, lowering the dimensionality of the feature vectors can lead to a significant gain in performance.

The CCCP is an extension to linear discriminant analysis (LDA), which is a well-known supervised dimensionality reduction technique from statistical pattern recognition [9,12]. LDA is capable of taking contextual information in the input variables into account, however contextual information in the output variables is not explicitly dealt with. The CCCP does take this information into account and therefore models this contextual information more accurately.

Another principal drawback of LDA is that it cannot extract more features than the number of classes minus one [7,9]. In the two-class case—often encountered in image segmentation or object detection—this means that we can only reduce the dimensionality of the data to one, and even though reducing the dimensionality could improve the performance it is not plausible that one single feature can describe class differences accurately. CCCP can avoid such extreme deterioration of the classification space and retain more than one dimension even in the case of two-class data.

LDA was originally proposed by Fisher [5,6] for the two-class case and extended by Rao [14] to the multi-class case. The technique is supervised, i.e., input and output patterns which are used for training have to be provided. Quite a few linear dimension reduction techniques have been proposed of which many are variations and extensions to Fisher's LDA, see [3,9,16]. Within the field of image classification, [1] and [13] show how classification performance can benefit from linear dimension reduction. The novel extension to LDA given in this paper explicitly deals with the contextual spatial characteristics of image data. To come to this extension of LDA, a formulation of this technique in terms of canonical correlation analysis (CCA, [11]) is used (see [9,16]), which enables us to not only to include the class labels of the pixel that is considered—as in classical LDA, but also to encode information from the surrounding class labelling structure. Related to our approach is the work of Borga [2] in which CCA is also used as a framework for image analysis.

Finally, it is mentioned that there is a close relationship of the LDA considered here and a form of LDA which is used for classification. The latter is also known as a linear discriminant classifier or Fisher's linear discriminant [9,16]. Here, however, LDA for dimensionality reduction is considered.

1.1 Outline

Section 2 formulates the general problem statement within the context of supervised image segmentation. However, we stress that the technique is not restricted to this task. Techniques like object detection or object classification can also benefit from the dimension reduction scheme proposed. Section 3 introduces LDA and discusses its link to CCA. Subsection 3.4 presents the CCCP. Subsection 3.5 discusses the drawback of obtaining too few dimensions with LDA, and explains how CCCP can overcome this limitation. Subsection 3.6 summarizes the main approach. Section 4 presents illustrative results on a lung filed segmentation task in chest radiographs. Finally, Section 5 provides a discussion and conclusions.

2 Problem Statement

To make the exposition more clear, the technique presented is directly related to the specific task of image segmentation, and it is not discussed in its full generality.

An image segmentation task in terms of pixel classification is considered—however, we may as well use other image primitives on a regular lattice. Based on image features associated to a pixel, it is decided to which of the possible classes this pixel belongs. Having classified all pixels in the image gives a segmentation of this image. Examples of features associated to a pixel are its gray level, gray levels of neighboring pixels, texture features, the position in the image, gray levels after linear or non-linear filtering of the image, etc.

Pixels are denoted by p_i and the features extracted from the image associated to p_i are represented in an n-dimensional feature vector $\mathbf{x}_i$. A classifier maps $\mathbf{x}_i$ to a certain class label coming from a set of K possibilities $\{l_1, \dots, l_K\}$. All pixels having the same label belong to the same segment and define the segmentation of the image. The classifier, e.g. a quadratic classifier, Fisher's linear discriminant, a support vector machine, or a k nearest neighbor classifier [9,12], is constructed using train data: example images and their associated segmentations should be provided beforehand from which the classifier learns how to map a given feature vector to a certain class label.

Before training the classifier, a reduction of dimensionality can be performed using the train data. This is done by means of a linear projection $\mathbf{L}$ from n to d ($d < n$) dimensions, which can be seen as a $d \times n$-matrix that is applied to the n-dimensional feature vectors $\mathbf{x}_i$ to get a d-dimensional feature representation $\mathbf{L}\mathbf{x}_i$. The matrix $\mathbf{L}$ is determined using the train data. Subsequently, the feature vectors of the train data are transformed to the lower dimensional feature vectors and the classifier is constructed using these transformed feature vectors. The following section presents a novel way to determine such a matrix $\mathbf{L}$.

3 Canonical Contextual Correlation Projections

3.1 Linear Discriminant Analysis

The classical approach to supervised linear dimensionality reduction is based on LDA. This approach defines the optimal transformation matrix $\mathbf{L}$ to be the one that maximizes the so-called Fisher criterion J

$$J(\mathbf{L}) = \mathrm{tr}((\mathbf{L}\mathbf{S}_W\mathbf{L}^t)^{-1}\mathbf{L}\mathbf{S}_B\mathbf{L}^t), \tag{1}$$

where $\mathbf{L}$ is the $d \times n$ transformation matrix, $\mathbf{S}_W$ is the mean within-class covariance matrix, and $\mathbf{S}_B$ is the between-class covariance matrix. The $n \times n$-matrix $\mathbf{S}_W$ is a weighted mean of class covariance matrices and describes the (co)variance that is (on average) present within every class. The $n \times n$-matrix $\mathbf{S}_B$ describes the covariance present between the several classes. In Equation (1), $\mathbf{L}\mathbf{S}_W\mathbf{L}^t$ and $\mathbf{L}\mathbf{S}_B\mathbf{L}^t$ are the $d \times d$ within-class and between-class covariance matrices of the feature data after reducing the dimensionality of the data to d using the linear transform $\mathbf{L}$.

When maximizing (1), one simultaneously minimizes the within-class covariance and maximizes the between-class covariance. The criterion tries to determine a transform $\mathbf{L}$ that maps the feature vectors belonging to one and the same class as close as possible to each other, while trying to keep the vectors that do not belong to the same class as far from each other as possible. The matrix that does so in the optimal way, as defined by (1), is the transform associated to LDA.

Once the covariance matrices $\mathbf{S}_W$ and $\mathbf{S}_B$ have been estimated from the train data, the maximization problem in (1) can be solved by means of a generalized eigenvalue decomposition involving the matrices $\mathbf{S}_B$ and $\mathbf{S}_W$. We do not discuss these procedures here, but refer to [3,4,7] and [9].

3.2 Canonical Correlation Analysis

This paper formulates LDA in a canonical correlation framework (see [9,16]) which enables the extension of LDA to CCCP. CCA is a technique to extract, from two feature spaces, those lower-dimensional subspaces that exhibit a maximum mutual correlation [11,2].

To be more precise, let X be a multivariate random variable, e.g. a feature vector, and let Y be another multivariate random variable, e.g. a numeric representation of the class label: $(1,0,\ldots,0)^t$ for class 1, $(0,1,\ldots,0)^t$ for class 2, etc. In addition, let $\mathbf{a}$ and $\mathbf{b}$ be vectors (linear transformations) having the same dimensionality as X and Y, respectively. Furthermore, define c to be the correlation between the univariate random variables $\mathbf{a}^tX$ and $\mathbf{b}^tY$, i.e.,

$$c = \frac{\mathsf{E}(\mathbf{a}^tX\mathbf{b}^tY)}{\sqrt{\mathsf{E}((\mathbf{a}^tX)^2)\mathsf{E}((\mathbf{b}^tY)^2)}}, \tag{2}$$

where E is the expectation. The first canonical variates $\mathbf{a}_1^tX$ and $\mathbf{b}_1^tY$ are obtained by those two vectors $\mathbf{a}_1$ and $\mathbf{b}_1$ that maximize the correlation in Equation (2). The second canonical variates are those variates that maximize c under the additional constraint that they are outside the subspace spanned by $\mathbf{a}_1$ and $\mathbf{b}_1$, respectively. Having the first two pairs of canonical variates, one can construct the third, by taking them outside the space spanned by $\{\mathbf{a}_1, \mathbf{a}_2\}$ and $\{\mathbf{b}_1, \mathbf{b}_2\}$, etc.

One way of solving for the canonical variates more easily is as follows. First estimate the matrices $\mathbf{S}_{XX}$, $\mathbf{S}_{YY}$, and $\mathbf{S}_{XY}$, that describe the covariance for the random variables X and Y, and the covariance between these variables, i.e., estimating $\mathsf{E}(XX^t)$, $\mathsf{E}(YY^t)$, and $\mathsf{E}(XY^t)$, respectively. Subsequently, determine the eigenvectors $\mathbf{a}_i$ of

$$\mathbf{S}_X := \mathbf{S}_{XX}^{-1}\mathbf{S}_{XY}\mathbf{S}_{YY}^{-1}\mathbf{S}_{XY}^t \tag{3}$$

and the $\mathbf{b}_j$ of

$$\mathbf{S}_Y = \mathbf{S}_{YY}^{-1}\mathbf{S}_{XY}^t\mathbf{S}_{XX}^{-1}\mathbf{S}_{XY} \,. \tag{4}$$

The two eigenvectors $\mathbf{a}_1$ and $\mathbf{b}_1$ associated with the largest eigenvalues of the matrices $\mathbf{S}_X$ and $\mathbf{S}_Y$, respectively, are the vectors giving the first canonical variates $\mathbf{a}_1^t X$ and $\mathbf{b}_1^t Y$. For the second canonical variates take the eigenvectors $\mathbf{a}_2$ and $\mathbf{b}_2$ with the second largest eigenvalues associated, etc. The number of canonical variates that can be obtained is limited by the smallest rank of both multivariate random variables considered.

3.3 LDA through CCA

LDA can be defined in terms of CCA (see for example [9] or [16]), hence avoiding the use of the Fisher criterion (1). To do so, let X be the random variable describing the feature vectors and let Y describe the class labels. Without loss of generality, it is assumed that X is centered, i.e., $\mathsf{E}(X)$ equals the null vector. Furthermore, as already suggested in Subsection 3.2, the class labels are numerically represented as K-dimensional standard basis vectors: for every class one basis vector.

Performing CCA on these random variables using $\mathbf{S}_X$ from (3), one obtains eigenvectors $\mathbf{a}_i$ that span the space (or part of this space) of n-dimensional feature vectors. A transformation matrix $\mathbf{L}$, equivalent to the one maximizing the Fisher criterion, is obtained by taking the d eigenvectors associated to the d largest eigenvalues and putting them as row-vectors in the transformation matrix:

$$\mathbf{L} = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_d)^t \,.$$

Linear dimensionality reduction performed with this transformation matrix gives results equivalent to classical LDA. Note that to come to this solution, an eigenvalue decomposition of $\mathbf{S}_Y$ is not needed.

The estimates of the covariance matrices used later on in our experiments are the well-known maximum likelihood estimates. Given N pixels p_i in our train data set, and denoting the numeric class label representation of pixel p_i by the K-dimensional vector $\mathbf{y}_i$, $\mathbf{S}_{XY}$ is estimated by the matrix

$$\frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i \mathbf{y}_i^t \,.$$

$\mathbf{S}_{XX}$ and $\mathbf{S}_{YY}$ are estimated in a similar way.

The CCA formulation of LDA enables us to extend LDA to a form of correlation analysis that takes the spatial structure of the class labelling in the neighborhood of the pixels into account.

3.4 Incorporating Spatial Class Label Context

In image processing, incorporating spatial gray level context into the feature vector is readily done by not only considering the actual gray level in that pixel as a feature, but by taking additional gray levels of neighboring pixels into account, or by adding large-scale filter outputs to the feature vector. However, on the class label side there is also contextual information available. Although two pixels could belong to the same class—and thus have the same class label, the configuration of class labels in their neighborhood can differ very much. LDA and other dimension reduction techniques, do not take into account this difference in spatial configuration, and only consider the actual label of the pixel.

The trivial way to incorporate these differences into LDA would be to directly distinguish more than K classes on the basis of these differences. Consider for example the 4-neighborhood label configurations in Figure 1. In a $K = 2$-class case, this 4-neighborhood could attain a maximum of $2^5 = 32$ different configurations (of which only four are displayed in the figure). These could then be identified as being different classes. Say we have M of them, then every configuration possible would get its own unique M-dimensional standard basis vector (as in Subsection 3.3) and one could subsequently perform LDA based on these classes, in this way indirectly taking more than a single class label into account when determining a dimension reducing matrix $\mathbf{L}$.

However, identifying every other configuration with a different class seems too crude. When two neighborhood label configurations differ in only a single pixel label, they should be considered more similar to each other then two label configurations differing in half of their neighborhood. Therefore, in our CCCP approach, a class label vector $\mathbf{y}_i$ is not encoded as a null vector with a single one in it, 1.e., a standard basis vector, but as a 0/1-vector in which the central pixel label and every neighboring label is encoded as a K-dimensional (sub)vector. Returning to our 2-class example from Figure 1, the

$$\begin{array}{ccc} & l_1 & \\ l_1 & l_1 & l_1 \\ & l_1 & \end{array} \quad \text{(a)} \qquad \begin{array}{ccc} & l_2 & \\ l_2 & l_2 & l_2 \\ & l_2 & \end{array} \quad \text{(b)}$$

$$\begin{array}{ccc} & l_2 & \\ l_1 & l_1 & l_2 \\ & l_1 & \end{array} \quad \text{(c)} \qquad \begin{array}{ccc} & l_1 & \\ l_2 & l_1 & l_1 \\ & l_2 & \end{array} \quad \text{(d)}$$

Fig. 1. Four possible class labellings in case a four-neighborhood context is considered. For this two-class problem the total number of possible contextual labellings equals $2^5 = 32$.

four label vectors that would give the proper encoding of the class labelling within these 4-neighborhoods (a), (b), (c), and (d) are

$$\begin{pmatrix}1\\0\\1\\0\\1\\0\\1\\0\\1\\0\end{pmatrix}, \begin{pmatrix}0\\1\\0\\1\\0\\1\\0\\1\\0\\1\end{pmatrix}, \begin{pmatrix}0\\1\\1\\0\\1\\0\\0\\1\\1\\0\end{pmatrix}, \text{and} \begin{pmatrix}1\\0\\0\\1\\1\\0\\1\\0\\0\\1\end{pmatrix}. \tag{5}$$

The five pixels (the four pixels in the neighborhood and the central pixel) are traversed left to right and top to bottom. So the first two entries of the four vectors correspond to the labelling of the top pixel and the last two entries correspond to the bottom pixel label.

Note that the label vectors are 10-dimensional, i.e., per pixel from the neighborhood (five in total) a sub-vector of size two is used to encode the two possible labellings per pixel. In general, if P is the number of pixels in the neighborhood including the central pixel, these (KP)-dimensional vectors contain P ones, and $(K-1)P$ zeros, because every pixel belongs to exactly one of K classes, and every pixels is thus represented by a K-dimensional sub-vector. In the foregoing example where $K = 2$ and $P = 5$, there are 5 ones and 5 zeros in the complete vector, and 1 one and 1 zero per sub-vector.

When taking the contextual label information into account in this way, gradual changes in the neighborhood structure are appreciated. In Figure 1, configurations (a) and (b) are as far from each other as possible (in terms of e.g. Euclidean or Hamming distance, cf. the vectors in (5)), because in going from one configuration to the other, all pixel sites have to change their labelling. Comparing a different pair of labellings from Figure 1 to each other, we see that their distance is less than maximal, because it needs less permutations to turn one contextual labelling into the other.

We propose the numeric class label encoding described above for incorporating contextual class label information into the CCA, resulting in the canonical correlation projection, CCCP, that can explicitly deal with gray value context—through the feature vectors $\mathbf{x}_i$—as well as with class label context—through our numeric class label encoding represented by the vectors $\mathbf{y}_i$. Note that CCCP encompasses classical LDA. Taking no class label context into account but only the class label of the central pixel clearly reduces CCCP to LDA.

3.5 Reduction to More than $K-1$ Dimensions

We return to one of the main drawbacks of LDA already mentioned: the fact that LDA cannot reduce the dimensionality to more than $K-1$, i.e., the number of classes minus 1. In many segmentation tasks K is not higher than 2 or 3, in which case LDA can only extract 1 or 2 dimensions. Starting with a high-dimensional image feature space, it is hardly to be expected that all relevant information is captured in this subspace.

The CCCP alleviates this limitation. The maximum number of canonical variates that can be extracted through CCA equals $\min\{\mathrm{rank}(\mathbf{S}_X), \mathrm{rank}(\mathbf{S}_Y)\}$. When dealing with as many as or fewer classes than the feature dimensionality, i.e., $K \leq n$, the limiting factor in the dimensionality reduction using LDA is the matrix $\mathbf{S}_Y$ which rank is equal to (or even smaller than) $K - 1$. However, by extending the class label context, the rank of $\mathbf{S}_Y$ increases and can even get larger than $\mathrm{rank}(\mathbf{S}_X)$.

So in general, CCCP can provide more canonical variates than classical LDA by incorporating more class label context. And consequently, for CCCP the resulting feature dimensionality can be larger than $K - 1$. In the experiments in Section 4, it is shown that this can significantly improve the segmentation results.

3.6 The CCCP Algorithm

The CCCP technique is summarized. A reduction of n-dimensional image data to d dimensions is considered.

- define what (contextual) image feature information to use (e.g. which filters), and which neighboring pixels to take for the class label context
- determine from the train images and associated segmentations the gray level feature vectors $\mathbf{x}_i$
- determine from the same data the class label feature vectors $\mathbf{y}_i$, i.e., determine for every pixel in the context its standard basis vector describing its class label and concatenate all these vectors
- estimate the covariance matrices $\mathbf{S}_{XX}$, $\mathbf{S}_{XY}$, and $\mathbf{S}_{YY}$
- do an eigenvalue decomposition of the matrix $\mathbf{S}_X := \mathbf{S}_{XX}^{-1}\mathbf{S}_{XY}\mathbf{S}_{YY}^{-1}\mathbf{S}_{XY}^t$ from (3)
- take the d rows of the $d \times n$ linear dimension reducing transformation matrix $\mathbf{L}$ equal to the d eigenvectors associated to the d largest eigenvalues
- transform all $\mathbf{x}_i$ using $\mathbf{L}$ to $\mathbf{L}\mathbf{x}_i$

4 Illustrative Experiments

This section exemplifies the theory by a simple illustrative example. The section is not intended to present a full-fledged state-of-the-art solution to the task, but merely to illustrate the possible improvements in performance when employing the CCCP instead of the original LDA or no dimensionality reduction at all. For this reason, the task considered is a lung field segmentation task in chest radiographs, which is based on a simple pixel classification technique. A segmentation scheme solving this problem properly may be based on snakes, active shape models, or some kind of Markov random field, taking more global contextual and/or geometric information into account (cf. [8]).

4.1 Chest Radiograph Data

The data used in the experiments consists of 20 standard PA chest radiographs taken from a tuberculosis screening programm. The size of the sub-sampled and digitized images

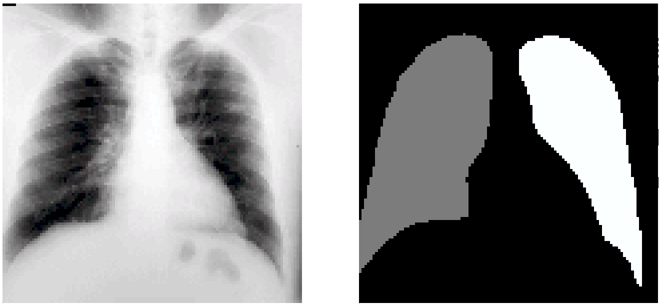

Fig. 2. The left image displays a typical PA chest radiograph as used in our experiments. The right image shows its expert lung field segmentation. The background is black, both lung fields are in different shades of gray.

equals 128×128. An examples of a typical chest radiographs is shown in Figure 2. The task is to segment, both lung fields.

In addition to the radiographs, the associated ground truth is given, i.e., in these images, the the lung fields are manually delineated by an expert and the delineation is converted to a 3-class pixel labelling. An example image is given in Figure 2 also.

4.2 Experimental Setup

In all experiment, 10 images were used for training and 10 for testing. The total number of feature vectors equals $20 \cdot (128 - 12)^2 = 269,120$ and train and test set both contain half of it. Note that pixel within a distance of 6 pixels from the border are not taken into account to avoid boundary problems in building up the contextual gray level features (see below).

Experiments were conducted using a nonparametric 1 nearest neighbor (1NN) classifier. We chose to use a 1NN classifier for its simplicity and because it offers suitable baseline results which makes a reasonable comparison possible [3,7,12]. Before the 1NN classifier was trained, the within-class covariance matrix $\mathbf{S}_W$ was whitened (cf. Subsection 3.1) based on the train data [7].

The variables in our experiments were the contextual class label information, and the dimensionality d to which the data is to be reduced. The contextual class label information belonging to one pixel p_i is defined by all pixels coming from within a radius of r pixels from p_i. Experiments were performed with $r \in \{0, 2, 3\}$; $r = 0$ means that only the central label belonging to p_i is taken into account (equal to classical LDA), $r = 2$ results in 13 contextual labels, and $r = 3$ in 29 contextual labels.

As contextual image features, we simply took the gray levels from neighboring pixels into account, so no filtering or other preprocessing is performed. The contextual information of pixel p_i consisted of all raw gray values within a radius of 5 from this pixel. In addition, the x and y coordinates were added to the image feature vector, which final dimensionality totals 81 + 2 = 83. (Choosing to set the radius for the contextual gray level information to 5 is based on a small pilot experiment using LDA. LDA performed best with these settings.)

The dimensionality d to reduce to were in the set $\{1,2,4,11,22,37,56,83\}$. N.B. setting d equal to 83 means no dimensionality reduction is performed.

Using the aforementioned d, image features and contextual class label features, the train set was used for determining the CCCP and training the 1NN classifier. Subsequently, using the test set, we determined the pixel classification error.

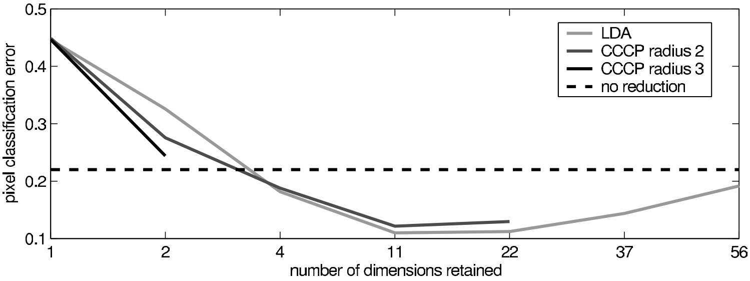

Fig. 3. The black dashed horizontal line indicates the performance of the pixel classification scheme if no dimensionality reduction is employed and the full 83-dimensional feature vector is used in the segmentation. The black solid line is the classification error obtained when using LDA. The gray lines give the performance for the two different instances of CCCP. The dark gray line uses a contextual radius of 2, while the light gray line uses a radius of 3. Their pixel classification error is plotted against feature dimensionality d. The optimal classification errors are 0.22, 0.24, 0.12, and 0.11, respectively.

4.3 Results

Figure 3 gives the results obtained by LDA, CCCP and no dimensionality reduction. Note that for LDA (solid black line), the dimensionality can only be reduced to 1 or 2, because the number of classes K is 3 (i.e., left lung field, right lung field, or background). Note also the peaking behavior [4,12] that is visible in the plots of the CCCP results.

Both instances of CCCP clearly outperforms LDA and they give a dramatic improvement over performing no dimensionality reduction as well. It should be noted, though, that CCCP does not outperform LDA for every dimensionality d.

Figure 4 gives for the example image in Figure 2 the segmentation obtained by the optimal LDA (left), the segmentation obtained by the optimal CCCP (middle), and the one obtained using no reduction (right). Comparing the three images, the main observations is that the CCCP-based segmentation gives much more coherent results than the other segmentations. Furthermore, there seems to be less confusion between left and right lung fields when CCCP is employed. The background classification error in comparison with the result in the right image, however, seems to go up a bit when using the CCCP approach. In the right image there are no misclassified background pixels, in both other images there are.

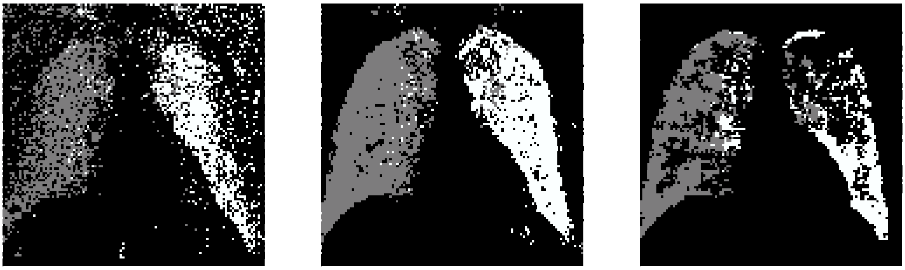

Fig. 4. The segmentation with optimal LDA ($d = 2$) is depicted on the left, the one with optimal CCCP in the middle ($d = 11$ and $r = 3$), and on the right is the segmentation obtained using no dimensionality reduction.

5 Discussion and Conclusions

In this work we extended classical LDA—as a dimensionality reduction technique—to incorporate the spatial contextual structure present in the class labelling. Our extension, called the canonical contextual correlation projection (CCCP), is based on a canonical correlation formulation of LDA that enables the encoding of these spatial class label configurations. Experiments on the task of segmenting the lung fields in chest radiographs demonstrated that in this way significant improvement over LDA or no dimension reduction is possible. Furthermore, these experiments show also that using a data-driven method for image segmentation—in which the dimension reduction is an essential part, good results can be obtained without the additional utilization of task-dependent knowledge. We expect that similar results hold in, for example, object detection, object classification or some other discriminative tasks in which CCCP can also be used to determine low-dimensional but still discriminative features.

Clearly, regarding the experiments, improving the segmentation results should be possible. For example, by using more complex pattern recognition techniques that can also handle contextual class label information in their classification scheme. Typically, such scheme employs a Markov random field approach, or something closely resembling this [10,15,17]. Here CCCP could also be a valuable tool in another way. Due to the iterative nature of these schemes they often are rather slow. In part, this may be attributed to the large contextual neighborhoods that are taken into account. Lowering the dimensionality of these neighborhoods can, in addition to improving the error rate, speed up the iterative process considerably.

An interesting way to further improve the dimensionality reduction scheme is the development of nonlinear CCCP. This is for example possible via a CCA-related technique called optimal scoring [9], which is, among other things, used for extending LDA to nonlinear forms. Nonlinear dimensionality reduction can of course lead to a better lower-dimensional representation of the image data, however the nonlinearity often makes such approaches computationally hard. Nonetheless, CCCP does (via CCA) provide a proper framework for these kind of extensions.

In conclusion, CCCP provides a general framework for linearly reducing contextual feature data in a supervised way, it is well capable of improving LDA and can be extended in several directions. It generalizes LDA by not only taking gray level context into account, but incorporating contextual class label information as well. In a small segmentation experiment, it was shown that CCCP can clearly give improvement performance compared to LDA and no dimensionality reduction.

References

1. P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.
2. M. Borga. *Learning Multidimensional Signal Processing*. Ph.D. Thesis, Linköping University, Sweden, 1998.
3. P. A. Devijver and J. Kittler. *Pattern Recognition: a Statistical Approach*. Prentice-Hall, London, 1982.
4. R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley & Sons, New York, second edition, 2001.
5. R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.
6. R. A. Fisher. The statistical utilization of multiple measurements. *Annals of Eugenics*, 8:376–386, 1938.
7. K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, New York, 1990.
8. B. van Ginneken, B. M. ter Haar Romeny, and M. A. Viergever. Computer-aided diagnosis in chest radiography: A survey. *IEEE Transactions on Medical Imaging*, 20(12):1228–1241, 2001.
9. T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer-Verlag, New York . Berlin . Heidelberg, 2001.
10. N. L. Hjort and E. Mohn. A comparison in some contextual methods in remote sensing classification. In *Proceedings of the 18th International Symposium on Remote Sensing of Environment*, pages 1693–1702, Paris, France, 1984. CNES.
11. H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:321–377, 1936.
12. A. K. Jain, R. P. W. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, 2000.
13. K. Liu, Y.-Q. Cheng, and J.-Y. Yang. Algebraic feature extraction for image recognition based on an optimal discriminant criterion. *Pattern Recognition*, 26(6):903–911, 1993.
14. C. R. Rao. The utilization of multiple measurements in problems of biological classification. *Journal of the Royal Statistical Society. Series B*, 10:159–203, 1948.
15. J. A. Richards, D. A. Landgrebe, and P. H. Swain. Pixel labeling by supervised probabilistic relaxation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 3(2):188–191, 1981.
16. B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, 1996.
17. G. Winkler. *Image Analysis, Random Fields and Dynamic Monte Carlo Methods*. Number 27 in Applications of mathematics. Springer-Verlag, Berlin . Heidelberg, 1995.