



ELSEVIER

# Dimensionality reduction of image features using the canonical contextual correlation projection

Marco Loog<sup>a,\*</sup>, Bram van Ginneken<sup>b</sup>, Robert P.W. Duin<sup>c</sup>

<sup>a</sup>Image Analysis Group, Department of Innovation, IT University of Copenhagen, Rued Langgaards Vej 7, 2300 Copenhagen S, Denmark

<sup>b</sup>Image Sciences Institute, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX Utrecht, The Netherlands

<sup>c</sup>Information and Communication Theory Group, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, P.O. Box 5046 2600 GA Delft, The Netherlands

Received 26 July 2004; accepted 29 April 2005

## 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 linear discriminant analysis (LDA), extending this technique 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 readily be dealt with in standard LDA.) The novel method is useful, for example, in supervised segmentation tasks in which high-dimensional feature vectors describe the local structure of the image.

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 the 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 extended LDA, that takes spatial label context into account, is derived from a formulation of standard LDA in terms of canonical correlation analysis. The novel technique is called the canonical contextual correlation projection (CCCP).

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 medical image segmentation problem in which it is shown that it may give significant improvement in segmentation accuracy.

© 2005 Pattern Recognition Society. Published by Elsevier Ltd. All rights reserved.

**Keywords:** Dimensionality reduction; Feature extraction; Contextual information; Canonical correlation analysis; Image data

## 1. Introduction

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) is presented. Besides contextual information from the input features, the dimension reducing technique can also take contextual label information into account (e.g. the local class label configuration in a segmentation task). The

\* Corresponding author. Tel.: +45 72 18 5072;  
fax: +45 72 18 5001.

E-mail addresses: [marco@itu.dk](mailto:marco@itu.dk) (M. Loog), [bram@isi.uu.nl](mailto:bram@isi.uu.nl) (B. van Ginneken), [r.p.w.duin@ewi.tudelft.nl](mailto:r.p.w.duin@ewi.tudelft.nl) (R.P.W. Duin)

URL: <http://www.itu.dk/people/marco/> (M. Loog).

<sup>1</sup> Larger parts of the work were carried out while M. Loog was affiliated to the Image Sciences Institute, Utrecht, The Netherlands.

technique is based on canonical correlation analysis (CCA) and dubbed the canonical contextual correlation projection (CCCP). The work presented is a fully revised version of Ref. [1].

Generally, the main goal of reducing the dimensionality of feature data, which is also called feature extraction, is to prevent the subsequently used model from over-fitting in the training phase [2,3]. 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 using label context 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 tasks the dimensionality, say  $n$ , of the feature data can be relatively large. Consider for example image analysis or computer vision tasks in which it is often not clear a priori what image information is needed for a good performance. As a consequence, focusing on supervised pixels classification tasks, many features per pixel may be included in the analysis, resulting in a high-dimensional feature vector. This already happens in two-dimensional image processing, but when processing large hyper-spectral images, medical three-dimensional volumes, or four-dimensional space/time image data, it may even be less clear what features to take into account and consequently even more features are added. However, high-dimensional data often leads to inferior results due to the curse of dimensionality [3] even if all relevant information for accurate classification is contained in the feature vector. Hence, lowering the dimensionality of the feature vectors in an appropriate way can lead to a significant gain in performance and mainly for this reason dimensionality reduction techniques have been developed.

The CCCP is an extension to linear discriminant analysis (LDA). The latter is a basic, well-known, and useful supervised dimensionality reduction technique from statistical pattern recognition [2,3]. 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. This class label context coming from the spatial configuration of images provides an additional source of classification information and therefore taking this contextual information into account can be beneficial.

The CCCP does take this latter information into account. Instead of associating a single output class with each sample, the output of the sample together with the output of neighboring samples is encoded in a multi-dimensional output vector. A simple coding scheme is proposed that maps similar neighborhoods to nearby positions in the output space. Subsequently, a CCA is performed employing these pairs of input and output vectors. In the limit of a neighborhood of zero size, this is equivalent to classical LDA.

Another principal drawback of LDA is that it cannot extract more features than the number of classes minus one [2,4]. In the two-class case—often encountered in image segmentation, e.g. object versus background—this means that one can reduce the dimensionality of the data merely to one, and even though this could improve the performance it is not plausible that one single feature can describe class differences accurately. The CCCP can avoid such extreme deterioration of the classification space and is able to 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 [7] 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 other supervised linear dimension reduction techniques have been proposed of which many can be interpreted as variations and extensions to LDA (see [2,4,8–10]). Within the field of image classification, in which the whole image is given a single label, e.g. in face or character recognition, Belhumeur et al. [11] and Liu et al. [12] show how classification performance can benefit from linear dimensionality reduction.

The novel extension to LDA given in this paper explicitly deals with the spatial contextual characteristics of image data. To come to this extension of LDA, a formulation of this technique in terms of CCA [13] is used (see Refs. [2,10]), which enables us to not only include the class labels of the pixel that is considered—as in classical LDA, but also to encode information from the surrounding class label structure. We are not aware of any other dimensionality reduction technique that takes such spatial label information into account and we expect that the principal idea presented in this paper may also be applicable in most of the other supervised dimension reducing techniques from Refs. [2,4,8–10]. We briefly return to this latter topic in Section 5.

### 1.1. Outline

The remainder of this article is as follows. Section 2 formulates the general problem within the context of pixel-based supervised image segmentation. Section 3 introduces LDA and discusses its link to CCA. Section 4.3 introduces the CCCP. Section 4 presents an illustrative example on medical image segmentation task in which the heart, the lung fields, and both clavicles are to be segmented within standard chest radiographs. Finally, Section 5 provides a discussion and conclusions.

## 2. Supervised image segmentation

Image segmentation in terms of pixel classification is considered. Based on one or more 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, and

thus having labelled all of them, 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 level outputs after linear or non-linear filtering of the image, etc. Furthermore, when dealing with full-color, multi-band, or hyperspectral images, features extracted from one or more different bands may be included as well.

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 class label coming from a set of  $K$  possibilities:  $\{\ell_1, \dots, \ell_K\}$ . All pixels having the same label belong to the same segment. The classifier is constructed using train data, i.e., example images and their associated segmentations are 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. This paper presents a novel way to determine such a matrix  $\mathbf{L}$ . Before doing so, the next section discusses standard LDA and a straightforward way to introduce extra information into the mapping  $\mathbf{L}$  using contextual output information.

### 3. LDA and a direct approach to incorporating context

#### 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  $J$  [2,4,8]:

$$\mathbf{L} = \underset{\mathbf{A}}{\operatorname{argmax}} J(\mathbf{A}) \quad (1)$$

with

$$J(\mathbf{A}) = \operatorname{tr}((\mathbf{A}\mathbf{S}_W\mathbf{A}^t)^{-1}\mathbf{A}\mathbf{S}_B\mathbf{A}^t), \quad (2)$$

where  $\mathbf{A}$  is a  $d \times n$  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 Eq. (2),  $\mathbf{A}\mathbf{S}_W\mathbf{A}^t$  and  $\mathbf{A}\mathbf{S}_B\mathbf{A}^t$  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  $\mathbf{A}$ .

When maximizing Eq. (2), one simultaneously minimizes the within-class covariance and maximizes the between-class covariance in the lower-dimensional space which is spanned by the rows of  $\mathbf{A}$ . 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 optimally, as defined by Eq. (2), 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 Eq. (2) can be solved by means of a generalized eigenvalue decomposition—related to maximizing a generalized Rayleigh quotient—involving the matrices  $\mathbf{S}_B$  and  $\mathbf{S}_W$  (see Refs. [2,4,8,14]). The eigenvalue problem to be solved is

$$\mathbf{S}_B\mathbf{V} = \mathbf{S}_W\mathbf{V}\Lambda \quad (3)$$

or equivalently

$$\mathbf{S}_W^{-1}\mathbf{S}_B\mathbf{V} = \mathbf{V}\Lambda \quad (4)$$

in which  $\mathbf{V}$  is an  $n \times n$  matrix consisting of  $n$  eigenvectors (as column vectors) and  $\Lambda$  is an  $n \times n$  diagonal matrix with the  $n$  eigenvalues  $\lambda_i$  associated to the eigenvectors  $\mathbf{v}_i$  in  $\mathbf{V}$  on the diagonal. A  $d \times n$  transformation matrix  $\mathbf{L}$  that maximizes the Fisher criterion is obtained by setting the rows of  $\mathbf{L}$  equal to the  $d$  transposed eigenvectors  $\mathbf{v}_i^t$  corresponding to the  $d$  largest eigenvalues.

#### 3.2. Incorporating spatial class label context: direct approach

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. Another option is to add 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 straightforward 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 Fig. 1. In a  $K=2$ -class case, this 4-neighborhood could attain a maximum of  $2^5 = 32$  different configurations (of which four possibilities are displayed in the figure). These could then be considered as being different classes. Say there are  $M$  of

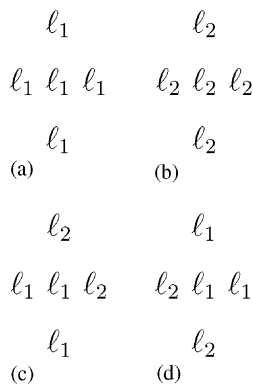


Fig. 1. Four possible class label configurations in case a four-neighborhood context is considered. For this two-class problem the total number of possible contextual configurations equals  $2^5 = 32$ .

them, then every configuration possible would get its own unique class label from the set  $\{\ell_1, \dots, \ell_M\}$  and one could subsequently perform LDA based on this extended set of classes, in this way, indirectly based on contextual class label information, taking more than the initial  $K$  labels into account when determining a dimension reducing matrix  $\mathbf{L}$ .

One may now simply use the aforementioned approach and determine dimension reducing transforms based on the suggested idea, however, identifying every other configuration with a different class seems too crude. (Let alone that it may result in an huge increase of the number of possible class labels, especially when the output context becomes relatively large.) When two neighborhood label configurations differ in only a single pixel label, they should be considered more similar to each other than two label configurations differing in half of their neighborhood. This is not the case in the foregoing. Because two class label contexts are considered as different or not. The procedure is ignorant of the fact that being different can be defined in a more gradual way. The CCCP approach, which is presented in the next section, distinguishes these grades of dissimilarity and models them.

## 4. Canonical contextual correlation projections

### 4.1. Canonical correlation analysis

To begin with, LDA is formulated in a canonical correlation framework (see Refs. [2,10]) which eventually 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 [13].

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 via a  $K$ -dimensional standard basis vector:  $(1, 0, \dots, 0)^t$  for class 1,  $(0, 1, \dots, 0)^t$  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  $\mathbf{a}^t X$  and  $\mathbf{b}^t Y$ , i.e.,

$$c = \frac{E(\mathbf{a}^t X \mathbf{b}^t Y)}{\sqrt{E((\mathbf{a}^t X)^2)E((\mathbf{b}^t Y)^2)}}, \quad (5)$$

where  $E$  is the expectation. The first canonical variates  $\mathbf{a}_1^t X$  and  $\mathbf{b}_1^t Y$  are obtained by those two vectors  $\mathbf{a}_1$  and  $\mathbf{b}_1$  that maximize the correlation in Eq. (5). 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  $E(XX^t)$ ,  $E(YY^t)$ , and  $E(XY^t)$ , respectively. Subsequently, determine the eigenvectors  $\mathbf{a}_j$  of

$$\mathbf{S}_X := \mathbf{S}_{XX}^{-1} \mathbf{S}_{XY} \mathbf{S}_{YY}^{-1} \mathbf{S}_{XY}^t \quad (6)$$

and the  $\mathbf{b}_j$  of

$$\mathbf{S}_Y = \mathbf{S}_{YY}^{-1} \mathbf{S}_{XY}^t \mathbf{S}_{XX}^{-1} \mathbf{S}_{XY}. \quad (7)$$

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 one covariance matrix,  $\mathbf{S}_{XX}$  or  $\mathbf{S}_{YY}$ , having the smallest rank. Note that in case one of the aforementioned matrices is singular, one could use the Moore–Penrose inverse in Eqs. (6) and (7) instead of the standard inverse. Because both inverses coincide if the matrices are full-rank, in our experiments, we used the Moore–Penrose inverse in all cases.

### 4.2. LDA through CCA

LDA can be defined in terms of CCA (see for example Ref. [2] or Ref. [10]). 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.,  $E(X)$  equals the null vector. Furthermore, as already suggested in Section 4.1, 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 Eq. (6), one obtains eigenvectors  $\mathbf{a}_j$  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 Eq. (7) is not needed.

The estimates of the covariance matrices used 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 such that the amount of (dis)similarity between label contexts is respected.

#### 4.3. Incorporating spatial class label context: label vector concatenation

Recalling the discussion at the end of Section 3.2, it is noted that identifying every other label 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 than two label configurations differing in for example half of their neighborhood. Therefore, in our approach, using the CCA formulation, a class label vector  $\mathbf{y}_i$  is not encoded as a null vector with a single one (1) in it, i.e., a standard basis vector (which would be equivalent to LDA through CCA as discussed in the previous subsection). The CCCP technique uses a more general 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 Fig. 1, Section 3.2, the four label vectors that give the proper CCCP encoding of the class labelling within these 4-neighborhoods (a)–(d) are

$$\begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 1 \\ 0 \\ 1 \\ 1 \\ 0 \\ 1 \end{pmatrix}, \text{ and } \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \\ 1 \end{pmatrix}. \quad (8)$$

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 resulting 0/1-vectors consist of a concatenation of the standard basis vectors numerically describing the class labels of the individual pixels in the neighborhood. The label vectors are 10-dimensional: per pixel from the neighborhood (five in total) a sub-vector of size two is used to encode the two possible label configurations 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, for every pixel belongs to exactly one of  $K$  classes, and every pixels is thus represented by a  $K$ -dimensional sub-vector.

When taking the contextual label information into account in this way, gradual changes in the neighborhood structure are appreciated. In Fig. 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 Eq. (8)), because in going from one configuration to the other, all pixel sites have to change their labelling. Comparing a different pair of labellings from Fig. 1 to each other, one sees 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 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.

#### 4.4. Reduction to more than $K-1$ dimensions

We return to one of the main drawbacks of LDA mentioned in the Introduction: 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  $\min\{\text{rank}(\mathbf{S}_{XX}), \text{rank}(\mathbf{S}_{YY})\}$  [2, 10]. 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 smaller than,  $K-1$ . However, by extending the class label context, the rank of  $\mathbf{S}_Y$  increases and can even get larger than  $\text{rank}(\mathbf{S}_X)$ .

Therefore, in general, CCCP can provide more canonical variates than classical LDA by incorporating more class

label context. Consequently, 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.

#### 4.5. Dimensionality reduction by means of the CCCP

The CCCP technique is summarized. Considered is the reduction of  $n$ -dimensional image data to a  $d$ -dimensional subspace:

- 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 contextual class label feature vectors  $\mathbf{y}_i$ , i.e., determine for every pixel within the output context the standard basis vector that encodes its class label and concatenate all these vectors;
- determine the matrices  $\mathbf{S}_{XX}$ ,  $\mathbf{S}_{XY}$ , and  $\mathbf{S}_{YY}$ ;
- perform an eigenvalue decomposition of the matrix<sup>2</sup>  $\mathbf{S}_X := \mathbf{S}_{XX}^{-1} \mathbf{S}_{XY} \mathbf{S}_{YY}^{-1} \mathbf{S}_{XY}^t$ ;
- 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  $\mathbf{x}_i$ , both from the train and the test set to  $\mathbf{Lx}_i$ .

### 5. An illustrative example

This section exemplifies the theory, illustrating the possible improvements in performance when employing the CCCP instead of the original LDA. Results for performing no dimensionality reduction at all are also provided. The task considered is a segmentation task concerning chest radiographs. In these images, the heart, both lung fields, and both clavicles are to be segmented. The objective is to minimize the pixel classification error, i.e., the number of mislabelled pixels.

#### 5.1. Chest radiograph data

The data used in the experiments consist of 20 digital standard posteroanterior chest radiographs randomly taken from the JSRT database.<sup>3</sup> The size of the sub-sampled

<sup>2</sup> We note at this point that computational burden of our method is dominated by the complexity of the matrix inversions in Eq. (6). Since both the  $n \times n$ -matrix  $\mathbf{S}_{XX}$  and the  $KP \times KP$ -matrix  $\mathbf{S}_{YY}$  need to be inverted, this boils down to a computational complexity of  $O(\max\{n, KP\}^3)$ . This means that as long as  $n \leq KP$ , the computational complexity of CCCP is comparable to that of standard LDA.

<sup>3</sup> The JSRT database is a publicly available chest radiograph database [15].

images equals  $128 \times 128$ . An examples of a typical chest radiographs is shown in Fig. 2. In addition to the radiographs, the associated segmentation is given, i.e., in these images, the heart, the lung fields, and both clavicles are manually delineated and the delineation is converted to a 6-class pixel labelling. An example of such a segmentation is given in Fig. 2 as well.

#### 5.2. Experimental setup

In all experiments, 10 images were used for training and 10 for testing. The total number of feature vectors equals  $20(128 - 16)^2 = 250,880$  of which both train and test set contain half. Note that pixel within a distance of 8 pixels from the border are not taken into account to avoid boundary problems in building up the contextual gray level or label features.

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 [8,4,3]. Before the 1NN classifier was trained, the within-class covariance matrix  $\mathbf{S}_W$  was whitened (cf. Section 3.1) based on the train data [4], i.e., the within-class covariance matrix is linearly transformed to the identity matrix.

As contextual image features, we simply took the gray levels from neighboring pixels into account, so no filtering or other preprocessing is performed. This contextual information of pixel  $p_i$  consisted of all raw gray values within a radius of 6 from this pixel. In addition, the  $x$  and  $y$  coordinates were added to the image feature vector, which final dimensionality totals  $113 + 2 = 115$ . Choosing to set the radius for the contextual gray level information,  $\gamma$ , to 6 is based on a small pilot experiment using LDA. Taking smaller values for  $\gamma$  resulted in a worse performance for LDA. Increasing  $\gamma$  further gave very little improvement in terms of accuracy and therefore  $\gamma$  is set to 6.

We note that there is of course a multitude of other choices possible regarding the initial pixel features. However, irrespective of the choice of initial features, CCCP provides the possibility of also incorporating contextual label information into the linear dimensionality reduction scheme, and as such the general capability of improving the performance of LDA.

The variables in our experiments are 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—similar to the gray value features—by all pixels coming from within a radius of  $\lambda$  pixels from  $p_i$ . Experiments were performed with  $\lambda \in \{0, \dots, 7\}$ . Taking  $\lambda = 0$  means that only the central label belonging to  $p_i$  is taken into account in which case CCCP equals classical LDA. For  $\lambda$  increasing from 1 to 7, the number of contextual labels are 5, 13, 29, 49, 81, 113, and 149, respectively. The dimensionality  $d$  to reduce to were in

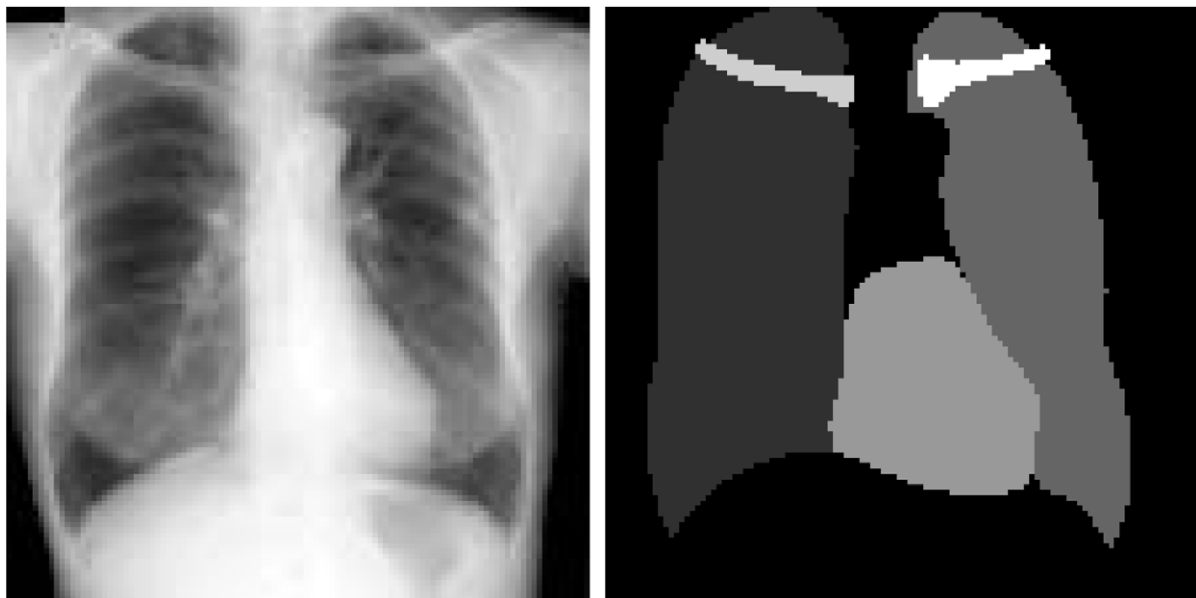


Fig. 2. Left: A typical posteroanterior chest radiograph as used in our experiments. Right: its corresponding segmentation. Background is black, left clavicle is white, and the four other segments—heart, lungs, and right clavicle—are in different shades of gray.

the set  $\{1, 3, 5, 10, 19, 35, 65, 115\}$ . Setting  $d$  equal to 115 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, a leave-one-out estimate of the classification error is obtained using the train set and in addition the pixel classification error is estimated based on the test set.

### 5.3. Results

Fig. 3 gives the results obtained by LDA, CCCP and no dimensionality reduction. Note that for LDA, the dimensionality can only be reduced to a maximum of 5 dimensions, because the number of classes  $K$  is 6. Note also the peaking behavior [3] that is visible in the plots of the CCCP results and the difference in error estimates for moderate dimensionalities when comparing corresponding left and right subfigures.

All instances of the CCCP clearly outperform LDA for certain dimensionalities  $d$ . Additionally, they give a dramatic improvement over performing no dimensionality reduction as well. It should be noted though, that the CCCP does not outperform LDA for every (fixed) dimensionality  $d$  as can be seen from Table 1 in which for  $\lambda \in \{0, 2, 4, 6\}$  all error estimates at  $d = 1, 3$ , and 5 are provided. LDA performs best when reducing to a single dimension. However, in most other cases CCCP seems to be the better one.

The optimal leave-one-out errors are 0.142 for LDA ( $d = 5$ ), 0.037 for CCCP ( $\lambda = 7, d = 19$ ), and 0.242 for no dimensionality reduction. The optimal classification errors on the

test set are 0.228 for LDA ( $d = 5$ ), 0.128 for CCCP ( $\lambda = 7, d = 19$ ), and 0.244 for no dimensionality reduction.

For the example image in Fig. 2, Fig. 4 gives the segmentation obtained after optimal LDA (left), the segmentation obtained using the CCCP ( $\lambda = 7$ ), and the one obtained using no reduction (right). Comparing the three images, it is readily perceived that the CCCP-based segmentation gives much more coherent results and better defined segment boundaries than the other segmentations. In addition to the actual segmentations, Fig. 4 shows images that merely indicate whether or not a pixel was misclassified. In these images, it may be easier to observe that the classification result obtained employing the CCCP is preferable over the other two.

## 6. Discussion and conclusions

In this work, the classical dimensionality reduction method linear discriminant analysis (LDA) is extended to incorporate spatial contextual structure present in the class labels. 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 a specific radiograph segmentation task 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—of which the dimensionality

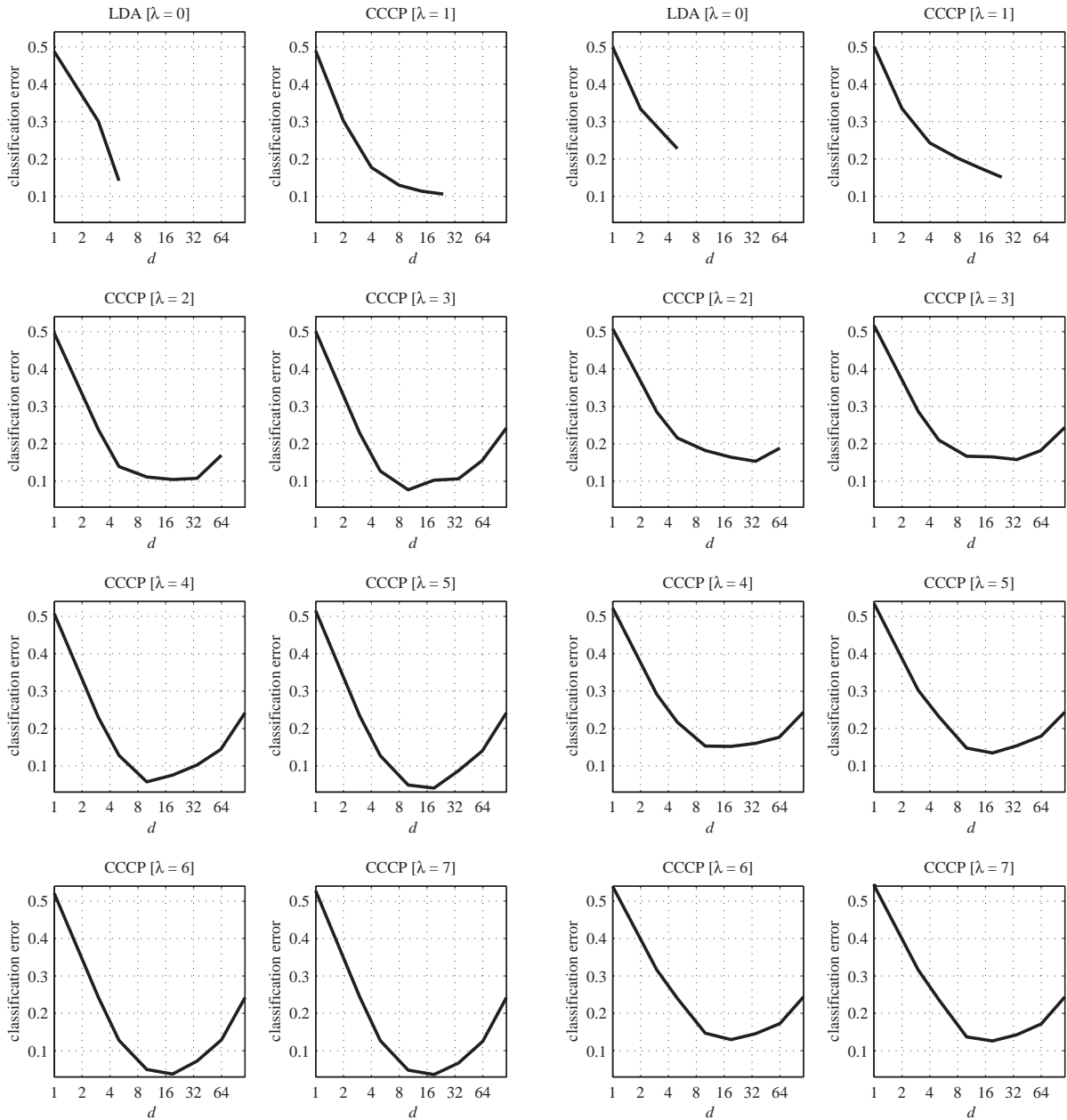


Fig. 3. The plots on the left give the leave-one-out estimates for the classification error vs the reduced dimensionality  $d$  for the eight different choices of  $\lambda$  (the value of  $\lambda$  is given above every subplot). The plots on the right give the error estimates based on the test data.

reduction is an essential part, very reasonable 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.

Regarding the experiments, it is on the other hand clear that further improvement of the segmentation results is pos-

sible. One could start by simply using a  $k$ NN classifier (instead of a 1NN) and determine an optimal  $k$ . More probably one may want to build a more intricate and dedicated classifier for the task at hand. Further improvements might then be obtained by using techniques that can also handle contextual class label information directly in their classification scheme. Typically, these latter schemes employ a Markov random field approach or something closely resembling this [16].

Table 1  
Error estimates for the dimensionalities  $d$  equal to 1, 3, and 5, and for  $\lambda$  set equal to 0 (LDA), 2, 4, and 6

$d$	Leave-one-out error				Error based on test data			
	LDA	CCCP			LDA	CCCP		
		$\lambda = 2$	$\lambda = 4$	$\lambda = 6$		$\lambda = 2$	$\lambda = 4$	$\lambda = 6$
1	0.487	0.496	0.507	0.521	0.499	0.507	0.522	0.539
3	0.301	0.238	0.230	0.243	0.334	0.285	0.291	0.316
5	0.142	0.139	0.128	0.128	0.228	0.215	0.217	0.240

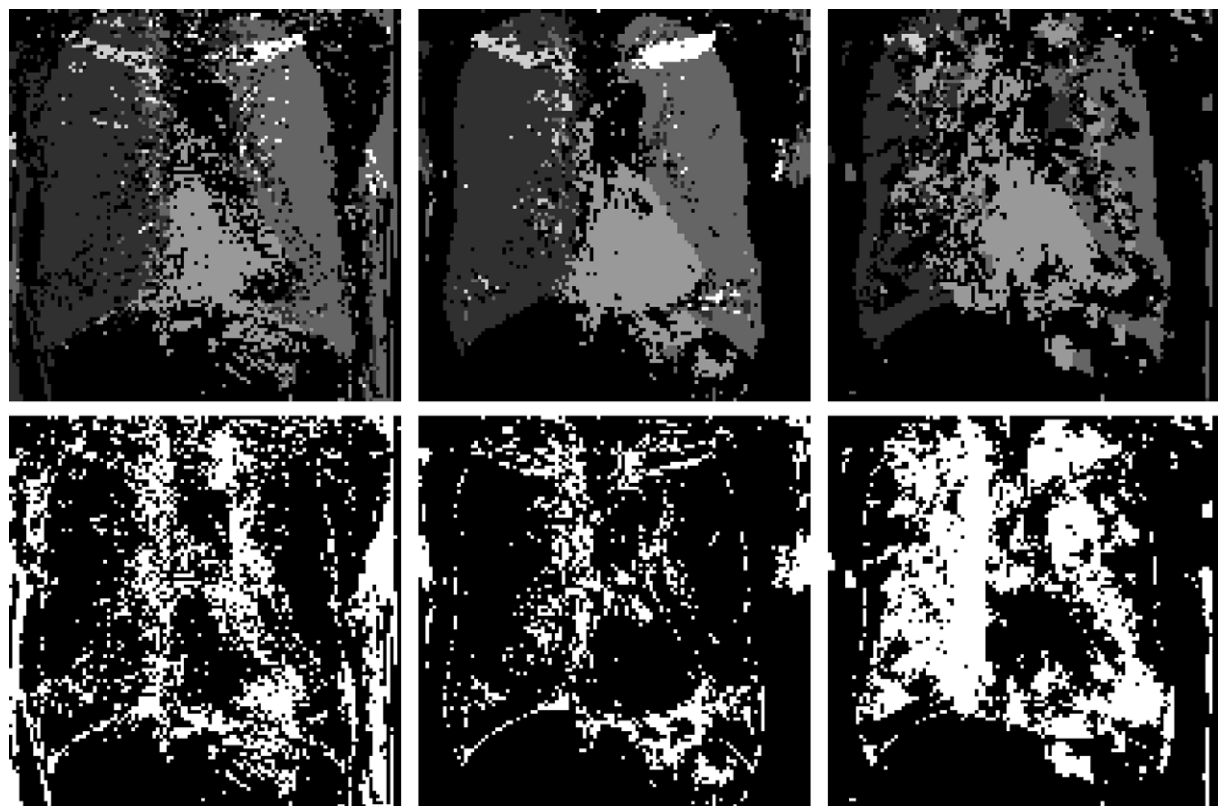


Fig. 4. Top row left: Segmentation with optimal LDA ( $d = 5$ ); middle: optimal one using CCCP ( $\lambda = 7$ ,  $d = 19$ ); right: Segmentation using no dimensionality reduction ( $d = 115$ ). Bottom row: Corresponding images indicate which pixels have been classified correctly (in black) and which incorrectly (in white).

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 [2], 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. More or less the same as for the optimal scoring approach holds for possi-

ble extensions via kernel methods [17]: the CCA framework may allow for a kernel extension of the CCCP, however, most of the time such extensions are bound to become computationally hard, which may restrict the applicability of such 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 result in clearly improved performance compared to LDA and no dimensionality reduction.

## References

- [1] M. Loog, B. van Ginneken, R.P.W. Duin, Dimensionality reduction by canonical contextual correlation projections, in: Proceedings of the European Conference on Computer Vision, Prague, Czech Republic, 2004, pp. 562–573.
- [2] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics, Springer, New York, Berlin, Heidelberg, 2001.
- [3] A.K. Jain, R.P.W. Duin, J. Mao, Statistical pattern recognition: a review, *IEEE Trans. Pattern Anal. Mach. Intell.* 22 (1) (2000) 4–37.
- [4] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, New York, 1990.
- [5] R.A. Fisher, The use of multiple measurements in taxonomic problems, *Ann. Eugen.* 7 (1936) 179–188.
- [6] R.A. Fisher, The statistical utilization of multiple measurements, *Ann. Eugen.* 8 (1938) 376–386.
- [7] C.R. Rao, The utilization of multiple measurements in problems of biological classification, *J. R. Statist. Soc. Ser. B* 10 (1948) 159–203.
- [8] P.A. Devijver, J. Kittler, *Pattern Recognition: a Statistical Approach*, Prentice-Hall, London, 1982.
- [9] G.J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition*, Wiley, New York, 1992.
- [10] B.D. Ripley, *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge, 1996.
- [11] P. Belhumeur, J. Hespanha, D. Kriegman, Eigenfaces vs. fisherfaces: recognition using class specific linear projection, *IEEE Trans. Pattern Anal. Mach. Intell.* 19 (7) (1997) 711–720.
- [12] K. Liu, Y.-Q. Cheng, J.-Y. Yang, Algebraic feature extraction for image recognition based on an optimal discriminant criterion, *Pattern Recognition* 26 (6) (1993) 903–911.
- [13] H. Hotelling, Relations between two sets of variates, *Biometrika* 28 (1936) 321–377.
- [14] G. Strang, *Linear Algebra and its Applications*, third ed., Harcourt, Brace and Jovanovich, New York, 1988.
- [15] J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K. Komatsu, M. Matsui, H. Fujita, Y. Kodera, K. Doi, Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules, *Am. J. Roentgenol.* 174 (2000) 71–74.
- [16] G. Winkler, *Image Analysis, Random Fields and Dynamic Monte Carlo Methods*, Number 27 in Applications of Mathematics, Springer, Berlin, Heidelberg, 1995.
- [17] B. Schölkopf, C.J.C. Burges, A.J. Smola, *Advances in Kernel Methods—Support Vector Learning*, MIT Press, Cambridge, 1999.

**About the Author**—MARCO LOOG studied mathematics at Utrecht University, The Netherlands, until 1997, the year he received his Master of Science degree in this discipline. Before pursuing a Ph.D., he started a two-year post-Master's program at the Department of Mathematics and Computer Science, Delft University of Technology, The Netherlands, which he finished in 1999, making him a full-blown Master of Technological Design. In 2004, he received a Ph.D. degree from the Image Sciences Institute, Utrecht, The Netherlands.

At present, he is an assistant professor at the Department of Innovation of the IT University of Copenhagen, Denmark. His ever-evolving research interests currently include probabilistic scale space theory, folklore theorems, dimensionality reduction methods, black math, blob machines, and general pattern analysis techniques for supervised image processing.

**About the Author**—BRAM VAN GINNEKEN develops computer-aided diagnosis systems for radiologists. He studied physics at Eindhoven University of Technology and at Utrecht University, in the Netherlands and he obtained his Ph.D. at the Image Sciences Institute of the University Medical Center Utrecht, where he is still employed. His favorite pastime is building endangered or extinct animals from papier-mâché.

**About the Author**—ROBERT P.W. DUIN studied applied physics at Delft University of Technology in the Netherlands. In 1978 he received the Ph.D. degree for a thesis on the accuracy of statistical pattern recognizers. In his research he included various aspects of the automatic interpretation of measurements, learning systems and classifiers. Between 1980 and 1990 he studied and developed hardware architectures and software configurations for interactive image analysis. After this period his interest was redirected to neural networks and pattern recognition. At this moment he is an associate professor of the Faculty of Electrical Engineering, Mathematics and Computer Science of Delft University of Technology. His present research is in the design, evaluation and application of algorithms that learn from examples. This includes neural network classifiers, support vector machines and classifier combining strategies. Recently he started to investigate alternative object representations for classification and became thereby interested in dissimilarity-based pattern recognition and in the possibilities to learn domain descriptions. Additionally he is interested in the relation between pattern recognition and consciousness.