Combining Reconstructive and Discriminative Subspace Methods for Robust Classification and Regression by Subsampling

Sanja Fidler, Danijel Skočaj, *Member*, *IEEE*, and Aleš Leonardis, *Member*, *IEEE*

Abstract—Linear subspace methods that provide sufficient reconstruction of the data, such as PCA, offer an efficient way of dealing with missing pixels, outliers, and occlusions that often appear in the visual data. Discriminative methods, such as LDA, which, on the other hand, are better suited for classification tasks, are highly sensitive to corrupted data. We present a theoretical framework for achieving the best of both types of methods: An approach that combines the discrimination power of discriminative methods with the reconstruction property of reconstructive methods which enables one to work on subsets of pixels in images to efficiently detect and reject the outliers. The proposed approach is therefore capable of robust classification with a high-breakdown point. We also show that subspace methods, such as CCA, which are used for solving regression tasks, can be treated in a similar manner. The theoretical results are demonstrated on several computer vision tasks showing that the proposed approach significantly outperforms the standard discriminative methods in the case of missing pixels and images containing occlusions and outliers.

Index Terms—Subspace methods, reconstructive methods, discriminative methods, robust classification, robust regression, subsampling, PCA, LDA, CCA, high-breakdown point classification, outlier detection, occlusion.

1 Introduction

Subspace methods have become a standard tool in the computer vision community for performing various types of visual learning and recognition/classification. These methods, which are based on principles originally used for statistical pattern recognition, fall into two categories, the first being *reconstructive* and the second *discriminative methods*, both of which exert distinct, yet equally important qualities. It is known that the class of reconstructive methods, such as PCA [25] (and, potentially, ICA [6] and NMF [17]), produce representations that enable sufficient reconstruction, thus being capable of dealing with the problem of missing pixels and outliers (occluded pixels) [18]. On the other hand, discriminative methods, such as LDA [10], enable the construction of flexible decision boundaries needed for classification. As both types of methods have been shown to be effective for recognition, the latter ones have often proven to yield better results [1], [21]. However, their usage is severely limited due to their nonrobust nature, preventing them from successfully coping with outliers and occlusions, which commonly appear in the visual data. We encounter the same problems when using subspace methods, such as CCA [4], which are related to regression task.

To be widely applicable, a method should have the ability to perform *robust learning* and *robust classification/ regression*. By robust, we mean the ability to detect outliers in images and, consequently, work on uncorrupted subsets of pixels, resulting in a high-breakdown point[1] method. Approaches to *robust learning* of discriminative models have been explored in the literature, although mainly focusing on detecting an image as a whole as a data outlier and discarding it from the learning process. The vast majority of these methods involve replacing the classical location and scatter matrix estimators by their robust counterparts, such as MVE estimators [5], MCD estimators [14], [16], [28], S-estimators [7], [15], M-estimators [23], and by the projection pursuit approach, as in [9], [27].

On the other hand, the problem of *robust classification/ regression* has rarely been addressed in the literature. This is mainly due to a highly nonrobust nature of discriminative methods (with the breakdown point zero) which contain too little information to successfully deal with outliers and occlusions appearing in the visual data. Specifically, the discriminative methods provide decision hyperplanes designed for optimal classification and do not, in general, offer good reconstruction of images, which is necessary for determining the pixels which are far away from their model values (i.e., are outliers [22]).

In contrast to discriminative methods, the reconstructive methods provide a principled way of performing robust recognition exploiting the redundancy in the visual data. These methods, which are known to produce good approximations of the data, have been proven successful in cases when images contain outliers, when the objects of interest in the images are occluded or appear on different backgrounds, and/or in the case where images are taken under varying illumination conditions. Several different robust

- *The authors are with the Faculty of Computer and Information Science, University of Ljubljana, Tržaška 25, SI-1001 Ljubljana, Slovenia. E-mail: {sanja.fidler, danijel.skocaj, ales.leonardis}@fri.uni-lj.si.*

Manuscript received 1 June 2004; revised 2 July 2005; accepted 18 July 2005; published online 13 Jan. 2006.
Recommended for acceptance by Y. Amit.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0278-0604.

[1] The breakdown point, as defined in statistics, is the worst-case measure. It represents the smallest fraction of pixel outliers in an image that can cause an estimator to produce arbitrarily bad results. Breakdown point zero only means that changing the value of a single pixel in an image can make an estimator fail. A high breakdown point refers to estimators that can tolerate a large amount of outliers [22].

0162-8828/06/$20.00 © 2006 IEEE Published by the IEEE Computer Society

versions of original methods have been developed, which work well under such nonideal conditions. Some of these approaches are based on substituting the standard least-squares metric by a robust one [8], while the others calculate the coefficients by utilizing a subsampling and hypothesize-and-test approach [18].

It is therefore evident that an ideal classifier should be able to combine the best of both, reconstructive and discriminative, approaches, contain information crucial for classification/regression, and also enable a calculation of the necessary coefficients by means of a robust subsampling approach. To the best of our knowledge, robust classification/regression by subsampling has not been tackled before.

In this paper, we present a method, novel in the field of robust classification, which makes the recognition of objects under nonideal conditions possible, i.e., in situations when objects are occluded or they appear on a varying background, or when their images are corrupted by outliers. The main idea behind the method is to combine the reconstructive and discriminative models by constructing a basis which, on the one hand, contains the *complete* discriminative information (of a particular discriminative model) necessary for the classification and, on the other hand, enables us to determine outliers in images and calculate the necessary coefficients by means of a subsampling approach resulting in a high breakdown point classification. The theoretical results are evaluated on several computer vision problems, showing that the proposed method significantly outperforms the standard discriminative and regression methods in the case of corrupted images.

The paper is organized as follows: We begin with a review of the related work in Section 2. In Section 3, we give a theoretical background on reconstructive and discriminative models. We formulate the problem in Section 4. In Section 5, we present our robust classification/regression method. The effectiveness of the proposed method is experimentally verified in Section 6 (in particular, we chose LDA and CCA for demonstration purposes, although our method is general and can be used with other linear discriminative methods as well). Finally, in the last section, we summarize the paper and give the conclusions.

2 Related Work

A number of approaches that combine different subspace methods already exist in the literature. The classical approach is to use PCA as a preprocessing step to LDA or CCA to overcome the singularity problems these two methods encounter when dealing with high-dimensional data such as images [1], [34], [33], [3], [31]. The fact that this can be done without losing any discriminative information [33] will serve us as an idea of how to combine both discriminative and reconstructive methods to achieve robustness to image degradations. The majority of other existing methods are concerned with improving the classification power of discriminative methods by incorporating the PCA information in different ways: In [19] and [20], the authors propose to add (or average) the output feature vectors obtained by PCA, ICA, and LDA or concatenating them into a single one upon which a designed RBF network returns the classification results. As these methods might outperform the classical discriminative methods under ideal conditions (when the images are "clean"), they still rely on calculating the feature vectors as a dot product between the different subspace bases and the testing image vector and, thereby, fail when dealing with images which contain outliers or are corrupted by noise.

An approach focusing on the classification of degraded images has been proposed by Stainvas et al. [30] and is, in its philosophy, closest to ours. The idea behind it is to improve classification of discriminative methods, which do not contain enough information to deal with corrupted data, by using the reconstruction property of the reconstructive methods. However, in their method, this combination is already done in the learning stage by minimizing concurrently the mean squared error (MSE) of the reconstruction and classification outputs resulting in an improved low-dimensional representation, which represents a trade-off between reconstruction and classification confidences. This differs greatly from our approach, which offers robustness in the *classification stage* by calculating the feature vectors on *subsets* of pixels in images and is consequently very robust to various image artifacts.

To the best of our knowledge, classification with linear discriminative subspace methods of corrupted images have not previously been done by the subsampling approach. The closest to this is the robust PCA method of [18], which makes use of the reconstruction property of PCA to successfully detect outliers and calculate the coefficients on the rest of the pixels. We will use this idea for the discriminative methods, although due to the fact that discriminative models do not approximate the data well, the implementation is not as straightforward as suggested in [13]. With this in mind, we will combine the discriminative and reconstructive methods to achieve perfect classification results (in the limits of each discriminative method) and, on the other hand, have the reconstruction abilities for detection of outliers and occlusions.

3 Theoretical Background

The central problem when working with high-dimensional data in a learning system is to find a suitable representation of the data by means of an optimal transformation. The definition of optimality varies from task to task, also depending on the knowledge one has about the learning database. In the case when little of such knowledge exists, the transformation is usually defined in the sense of optimal dimension reduction, statistical "interestingness," positiveness, or simplicity of the transformation. The representation has to be as informative as possible, thus mainly having the property of approximating the original data well. These methods are referred to as the *reconstructive methods*. On the other hand, in the case when one has prior knowledge of the class labels, hyperplanes that best separate the classes are usually sought for. The representation obtained does not usually provide good reconstruction of the data, is more task dependent, but spatially and computationally much more efficient and often gives superior classification results compared to the reconstructive methods. These methods are referred to as *discriminative methods*.

In the following, we will present a general theoretical background for both the reconstructive and discriminative methods and introduce the notation. We would like to emphasize that we shall only consider linear subspace methods.

3.1 Notation

Let n be the number of images in the training data set, each of them containing m pixels, and let c be the number of classes the images belong to. We write images as vectors and arrange them (according to the classes) in columns of a matrix $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$, $\mathbf{x}_i \in \mathbb{R}^m$. For a simpler notation, we will assume X to be centered, i.e., having zero mean, unless otherwise specified.

We will use the notation U for the basis of the reconstructive methods and W for the basis of the discriminative methods. With A, we will denote the feature matrix composed of feature vectors $\mathbf{a}_i$ expressed in the basis of the reconstructive model. We will also frequently operate with submatrices. For $\mathcal{I}$ and $\mathcal{J}$, being nonempty subsets of the set $\{1, 2, \ldots, n\}$, the symbol $M_{\mathcal{I}:}$ will be used to denote the submatrix of M containing only those *rows* of M whose indices are in the set $\mathcal{I}$, arranged in their natural order and, similarly, $M_{:\mathcal{J}}$ will stand for the submatrix of M containing only those *columns* of M whose indices are in the set $\mathcal{J}$. For a vector $\mathbf{a}$, the symbol $\mathbf{a}_{\mathcal{I}}$ will be used to denote those elements of $\mathbf{a}$ whose indices are in $\mathcal{I}$. In our calculations, we will mainly be operating with two sets, $\mathcal{K} = \{1, 2, \ldots, k\}$ and $\mathcal{N} - \mathcal{K} = \{k+1, k+2, \ldots, n\}$. Thus, if a matrix M consists of n *columns*, M can be written as $M = \left[M_{:\mathcal{K}}, M_{:(\mathcal{N}-\mathcal{K})}\right]$ and, if it consists of n *rows*,

$$M = \begin{bmatrix} M_{\mathcal{K}:} \\ M_{(\mathcal{N}-\mathcal{K}):} \end{bmatrix}.$$

3.2 Reconstructive Methods

The main goal of the reconstructive methods is to find a linear representation

$$X = UA \tag{1}$$

$$\text{where} \qquad U = [\mathbf{u}_1, \ldots, \mathbf{u}_n] \in \mathbb{R}^{m \times n},\ A \in \mathbb{R}^{n \times n} \tag{2}$$

that best describes the data subject to different criteria. Here, U is called the *basis matrix*, while the matrix of coefficients, $A = (U^T U)^{-1} U^T X$, is referred to as the *feature matrix*. If U is an orthonormal matrix, A simplifies to $A = U^T X$.

By far, most widely known and used method is Principal Component Analysis (PCA) which seeks for a low-dimensional representation of the data which minimizes the squared reconstruction error [11]. PCA can also be interpreted as searching for a linear transformation that minimizes the statistical dependencies of second order between the transformed data, thus PCA finds a basis $\{\mathbf{u}_i\}_{i=1}^n$ that yields mutually uncorrelated coefficient vectors. This is in contrast to Independent Component Analysis, which finds a basis of mutually independent vectors by also minimizing higher-order statistical dependencies [2], [6]. Another method that recently gained attention is Nonnegative Matrix Factorization (NMF), which seeks a representation that has all the coefficients and the basis vectors nonnegative (here, the data matrix X is obviously not centered) [17].

After the optimal basis is obtained, it can then be reduced to $U_{:\mathcal{K}}$, where $k := |\mathcal{K}|$ indicates that usually only k, $k \ll n$, basis vectors (those that take up the most variance) are needed to represent $\mathbf{x}$ to a sufficient degree of accuracy as their linear combination

$$\tilde{\mathbf{x}} = \sum_{j=1}^{k} a_j(\mathbf{x}) \mathbf{u}_j = U_{:\mathcal{K}} \left(\left(U_{:\mathcal{K}}^T U_{:\mathcal{K}} \right)^{-1} U_{:\mathcal{K}}^T \mathbf{x} \right). \tag{3}$$

Here, $\tilde{\mathbf{x}}$ denotes the approximation to $\mathbf{x}$ and $a_j(\mathbf{x})$ are the coefficients obtained by projecting $\mathbf{x}$ onto the selected basis, $\mathbf{a}_{\mathcal{K}} := [a_1, a_2, \ldots, a_k]^T = (U_{:\mathcal{K}}^T U_{:\mathcal{K}})^{-1} U_{:\mathcal{K}}^T \mathbf{x}$. If the basis U is orthonormal, $\mathbf{a}_{\mathcal{K}} = U_{:\mathcal{K}}^T \mathbf{x}$.

In the theory to come, we will not choose any of the mentioned methods in particular, but, rather, try to stay general throughout the paper. It might, though, be worth emphasizing that the most appealing of the stated methods is PCA since it is optimal in reconstruction error and can therefore detect outlying pixels and occlusions [18] to a larger degree of accuracy than the other methods.

3.3 Discriminative Methods

Discriminative methods were designed particularly for classification tasks. They assume that prior knowledge about classes of the training data is available, which is then integrated in the supervised learning process to produce a small number of hyperplanes that are capable of separating the training data with no (or little) error.

To be more specific, the objective of discriminative methods is to find a linear function,

$$\begin{aligned} g(\mathbf{x}) &= W^T \mathbf{x}, \\ \text{where} \quad W &= [\mathbf{w}_1, \ldots, \mathbf{w}_c] \in \mathbb{R}^{m \times c}, \end{aligned} \tag{4}$$

which is used for transforming the data into a lower-dimensional classification space upon which it is decided, according to some chosen metric, to which class a given sample $\mathbf{x}$ belongs.[2] To find an optimal decision function, a number of different criteria can be employed.

Probably the most widely used for classification is Linear Discriminant Analysis, which, in the training stage, finds the projection directions on which the intraclass scatter is minimized while the interclass scatter is maximized. Specifically, LDA maximizes the objective function, also called the Fisher criterion function [12], which is defined as

$$J(\mathbf{w}) = \frac{\mathbf{w}^T S_b \mathbf{w}}{\mathbf{w}^T S_w \mathbf{w}}, \tag{5}$$

where S_b denotes the between-class and S_w the within-class scatter matrix of the training data. In the classification stage, the new image samples are projected onto these directions to form feature vectors according to which samples are classified to a certain class.

The subspace methods, such as Canonical Correlation Analysis (CCA) [24], [4], which are used for regression tasks, mathematically follow similar concepts and can therefore be addressed by our proposed approach in a similar fashion, as we will demonstrate in Section 6. Here, we briefly review the theory of CCA, which is a supervised method relating two sets of observations, one set being composed of training images and the other set of the corresponding measurements (e.g., orientations or positions of an object). In the *training stage*, CCA finds pairs of directions (canonical correlation vectors) that yield maximum correlation between the projections of input vectors. This can be followed by performing linear regression on the obtained projections (canonical correlation coefficients). More specifically, given n pairs of

[2] To be exact, W is $(c-1)$-dimensional in the LDA case [1].

mean-centered observations $(\mathbf{x}_i \in \mathbb{R}^p, \mathbf{y}_i \in \mathbb{R}^q), i = 1, \ldots, n$, aligned in the data matrices $X = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$ and $Y = [\mathbf{y}_1, \ldots, \mathbf{y}_n] \in \mathbb{R}^{q \times n}$, CCA finds $c = \min(p, q)$ pairs of directions $\mathbf{w_x} \in \mathbb{R}^p$ and $\mathbf{w_y} \in \mathbb{R}^q$ that maximize the correlation between the projections $\mathbf{w}_\mathbf{x}^T \mathbf{x}_i$ and $\mathbf{w}_\mathbf{y}^T \mathbf{y}_i$. CCA maximizes the function

$$\rho(\mathbf{w_x}, \mathbf{w_y}) = \frac{\mathbf{w}_\mathbf{x}^T C_{\mathbf{xy}} \mathbf{w_y}}{\sqrt{\mathbf{w}_\mathbf{x}^T C_{\mathbf{xx}} \mathbf{w_x} \mathbf{w}_\mathbf{y}^T C_{\mathbf{yy}} \mathbf{w_y}}}, \tag{6}$$

where $C_{\mathbf{xx}}$, $C_{\mathbf{yy}}$, and $C_{\mathbf{xy}}$ are within-set and between-set covariance matrices of the input data. In the *regression stage*, the orientation (or position) of the object is estimated by using canonical correlation coefficients obtained from a novel image of the object.

4 Defining the Problem

The comparison between discriminative and reconstructive methods for classification tasks has been a subject of extensive research and testing [1], [21]. The general conclusion was that, under ideal circumstances (when images do not contain artifacts such as noise or outliers), discriminative methods outperform the reconstructive methods. The explanation for this is rather obvious: The discriminative methods focus more on specific prior knowledge, which can thus be more efficiently integrated into the learning process. These methods, in most cases, offer linear transformations of much lower dimensions than the reconstructive methods. But, there is a trade-off to this: By having fewer basis vectors, these methods do not usually provide good approximation of the data which is necessary for successful detection of outliers and occlusions.

Images often contain noise or outliers, that is, pixels that do not belong to objects being depicted. Therefore, tools must exist that enable us to extract reliable information based on only uncorrupted subsets of pixels. Since both reconstructive as well as discriminative methods rely on calculating the dot product $U^T\mathbf{x}$ (appearing in the calculation of coefficients in (3)) and $W^T\mathbf{x}$ (needed in a linear classifier g in (4)), respectively, they obviously take into account *all* pixels in an image $\mathbf{x}$. The results can therefore be unreliable when $\mathbf{x}$ contains even a small amount of outlying pixels.

However, this undesirable property of linear methods has been successfully overcome for reconstructive methods (in particular PCA) by employing the *robust coefficient estimation procedure* [18]. The basic idea behind the approach is to translate the original dot product calculation into solving an overdetermined linear system using only subsets of pixels. The obtained coefficients are used for backprojection and the pixels that deviate the most from the expected approximation error are pronounced to be outliers. The better the reconstruction the given basis provides, the more reliable the detection of outliers is and, consequently, the more exact the obtained coefficients are. The details of the approach are given in the Appendix.

As was already mentioned, the discriminative methods usually give only a small number of basis vectors which do not offer a satisfactory reconstruction property that would enable a correct detection of outlying pixels in an image and, consequently, a reliable calculation of the linear classifier g. In this paper, we will present a method which shows how to construct a basis that, on one hand, contains sufficient reconstructive information to enable the use of the subsampling approach [18] and, on the other hand, contains all of the discriminative information for classification.

We would like to emphasize that our aim is not to improve classification power of the standard discriminative methods in the case of ideal data, but to also achieve similar results in the case of corrupted data. We will also not deal with robustness in the training stage and presume that the training stage of each method used was already performed. Our main concern is *robustness in the classification or regression stage*. By robust, we mean the ability to correctly classify a novel image which contains a large number of corrupted pixels (outliers).

5 Robust Classification/Regression Approach

The classification of discriminative models is based on a linear function

$$g(\mathbf{x}) = W^T\mathbf{x}, \tag{7}$$

which is used for transforming the data into a lower-dimensional classification space upon which it is decided, according to some chosen metric, to which class a novel image $\mathbf{x}$ belongs.

What we will do is rewrite the dot product in (7) into a form that will enable robust estimation. This will be done by incorporating the basis of a reconstructive model into this classification function by employing a few linear algebra operations.

To begin with, let $U \in \mathbb{R}^{m \times n}$ denote the complete basis of a reconstructive model. Let us first point out that our robust method will not need the complete reconstructive basis; this is meant exclusively to give a justification to our final calculations. For the purpose of clarity, let us also assume that the reconstructive basis is orthonormal with the extension to a nonorthogonal basis being just a matter of matrix manipulation.

To rewrite the expression in (7), we will use the fact that both bases, U and W, lie in the span of the training data vectors. Since $X = UA$ (by definition in (1)), this obviously holds for the reconstructive models. Moreover, because U has rank n, it spans *exactly* the same space as the training data.[3] As it might be intuitively obvious that the discriminative basis also "lives" in the learning data space, there is no proof to cover all the discriminative models; therefore, each of them has to be dealt with separately.[4] Since U spans the same space as X and W is a subspace of this space, the immediate consequence is that the discriminative basis can be written in the basis of the reconstructive model, i.e., $W = UV$, where $V \in \mathbb{R}^{n \times c}$. Note that V is a matrix that can be already calculated in the training stage as a projection of

[3] To be exact, U is usually of rank $n-1$ (because the vectors in X are mean-centered) or even smaller if some input images are linearly dependent. However, the same observation also holds for X, thus this fact does not influence our derivation.

[4] The proof in the LDA case is due to the recent paper of Yang and Yang [33] who showed that LDA can be performed in the PCA transformed space and then backprojected to the PCA space to get the image LDA vectors without losing any discriminative information. This automatically implies that the LDA vectors lie in the span of the PCA basis which spans the same linear space as the learning data, which concludes our argument. For the CCA, it was also shown that the CCA vectors lie in the span of the training data [3], [24].

W onto U. If U is orthogonal, then $V = U^T W$; otherwise, $V = (U^T U)^{-1} U^T W$.

Similarly, assuming a novel image $\mathbf{x}$ (which we want to classify) follows a distribution of one of the classes in the training set, it can therefore be well approximated with a linear combination of the basis vectors of the reconstructive model, $\mathbf{x} \approx a_1 \mathbf{u}_1 + \cdots + a_n \mathbf{u}_n = U\mathbf{a}$. The classification function now takes the following form:

$$g(\mathbf{x}) = W^T \mathbf{x} = \left(V^T U^T\right)(U\mathbf{a}) = V^T \mathbf{a}. \tag{8}$$

We have rewritten the function into an expression that uses the feature vectors corresponding to the reconstructive model. This is a promising start since an efficient algorithm already exists for robust calculation of the coefficient vector $\mathbf{a}$ of the reconstructive model in cases when the data contains outliers, occlusion, or non-Gaussian noise [18]. We will employ this method, but not just yet. Notice that the expression in (8) demands *all* n coefficients for its calculation. Since, in most computer vision applications, the number of training images n is large, the computational complexity of the robust estimation of all the coefficients would be too prohibitive to make this method applicable in practice. The idea is to use only a truncated basis $U_{:\mathcal{K}}$ of $k \ll n$ basis vectors, which is usually used for calculations involving reconstructive methods. The number k is chosen so that the truncated basis approximates the data to a good degree of accuracy.

However, this truncated reconstructive basis is not sufficient for our classification task. In particular, it provides only k coefficients, $\mathbf{a}_{\mathcal{K}} = [a_1, \ldots, a_k]^T$, which makes the calculation of (8) impossible. To see this, we rewrite the expression in (8):

$$\begin{aligned} g(\mathbf{x}) = V^T \mathbf{a} &= \left[V_{\mathcal{K}:}^T,\, V_{(\mathcal{N}-\mathcal{K}):}^T\right] \begin{bmatrix} a_{\mathcal{K}} \\ a_{(\mathcal{N}-\mathcal{K})} \end{bmatrix} = \\ &= V_{\mathcal{K}:}^T \mathbf{a}_{\mathcal{K}} + V_{(\mathcal{N}-\mathcal{K}):}^T \mathbf{a}_{(\mathcal{N}-\mathcal{K})}. \end{aligned} \tag{9}$$

The function g could, in principle, be estimated only according to the truncated coefficient vector (by calculating only the first term in the sum), but by doing so we would very likely be losing valuable discriminative information contained in the last $n-k$ coefficients. While the first k coefficients contain most of the reconstructive information, there is no guarantee that most of the discriminative information is present in the first k of them as well. This is demonstrated in the first experiment in Section 6.

Obviously, some extra information needs to be added to the truncated reconstructive basis $U_{:\mathcal{K}}$ to retain the complete discrimination power of g (i.e., also enable the calculation of the second term of the sum in (9)). These will be done by augmenting the truncated reconstructive basis with a small number of additional vectors.

Let us define $\tilde{W} := U_{:(\mathcal{N}-\mathcal{K})} V_{(\mathcal{N}-\mathcal{K}):} \in \mathbb{R}^{m \times c}$. The matrix $\tilde{W}$ is composed of c vectors arranged in its columns which are linear combinations of the last $n-k$ basis vectors of a reconstructive model. Each of them is orthogonal to all of the first k vectors of the reconstructive basis; however, they are not mutually orthogonal. In order to enable easier calculations later on, the matrix $\tilde{W}$ can be orthogonalized adequately:

$$\tilde{W}_{\perp} = \tilde{W}(\tilde{W}^T \tilde{W})^{-1/2}. \tag{10}$$

Next, let us define $\widehat{U}$ and $\widehat{V}$ as:

$$\begin{aligned} \widehat{U} &= \left[U_{:\mathcal{K}}, \tilde{W}_{\perp}\right] = \left[U_{:\mathcal{K}}, \tilde{W}(\tilde{W}^T \tilde{W})^{-1/2}\right] \in \mathbb{R}^{m \times (k+c)} \\ \widehat{V} &= \begin{bmatrix} V_{\mathcal{K}:} \\ (\tilde{W}^T \tilde{W})^{1/2} \end{bmatrix} \in \mathbb{R}^{(k+c) \times c}. \end{aligned} \tag{11}$$

The new basis $\widehat{U}$ is the basis $U_{:\mathcal{K}}$ extended by $c \ll n$ additional vectors, while $\widehat{V}$ is the matrix $V_{\mathcal{K}:}$ also extended by c row vectors.

Now, it is easy to show that $\widehat{U}$ and $\widehat{V}$ contain *all* of the discriminative information contained in W. The new classification function $\widehat{g}(\mathbf{x}) := (\widehat{U}\widehat{V})^T \mathbf{x}$ can be expressed as

$$\begin{aligned} \widehat{g}(\mathbf{x}) &= (\widehat{U}\widehat{V})^T \mathbf{x} = \\ &= \left(\left[U_{:\mathcal{K}}, \tilde{W}(\tilde{W}^T \tilde{W})^{-1/2}\right] \begin{bmatrix} V_{\mathcal{K}:} \\ (\tilde{W}^T \tilde{W})^{1/2} \end{bmatrix}\right)^T \mathbf{x} = \\ &= (U_{:\mathcal{K}} V_{\mathcal{K}:} + \tilde{W})^T \mathbf{x} = \\ &= (U_{:\mathcal{K}} V_{\mathcal{K}:} + U_{:(\mathcal{N}-\mathcal{K})} V_{(\mathcal{N}-\mathcal{K}):})^T \mathbf{x} = \\ &= \left(\left[U_{:\mathcal{K}}, U_{:(\mathcal{N}-\mathcal{K})}\right] \begin{bmatrix} V_{\mathcal{K}:} \\ V_{(\mathcal{N}-\mathcal{K}):} \end{bmatrix}\right)^T \mathbf{x} = \\ &= (UV)^T \mathbf{x} = W^T \mathbf{x} = g(\mathbf{x}) \end{aligned} \tag{12}$$

and is thus equivalent to the original classification function $g(\mathbf{x})$. This concludes our argument.

Since the new basis $\widehat{U} := [\widehat{\mathbf{u}}_1, \widehat{\mathbf{u}}_2, \ldots, \widehat{\mathbf{u}}_{k+c}]^T$ contains the truncated reconstructive basis $U_{:\mathcal{K}}$ and, thus, offers a good reconstruction of images, the image $\mathbf{x}$ can be well approximated[5] in this extended basis:

$$\mathbf{x} \approx \widehat{a}_1 \widehat{\mathbf{u}}_1 + \cdots + \widehat{a}_k \widehat{\mathbf{u}}_k + \cdots + \widehat{a}_{k+c} \widehat{\mathbf{u}}_{k+c} := \widehat{U}\widehat{\mathbf{a}}. \tag{13}$$

As the matrix $\widehat{U}$ is orthogonal, the coefficients can be obtained in the least square sense as:

$$\widehat{\mathbf{a}} = \widehat{U}^T \mathbf{x}. \tag{14}$$

But, when the image $\mathbf{x}$ is corrupted by outliers, the reconstructive property of $\widehat{U}$ in (13) enables us to employ the method of [18] to successfully detect outliers of $\mathbf{x}$ and calculate the coefficient vector $\widehat{\mathbf{a}} = [\widehat{a}_1, \ldots, \widehat{a}_{k+c}]^T$ robustly using only the nonoccluded pixels in the image $\mathbf{x}$, as described in the Appendix. We have thus reduced the estimation of the n-dimensional coefficient vector $\mathbf{a}$ in (8) down to calculating a $(k+c)$-dimensional vector $\widehat{\mathbf{a}}$, where c is, in most applications, much smaller than n.

Since the robust estimation of $\widehat{\mathbf{a}}$ is a good approximation to (14) and the following holds,

$$g(\mathbf{x}) = \widehat{V}^T (\widehat{U}^T \mathbf{x}) = \widehat{V}^T \widehat{\mathbf{a}}, \tag{15}$$

the classification function g can therefore be calculated as the dot product of the extended matrix $\widehat{V}$ and $\widehat{\mathbf{a}}$, but since the coefficient vector $\widehat{\mathbf{a}}$ can be obtained in a robust manner, the new calculation of g is also robust to outliers and occlusions.

To summarize, we constructed a basis $\widehat{U}$ of $k+c$ vectors (where $k+c$ is usually far smaller than n) which carries the *complete* discriminative information (in the limits of a chosen

[5] The approximation (13) is even more exact than using only k reconstructive basis vectors since the new basis contains a few more extra vectors (linear combinations of the last $n-k$ reconstructive basis vectors) carrying some additional variance.

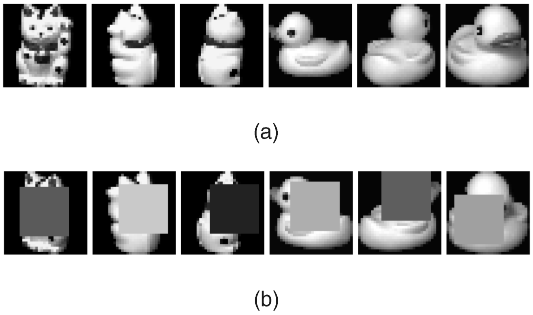


Fig. 1. (a) A few training images and (b) a few occluded test images of two COIL objects.

discriminative method), enables the detection of outliers and occlusions, and offers the calculation of the coefficient vector $\widehat{\mathbf{a}}$ on the uncorrupted pixels. Once the coefficients are obtained, the classification function g can be calculated with (15), as a dot product of the coefficient vector $\widehat{\mathbf{a}}$ and the extended matrix $\widehat{V}$. We must emphasize that both $\widehat{U}$ and $\widehat{V}$ are already calculated in the training stage and need no further calculations in the classification stage. These procedures are summarized in Algorithms 1 and 2.

```
Algorithm 1: Learning Phase
Input: $X \in \mathbb{R}^{m\times n}$
Output: $\widehat{U} \in \mathbb{R}^{m\times(k+c)}$, $\widehat{V} \in \mathbb{R}^{(k+c)\times c}$
  recMethod$(X) \rightarrow U \in \mathbb{R}^{m\times n}, A \in \mathbb{R}^{n\times n}$
  discMethod$(A) \rightarrow V \in \mathbb{R}^{n\times c}$
  $\tilde{W} = U_{:(\mathcal{N}-\mathcal{K})} V_{(\mathcal{N}-\mathcal{K}):} \in \mathbb{R}^{m\times c}$
  $\widehat{U} = \left[U_{:\mathcal{K}}, \tilde{W}(\tilde{W}^T\tilde{W})^{-1/2}\right] \in \mathbb{R}^{m\times(k+c)}$
  $\widehat{V} = \begin{bmatrix} V_{\mathcal{K}:} \\ (\tilde{W}^T\tilde{W})^{1/2} \end{bmatrix} \in \mathbb{R}^{(k+c)\times c}$
```

```
Algorithm 2: Classification Phase
Input: $\mathbf{x} \in \mathbb{R}^{m}$, $\widehat{U} \in \mathbb{R}^{m\times(k+c)}$, $\widehat{V} \in \mathbb{R}^{(k+c)\times c}$
Output: $g(\mathbf{x}) \in \mathbb{R}^{c}$
  Project $\mathbf{x}$ into $\widehat{U}$ in a robust way to obtain $\widehat{\mathbf{a}} \in \mathbb{R}^{k+c}$.
  $g(\mathbf{x}) = \widehat{V}^T\widehat{\mathbf{a}} \in \mathbb{R}^{c}$
```

6 Experimental Results

In this section, empirical evaluation of the proposed approach is presented to demonstrate its advantages over the standard discriminative methods for robust classification. Specifically, we chose to evaluate our extended basis principle for LDA and CCA, but, as emphasized in the theoretical part of the paper, the framework is general and can be applied to any other linear subspace method to give similar results as will be presented here. We tested our proposed robust technique on three traditional computer vision tasks: *object* and *face recognition*, for which we used the extended LDA basis approach, and *estimation of objects' orientation*, which is a regression task and is addressed using the extended CCA approach. These problems were selected because they clearly show the sensitivity of the standard subspace approaches to non-Gaussian noise (occlusions) and demonstrate the ability of the proposed method to overcome this shortcoming on different image domains.

TABLE 1
Results on Two COIL Objects

	Fisher criterion				MARE
	ground truth	non-robust	missing pixels	robust	ground truth
LDA	4.84	0.04	0.45	0.03	40.17
LDAonK	1.47	0.05	1.08	0.39	17.29
LDAaPCA	4.84	0.04	2.65	1.06	17.65

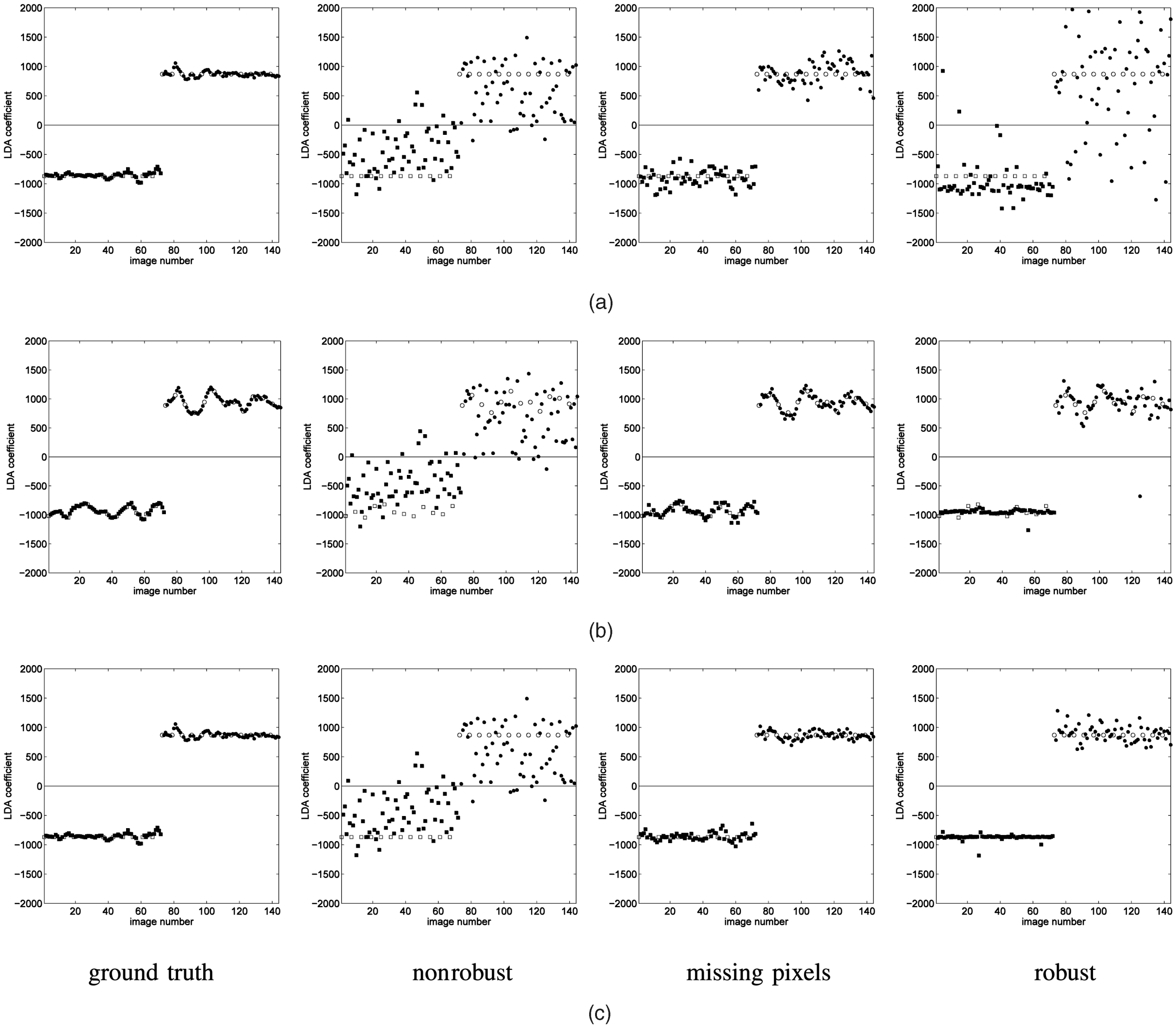
6.1 Robust Estimation of LDA Coefficients

6.1.1 Object Recognition

We first demonstrate the performance of the extended LDA approach on a simple two-class problem, where the task is to correctly classify a novel image to one of two classes. We performed the experiment on two objects from the COIL database [26]. In the training stage, 12 images of each object were used (altogether, 24 images of size 32×32, some of them are shown in Fig. 1a), while the remaining 60 were used for testing. In the first part of the experiment, the nonoccluded test images were used, while, in the second part, when we evaluated the robustness of the proposed approach, each test image was occluded with a square of a random intensity at a randomly chosen position (Fig. 1b).

In this experiment, we show two major issues that we want to emphasize in this paper. First, it is demonstrated that, by performing LDA classification in a truncated PCA space (using only the first term in the classification function (9) and thereby discarding the discriminative information contained in the second term), the results are very unreliable. This clearly indicates that the proposed extended basis is crucial to obtain quality results. Second, we show how well the proposed representation, which holds discriminative and reconstructive information, can deal with occlusions. When evaluating robustness we will also show that the straightforward application of the robust method of [18] to the standard LDA basis does not give satisfactory results since the LDA basis does not provide enough information for reconstruction to successfully deal with outliers. This again leads to a conclusion that the extended LDA representation is necessary for a reliable robust classification.

To demonstrate these issues, three different approaches were tested and compared: The standard LDA (denoted as *LDA* in the tables and figures to follow), the LDA classification performed in the truncated PCA basis (denoted as *LDAonK*), and classification using the extended LDA basis (referred to as *LDAaPCA*). The results for the three different approaches are shown in Table 1 and Fig. 2. In the first row of Table 1 and Fig. 2 (denoted as *LDA*), the results for the standard LDA approach are displayed. The second row (*LDAonK*) shows the results of the estimation of LDA coefficients in the truncated PCA subspace (where only the first 12 principal vectors and components were retained). Finally, the results in the third row (*LDAaPCA*) were obtained using the proposed method; in this case, the 11-dimensional principal subspace was augmented with one additional basis vector holding discriminative information contained in the discarded principal vectors (yielding a 12-dimensional subspace—a subspace of the same dimensionality as in the *LDAonK* case to enable a fair comparison of the two methods). The estimation of LDA coefficients was then performed in



ground truth

nonrobust

missing pixels

robust

Fig. 2. Results on two COIL objects. (a) LDA, (b) LDAonK, and (c) LDAaPCA.

this augmented subspace. Table 1 displays quantitative results—the values of the Fisher criterion (5) calculated for the test images, while, in Fig. 2, the values of the LDA coefficient are presented in a graphical form. The projections of the images of the first object are depicted as squares (empty squares for training images and filled squares for test images), while the projections of the images of the second object are denoted as circles.

The first column of Table 1 and Fig. 2 (denoted as *ground truth*) shows the results of the three approaches applied to the nonoccluded test images. Since the test images were "clean," the standard LDA approach performed very well. The projections of training images of two objects are perfectly separated and the generalization to test images is rather good as well; the recognition rate is 100 percent and the values of the Fisher criterion are very high. When the estimation of LDA coefficients was performed in the truncated PCA subspace, the projections of training images were still separated for the two classes (depicted in the leftmost plot in Fig. 2b), but not as well as in the standard LDA case and the value of the Fisher criterion was also significantly smaller (second row in Table 1). This indicates that, by truncating the full principal subspace, some significant information which is necessary for the optimal calculation of LDA coefficients is lost. This is even more evident in Fig. 3, which depicts the results of LDA classification performed in different dimensions of the truncated PCA space. By increasing the subspace dimension (k), the discarded discriminative information decreases and the results of the *LDAonK* approach converge to the optimal ones. The optimal results were also achieved when the truncated principal subspace was extended with the additional vector as proposed in the paper. The *LDA* and *LDAaPCA* approaches produced equivalent results. This clearly shows that the appended basis vector captures all the LDA-relevant information, which is contained in the discarded principal vectors and which is disregarded by *LDAonK*.

In the second to fourth columns of Table 1 and Fig. 2, the results of robustness performance of the different approaches on the occluded test images are presented. Specifically, the second column (indicated as *nonrobust*) shows the poor performance of the standard nonrobust way of calculating the LDA coefficients (using the dot product as stated in (4)) for all three approaches. The occlusions on the test images seriously affected the calculation of the LDA coefficients, thus the recognition of the objects was very unreliable. For the next two columns, the robust procedure as described in the Appendix was used for estimation of the

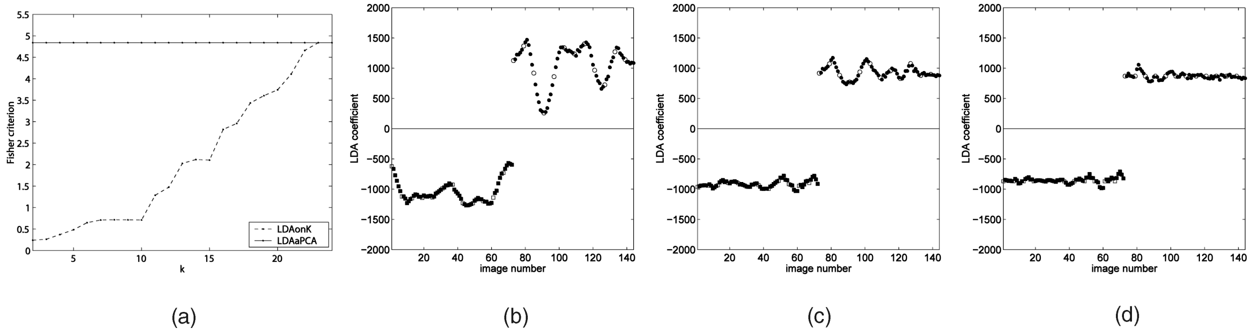


Fig. 3. Results on ground truth images of two COIL objects: (a) for various k, (b) LDAonK approach for $k = 3$, (c) for $k = 13$, and (d) for $k = 23$.

necessary coefficients for all three approaches. First, outliers were treated as *missing pixels* and were omitted during the robust computation of the subspace coefficients giving the results shown in the third column (denoted as *missing pixels*) of Table 1 and Fig. 2. The values of the Fisher criterion and the plots in Fig. 2 of the LDA coefficients show that *LDAonK* and *LDAaPCA* produced more reliable results than the *LDA* approach. This is due to the fact that both bases, *LDAonK* and *LDAaPCA*, carry more variance than the standard LDA basis, which is a prerequisite for the robust procedure to perform well. Furthermore, it is also evident that *LDAaPCA* approach outperformed *LDAonK* approach due to an even richer representation. Finally, the fourth column (denoted as *robust*) presents the results of the robust method run for all three approaches where no information about the occlusions were presumed to be known with the robust procedure detecting the outliers as described in the Appendix. Since the outlier detection largely depends on the reconstruction error (which is shown in the last column of Table 1 in terms of mean absolute reconstruction error (MARE)), the robust approach for the standard *LDA* basis produced inferior results. In the case of two-class classification, a linear discriminant vector spans only a one-dimensional subspace, which does not enable sufficient reconstruction of images and, consequently, a reliable detection of outliers. The robust procedure using the basis of *LDAonK* and *LDAaPCA* yielded significantly better results utilizing reconstructive properties of the PCA method performed on a 12-dimensional subspace. It is therefore evident that, for a successful detection of outliers, the reconstructive property of the basis is crucial and is usually not provided by the discriminative methods.

6.1.2 Face Recognition

With this experiment, our goal was three-fold: 1) to demonstrate the performance of the robust procedure in the classification task for various amounts of degraded pixels, 2) to show how the chosen value for k in the $(k + c)$-dimensiona extended (LDA) basis influences the performance of the robust method for classification in the presence of occlusion, and 3) to give a visual idea of how well the robust procedure selects the "good" pixels and discards the occluded ones from the calculation process.

The face recognition experiment was performed on two testbeds: ORL database and AR face database.

In the first experiment, the robustness of the method was tested on the ORL face database from Olivetti Research Laboratory in Cambridge, United Kingdom [29]. The database contains 10 different images of 40 distinct subjects. One half of the images, resized to 64×64 pixels, was used for training (five images per person, see Figs. 4a and 4b), while the other half was occluded with varying amount of occlusion (Fig. 4c) and used for testing.

Fig. 5 shows the results obtained by performing the LDA classification on the ORL test images with different amount of occlusions (0-95 percent) in three different ways: using the standard nonrobust approach (indicated as *nonrobust*), and by the proposed robust approach with known (*miss.pix.*), and unknown (*robust*) positions of outliers. In the first part of the experiment, five-dimensional principal subspace ($k = 5$) was augmented with additional basis vectors. One can observe that the standard nonrobust approach was considerably affected by the occlusions and its efficiency rapidly decreased with the increase of the percentage of outliers. In contrast, the proposed robust method performed very well. When the positions of the outliers were known, the nonmissing pixels contained information sufficient for reliable discrimination between the 40 subjects for almost all levels of occlusion. Even when the robust method had to automatically detect outliers, the results were degraded only after the occlusion reached more than 50 percent, which demonstrates the high breakdown point of the proposed method.

In the second part of the experiment, we tested how different dimensionalities of the PCA subspace used in the extended LDA basis influence the results. The results are depicted in Fig. 6. The values of the Fisher criterion and recognition rates obtained on the test images with

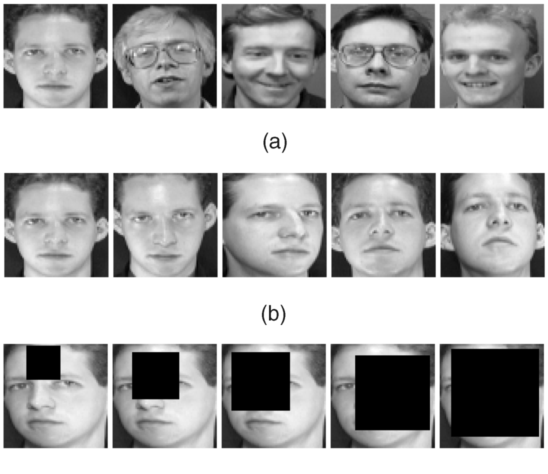
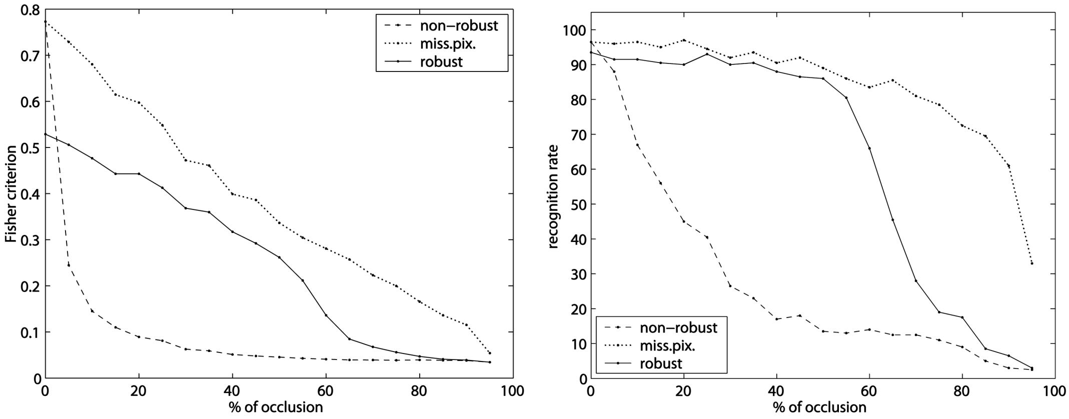
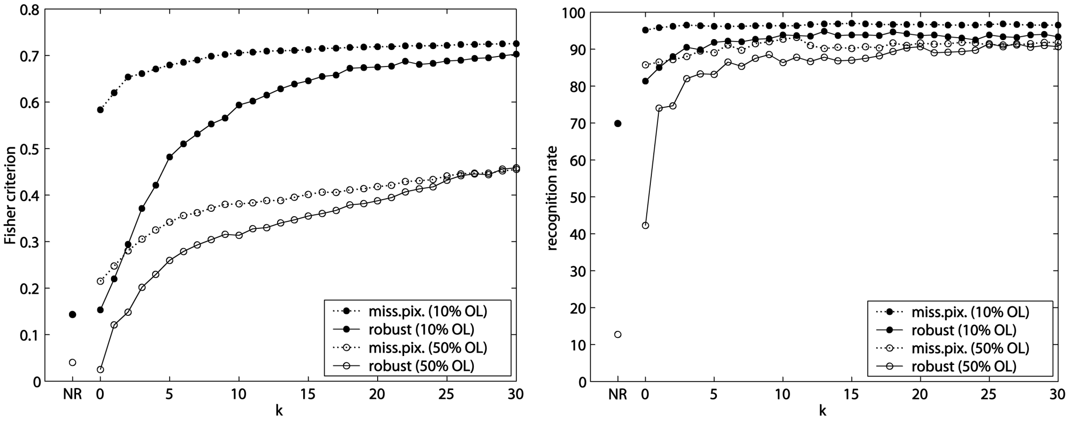


Fig. 4. Images from the ORL face database. (a) and (b) Training images. (c) Test image occluded with 10 percent, 20 percent, 30 percent, 50 percent, and 70 percent occlusion.



(a) (b)

Fig. 5. Results on the ORL face database with test images containing different amounts of occlusion and $k = 5$: (a) Fisher criterion. (b) Recognition rate.



(a) (b)

Fig. 6. Results on the ORL face database with test images containing 10 percent and 50 percent of occlusion for standard LDA approach (NR) and for the proposed method with different k: (a) Fisher criterion. (b) Recognition rate.

10 percent (filled circles) and 50 percent (empty circles) of occlusion are presented for known (dotted line) and unknown (solid line) positions of outliers. As expected, the nonrobust method (indicated as *NR*) did not perform well. When we applied the method for robust estimation of coefficients in the subspace spanned by the LDA vectors only ($k = 0$), the results improved, but they were still significantly inferior to the results of the proposed method. By also having the reconstructive basis (augmented with additional vectors), the reconstructive power of the method increased and the detection of the outliers became more reliable. The results improved as the value of k increased. This indicates that, by improving the reconstruction power of the extended LDA basis, the robust method is better able to correctly detect the outliers and is consequently capable of a more exact estimation of the LDA coefficients, resulting in a reliable classification on highly occluded images. It is worth noting that a satisfying level of the evaluation criteria was already achieved when only a few (i.e., five) principal vectors were used.

Last, the performance of the proposed approach was evaluated on images containing real occlusions. The experiment was conducted on the AR face database [21], which contains over 4,000 color images of 126 persons taken during two distinct photo sessions (separated by two weeks), with different facial expressions, illumination conditions, and occlusions (sun glasses and scarf). This database is commonly used by researchers for performance evaluation of robust face recognition algorithms and, therefore, has a comparative value. Following Martinez and Kak [21], images of 50 people, in our case, the first 25 males and 25 females, were taken. In the preprocessing step, the original images were converted to gray scale, aligned by the eyes, resized, and cropped to size 100×52. In the training stage, the images of neutral, smile, and anger face expressions were used from both sessions (six images per person, 300 images altogether), while the images of occlusion by glasses and scarf and images of the scream face expression were used for testing (also six images per person). Only those images were taken as the training as well as the test images that were captured under the same illumination conditions. Fig. 7a shows examples of training images for two people from the database, while Fig. 7b contains the test images for the same people.

First, we visually demonstrate how well the robust approach copes with occlusion and how well it handles changes in appearance of the faces due to different facial expressions. Fig. 8 shows the pixel selection process run on the test images shown in Fig. 8a. A few iterations of the robust

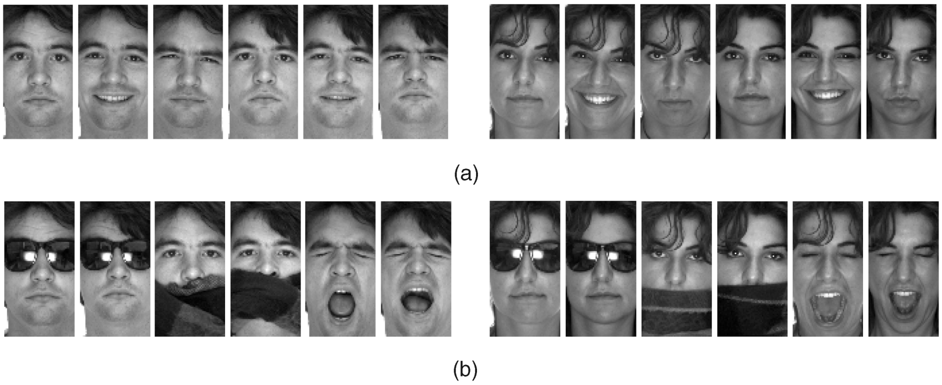


Fig. 7. Images from the AR face database. (a) Training images, from left to right: first three images in both columns are neutral, smile, and anger expressions from the first session, while the second three images depict these expressions in the same order from the second session. (b) Test images, from left to right: occlusion by sunglasses, scarf, and an image of a scream face expression from both sessions.

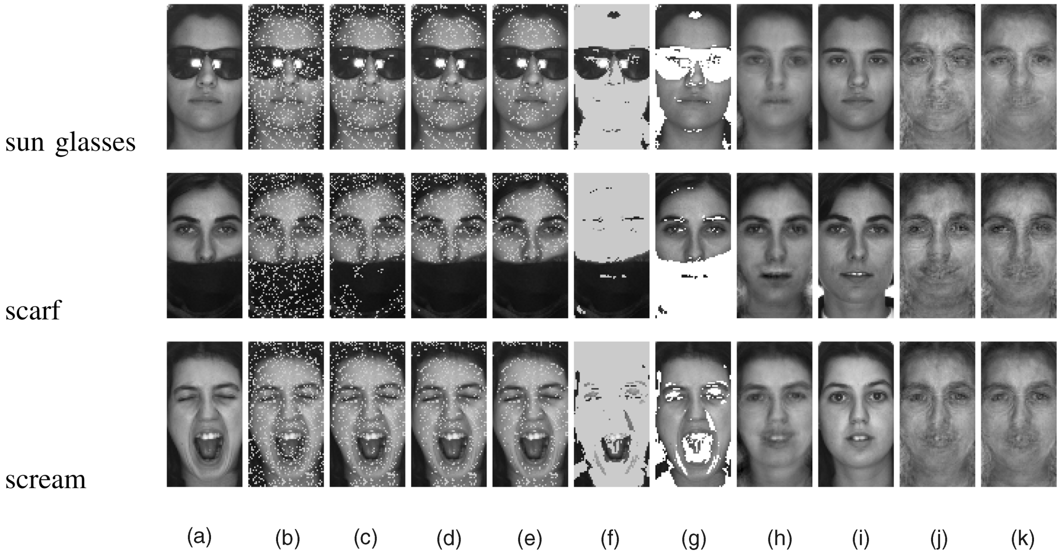


Fig. 8. Results of the pixel selection procedure. (a) Test image. (b) Random initialization in the pixel selection procedure (light-gray pixels denote the selected pixels). (c), (d), and (e) A few stages of the α-trimming method. (f) All compatible points. (g) The finally selected pixels used for coefficient calculation (denoted with the gray-value of the original test image). (h) Reconstructed image. (i) Training image of the person in (a). (j) Reconstructed image obtained with the standard LDA method. (k) Reconstructed image obtained using naive ($k = 0$) robust LDA method (see text for details).

algorithm for estimation of subspace coefficients (see the Appendix for details) are depicted in Figs. 8b, 8c, 8d, and 8e. Initially, a number of pixels was chosen randomly (denoted as light-gray pixels in Fig. 8b). The values of subspace coefficients were calculated from the selected pixels. These pixels were then subject to a few α-trimming iterations (Figs. 8c, 8d, and 8e). Finally, the remaining selected pixels were used to determine all compatible points (the ones consistent with the appearance of the training images), which are shown in Fig. 8f. As one can observe, the pixels on the sun glasses, scarves, mouth and eyes (in the "scream" test cases) were discarded since these regions significantly differ from the appearance in the training images. The subspace coefficients were afterward calculated, taking into account only the compatible pixels (Fig. 8g), thus the reconstructed image (Fig. 8h) is very similar to the training images of the (correct) subject (Fig. 8i), which results in a reliable classification. For comparison, the reconstructed images given by the standard LDA method and the method of robust coefficient estimation of [18] applied directly to the LDA basis (which is the same as taking $k = 0$ in our extended LDA approach) are given in Figs. 8j and 8k, respectively. Since these two images do not resemble the training images of the correct person, the calculated LDA coefficients subsequently also differ from the optimal ones and are, as such, more likely to incur misclassification of the test image. More examples for robust classification procedure are depicted in Fig. 9.

To quantitatively evaluate the robustness, the proposed approach was compared to the standard LDA method. Table 2 shows the results—the values of the Fisher criterion and the recognition rates for both of the approaches used. Here, the value for k was set to 37, which ensured the 85 percent variance captured by the eigenvectors. One can observe that the robust approach significantly outperformed the standard LDA (denoted as *nonrobust* in Table 2) in the case of images of subjects wearing sunglasses and scarves. The recognition rates differ more than 30 percent in favor of the proposed approach and the Fisher criterion values are also much higher for the robust LDA, confirming its stability also in the presence of real occlusion. For the "scream" images, both methods perform equally well, which is due to the following reason: Since the variations of the mouth

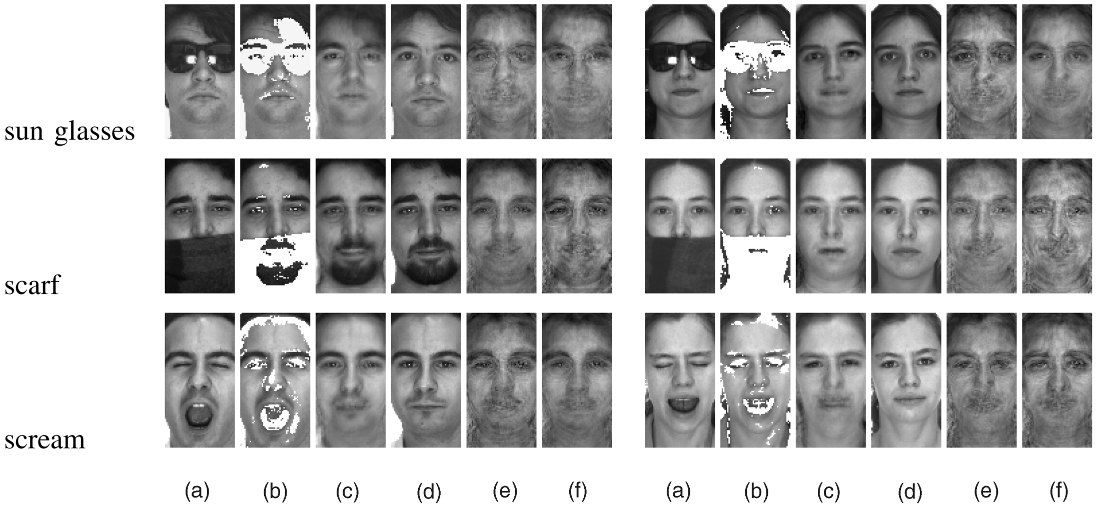


Fig. 9. Results of the pixel selection procedure for multiple people. (b)-(f) are obtained in the same way as (g)-(k) in Fig. 8, respectively.

regions are already rather large within images of each person due to different expressions in the training stage, the LDA training produced a basis that is relatively insensitive around the mouth region. This explains the rather good performance of the standard LDA for the "scream" expression. On the other hand, the nonlocal changes of the faces prevented the robust method from performing a better estimation of the coefficients, even if the outliers were determined correctly (Figs. 8 and 9, bottom row).

This experiment demonstrates the robustness power of the proposed method also in the case of images containing real occlusion. Moreover, the visual results in Figs. 8 and 9 show how well the method actually detects the outliers and how reliable the estimation of coefficients used for classification is (which is presented in terms of the reconstructed images in both figures).

6.2 Robust Estimation of CCA Coefficients

In the final experiment, we demonstrate the generality of the proposed concept by applying it to a regression method, namely CCA, and show its effectiveness in performing regression in the presence of outliers. Within this, we want to elaborate on the following issues: 1) By performing CCA estimation in the truncated PCA space (using only the first term in the classification function (9)), some significant information necessary for regression is lost, 2) to show how well the robust method performs regression in the case of occluded objects in images, and 3) to demonstrate how the robust CCA method behaves for different values of k (different amount of reconstruction information added to the CCA basis).

TABLE 2
Results on the AR Database

	Fisher criterion		recognition rate (%)	
	non-robust	robust	non-robust	robust
sun glasses	0.28	0.99	43.00	84.00
scarf	0.46	1.20	59.00	93.00
scream	0.85	0.86	87.00	87.00

The experiment was performed on a set of 120 images of a toy fish, which were taken from the views evenly distributed around the object (some examples are depicted in Fig. 10a) The task was to learn the relation between the appearances of the object and their orientations using CCA and then to use this knowledge to estimate the orientation of the object in a novel image. Every fourth image (30 images of size 64×64) was used for training, while the remaining 90 were used for testing. For each training image, its orientation (two-dimensional vector indicating the direction from which the image was taken—sine and cosine of the angle) was known. After CCA was performed on the training data, the input images were projected onto the obtained CCA vectors yielding the corresponding two-dimensional CCA coefficient vectors. A linear mapping from these coefficients to orientation vectors was estimated using the least squares minimization method. This mapping function was then used for estimation of the orientations of the test images from their canonical correlation coefficients.

To show that performing regression in the truncated PCA space does not yield reliable results, we first used the nonoccluded images also in the test stage. The results are shown in Fig. 11a and Table 3a. The plots show the actual orientations (abscissa) and the estimated orientations

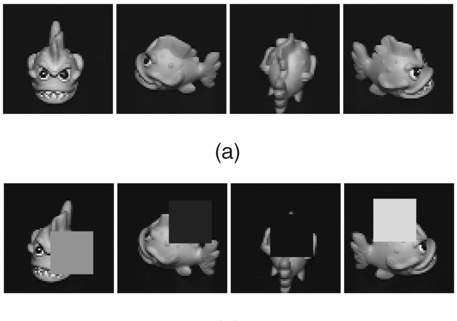
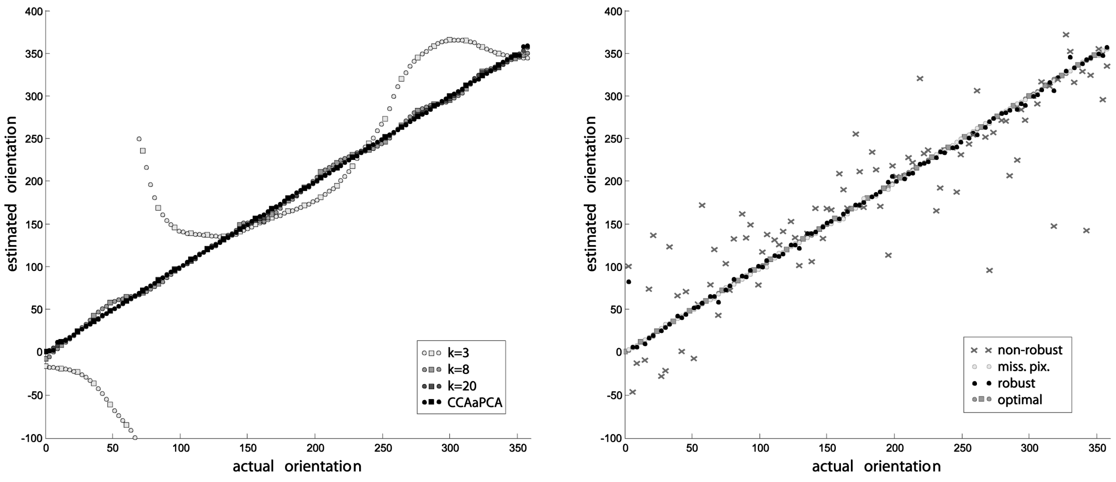


Fig. 10. (a) Four nonoccluded images. (b) Four occluded test images.



(a)

(b)

Fig. 11. Estimation of orientation: Results on (a) nonoccluded images for different k and (b) occluded test images.

(ordinate) of the object in the training images (denoted with squares) and test images (circles). The results were far from the optimal ones when only the truncated vectors of principal components were used (e.g., for $k = 3$ or $k = 8$). By increasing the number of preserved principal vectors, the results converged to the optimal ones achieved either by the proposed extended CCA basis (denoted by *CCAaPCA*) or by the standard CCA.

In the second part of the experiment, a square of a random intensity was added at a randomly chosen position in each test image (Fig. 10b). The results are presented in Fig. 11b. Here, k was set to 15 to ensure the 85 percent variance captured by the PCA basis. When the standard CCA method was used, the obtained projections of occluded test images were severely affected by the outlying pixels, thus the estimates were very inaccurate (denoted by *nonrobust*). The robust method performed significantly better. When the positions of outliers were presumed to be known and the coefficients were estimated from inliers only (denoted as $miss.pix$), the results were very close to the optimal ones, while the results were only slightly worse when the robust method also performed the outlier detection (denoted as *robust*).

TABLE 3
Mean Absolute Orientation Errors:
(a) Nonoccluded and (b) Occluded Test Images

k	error
3	42.08
5	5.17
8	3.61
10	2.25
15	1.25
20	0.97
CCAaPCA	0.67

(a)

k	miss. pix.	robust
3+2	2.62	20.33
5+2	2.33	14.71
8+2	1.72	6.81
10+2	1.56	4.51
15+2	1.22	3.40
20+2	1.14	2.56
	non-robust	36.92

(b)

Table 3b presents mean absolute orientation errors in degrees for different dimensions of the principal subspace for the cases where the positions of outliers were assumed to be known ($miss.pix.$) and not known ($robust$). One can observe that the errors in the first case are very small even when a small number of principal components were used. As expected, the results were slightly worse when the outliers were not known since they first had to be detected by the robust procedure. In this case, a higher number of principal components should be used since the top few principal components do not contain enough information for a reliable detection of outliers. However, these results are still significantly better than the results of the standard nonrobust method.

7 Conclusions

The importance of the discriminative methods has been emphasized in the literature for their strong ability of classification, which is one of the main tasks of computer vision. However, the fact that they cannot successfully cope with outliers and occlusions that commonly appear in real-world settings has severely limited their domain of applicability. The concept of robust classification using global subspace methods appears to be an elusive task and, thus, has rarely been tackled in literature.

In this paper, we proposed a method that is novel in the area of robust classification/regression. It combines the properties of both discriminative and reconstructive methods, preserving the classification power from the former and enabling robust behavior stemming from the latter. The robust approach exploits several techniques, i.e., robust estimation and the hypothesize-and-test paradigm, which, combined together in a general framework, achieve the goal. We evaluated the theoretical results on several computer vision tasks, showing that the proposed method significantly outperforms the standard discriminative methods. A general conclusion drawn from these experiments is that our robust method can tolerate much higher levels of outliers (occlusions) than the standard discriminative methods.

The applications of the proposed method are numerous. All tasks that can be accomplished by the classical linear subspace discriminative methods can also be achieved within the framework of our proposed approach, only more robustly and on more complex scenes.

Appendix

Robust Coefficient Estimation

For completeness of the paper, we briefly summarize the robust coefficient estimation procedure as developed in [18].

Let $\mathbf{x} \in \mathbb{R}^m$ be an image vector containing m pixels and U a basis matrix of a reconstructive model which provides the approximation of $\mathbf{x}$ by its reduced basis:

$$\mathbf{x} \approx a_1\mathbf{u}_1 + \cdots + a_k\mathbf{u}_k,$$

where the coefficient vector $\mathbf{a}_{\mathcal{K}} = [a_1, \cdots, a_k]^T$ is usually calculated as $\mathbf{a}_{\mathcal{K}} = (U_{:\mathcal{K}}^T U_{:\mathcal{K}})^{-1} U_{:\mathcal{K}}^T \mathbf{x}$. The problem appears when $\mathbf{x}$ contains outliers since the product $U_{:\mathcal{K}}^T \mathbf{x}$ takes into account *all* pixels, thereby also the corrupted ones, and can consequently give a wrong value for $\mathbf{a}_{\mathcal{K}}$. In order to overcome this problem, we need to robustly solve the overdetermined linear system of equations

$$\begin{aligned} x^1 &= a_1u_1^1 + a_2u_2^1 + \cdots + a_ku_k^1 \\ x^2 &= a_1u_1^2 + a_2u_2^2 + \cdots + a_ku_k^2 \\ &\vdots \\ x^m &= a_1u_1^m + a_2u_2^m + \cdots + a_ku_k^m, \end{aligned} \tag{16}$$

where $image^i$ denotes the ith pixel in an $image$ vector. Hypothetically speaking, if the approximation of $\mathbf{x}$ would be of zero error, only k equations would be needed to calculate $\mathbf{a}_{\mathcal{K}}$. But, since generally the approximation error is not zero, yet still very small, we could take into account only p, where $k < p \ll m$, equations from (16) to determine the coefficient vector $\mathbf{a}_{\mathcal{K}}$ to a satisfactory degree of accuracy. This is because the methods that provide good reconstruction of the data can exploit the redundancy present in the visual data.

The robust coefficient estimation procedure is based on a hypothesize-and-test paradigm using subsets of image pixels. The basic idea is to randomly choose a set of p pixels $\mathcal{H} \subset \{i \mid i = 1, \ldots, m\}$, $|\mathcal{H}| = p$, in the image $\mathbf{x}$ (each such choice is called a hypothesis), and take into account only these pixels when determining the coefficients. The task is to find values $\{a_j\}_{j=1}^k$ such that the expression

$$E(\mathcal{H}) = \sum_{i\in\mathcal{H}} \left(x^i - \sum_{j=1}^{k} a_j u_j^i \right)^2$$

is minimal. This is achieved by solving the linear system $G\mathbf{a} = \mathbf{d}$, where G is the $k \times k$ matrix with the entries $g_{ij} = \langle \mathbf{u}_i^{\mathcal{H}}, \mathbf{u}_j^{\mathcal{H}} \rangle$ and $\mathbf{d}$ is the vector with entries $d_j = \langle \mathbf{x}^{\mathcal{H}}, \mathbf{u}_j^{\mathcal{H}} \rangle$. Here, $\mathbf{x}^{\mathcal{H}}$ denotes the pixels in $\mathbf{x}$, which belong to hypothesis $\mathcal{H}$, i.e., $\mathbf{x}^{\mathcal{H}} = \{x^i \mid i \in \mathcal{H}\}$. The obtained coefficients $\{a_j\}_{j=1}^k$ are then used to calculate the reconstruction in the selected pixels $\mathbf{x}^{\mathcal{H}}$. Based on the error distribution in these points, their number is reduced by a factor α (α-trimming), until the maximum error falls below a predefined threshold Θ. This value is determined by considering the average reconstruction error in a single pixel of an image in the training set.

To increase the probability of determining correct coefficients, h hypotheses are generated. For each hypothesis, the robust coefficient estimation is performed. Finally, the best hypothesis is chosen based on the number of compatible pixels and the related reconstruction errors.

Acknowledgments

This research has been supported, in part, by the following funds: Research program Computer Vision P2-0214 (Slovenian Ministry of Higher Education, Science, and Technology), EU FP6-004250-IP project CoSy, EU FP6-511051 project MOBVIS, and CONEX project.

References

[1] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, “Eigenfaces versus Fisherfaces: Recognition Using Class Specific Linear Projection,” *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 711-720, July 1997.

[2] A.J. Bell and T.J. Sejnowski, “An Information Maximisation Approach to Blind Separation and Blind Deconvolution,” *Neural Computation,* vol. 7, no. 6, pp. 1129-1159, 1995.

[3] M. Borga and H. Knutsson, “Canonical Correlation Analysis in Early Vision Processing,” *Proc. Ninth European Symp. Artificial Neural Networks,* pp. 309-314, 2001.

[4] M. Borga, “Learning Multidimensional Signal Processing,” PhD thesis, Linköping Univ., Sweden, 1998.

[5] C.V. Chork and P.J. Rousseeuw, “Integrating a High-Breakdown Option into Discriminant Analysis in Exploration Geochemistry,” *J. Geochemical Exploration,* vol. 43, pp. 191-203, 1992.

[6] P. Comon, “Independent Component Analysis—A New Concept?” *Signal Processing,* vol. 36, pp. 287-314, 1994.

[7] C. Croux and C. Dehon, “Robust Linear Discrimination Analysis Using S-Estimators,” *Canadian J. Statistics,* vol. 29, pp. 473-492, 2001.

[8] F. De la Torre and M.J. Black, “A Framework for Robust Subspace Learning,” *Int'l J. Computer Vision,* vol. 54, no. 1, pp. 117-142, 2003.

[9] C. Dehon, P. Filzmoser, and C. Croux, “Robust Methods for Canonical Correlation Analysis,” *Data Analysis, Classification, and Related Methods,* pp. 321-326, Berlin: Springer-Verlag, 2000.

[10] RO. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification,* second ed. Wiley-Interscience, 2000.

[11] H. Hotelling, “Analysis of a Complex of Statistical Variables into Principal Components,” *J. Educational Psychology,* vol. 24, pp. 417-441, 1933.

[12] R.A. Fisher, “The Use of Multiple Measurements in Taxonomic Problems,” *Annals of Eugenics,* vol. 7, pp. 179-188, 1936.

[13] R. Gross, I. Matthews, and S. Baker, “Fisher Light-Fields for Face Recognition across Pose and Illumination,” *Proc. German Symp. Pattern Recognition (DAGM),* pp. 481-489, 2002.

[14] D.M. Hawkins and G.J. McLachlan, “High-Breakdown Linear Discriminant Analysis,” *J. Am. Statistical Assoc.,* vol. 92, pp. 136-143, 1997.

[15] X. He and W.K. Fung, “High Breakdown Estimation for Multiple Populations with Applications to Discriminant Analysis,” *J. Multivariate Analysis,* vol. 72, no. 2, pp. 151-162, 2000.

[16] M. Hubert and K. Van Driessen, “Fast and Robust Discriminant Analysis,” *Computational Statistics and Data Analysis,* vol. 45, pp. 301-320, 2003.

[17] D.D. Lee and H.S. Seung, “Algorithms for Non-Negative Matrix Factorization,” *Advances in Neural Information Processing Systems,* vol. 13, pp. 556-562, 2001.

[18] A. Leonardis and H. Bischof, “Robust Recognition Using Eigenimages,” *Computer Vision and Image Understanding,* vol. 78, no. 1, pp. 99-118, 2000.

[19] X. Lu, Y. Wang, and A.K. Jain, “Combining Classifiers for Face Recognition,” *Proc. IEEE Int'l Conf. Multimedia and Expo,* vol. 3, pp. 13-16, 2003.

[20] G.L. Marcialis and F. Roli, “Fusion of PCA and LDA for Face Verification,” *Proc. Post-ECCV Workshop Biometric Authentication (BIOMET),* pp. 30-37, 2002.

[21] A.M. Martinez and A.C. Kak, “PCA versus LDA,” *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 23, no. 2, pp. 228-233, Feb. 2001.

[22] P. Meer, C.V. Stewart, and D.E. Tyler, "Robust Computer Vision: An Interdisciplinary Challenge, Guest Editorial," *Computer Vision and Image Understanding 78,* vols. 1-7, 2000.

[23] O.L. Mangasarian and D.R. Musicant, "Robust Linear and Support Vector Regression," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 9, pp. 950-955, Sept. 2000.

[24] T. Melzer, M. Reiter, and H. Bischof, "Appearance Models Based on Kernel Canonical Correlation Analysis," *Pattern Recognition,* vol. 36, no. 9, pp. 1961-1973, 2003.

[25] S.K. Nayar, H. Murase, and S.A. Nene, "Parametric Appearance Representation," *Early Visual Learning,* pp. 131-160, Oxford Univ. Press, 1996.

[26] S.A. Nene, S.K. Nayar, and H. Murase, "Columbia Object Image Library (COIL-20)," Technical Report CUCS-005-96, Feb. 1996.

[27] A.M. Pires, "Robust Linear Discriminant Analysis and the Projection Pursuit Approach, Practical Aspects," *Proc. Int'l Conf, Robust Statistics,* 2001.

[28] PJ. Rousseeuw, "Multivariate Estimation with High Breakdown Point," *Math. Statistics and Applications,* vol. B, pp. 283-297, 1985.

[29] F. Samaria and A. Harter, "Parameterisation of a Stochastic Model for Human Face Identification," *Proc. Second IEEE Workshop Applications of Computer Vision,* Dec. 1994.

[30] I. Stainvas, N. Intrator, and A. Moshaiov, "Improving Recognition via Reconstruction," submitted.

[31] D.L. Swets and J. Weng, "Using Discriminant Eigenfeatures for Image Retrieval," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 18, no. 8, pp. 831-837, Aug. 1996.

[32] M. Turk and A. Pentland, "Eigenfaces for Recognition," *J. Cognitive Neuroscience,* vol. 3, no. 1, pp. 71-86, 1991.

[33] J. Yang and J.-Y. Yang, "Why Can LDA Be Performed in PCA Transformed Space?" *Pattern Recognition,* vol. 36, pp. 563-566, 2003.

[34] W. Zhao, A. Krishnaswamy, R. Chellappa, D. Swets, and J. Weng, "Discriminant Analysis of Principal Components for Face Recognition," *Face Recognition: from Theory to Applications,* Springer-Verlag, pp. 73-85, 1998.

Sanja Fidler is a doctoral student with the Faculty of Mathematics and Physics, University of Ljubljana. She received the diploma degree in mathematics from the same institution in 2002. For her thesis, she received the Faculty Award. Her research interests are probabilistic and statistical methods and their use in computer vision.

Danijel Skočaj is a researcher with the Faculty of Computer and Information Science, University of Ljubljana. He received the PhD degree from the same institution in 2003. His main research interests lie in the field of cognitive vision and include automatic modeling of objects from visual information with the emphasis on the robust and incremental appearance-based visual learning and recognition. He is a member of the IEEE and the IEEE Computer Society.

Aleš Leonardis is a full professor and the head of the Visual Cognitive Systems Laboratory with the Faculty of Computer and Information Science, University of Ljubljana. He is also an adjunct professor in the Faculty of Computer Science, Graz University of Technology. From 1988 to 1991, he was a visiting researcher in the General Robotics and Active Sensory Perception Laboratory at the University of Pennsylvania. From 1995 to 1997, he was a postdoctoral associate at the PRIP, Vienna University of Technology. He was also a visiting researcher and a visiting professor at the Swiss Federal Institute of Technology ETH in Zürich and at the Technische Fakultät der Friedrich-Alexander-Universität in Erlangen, respectively. His research interests include robust and adaptive methods for computer vision, object and scene recognition, learning, and 3D object modeling. He is an author or coauthor of more than 130 papers published in journals and conferences and he coauthored the book *Segmentation and Recovery of Superquadrics* (Kluwer, 2000). He is an associate editor of *Pattern Recognition*. He has served on the program committees of major computer vision and pattern recognition conferences. He is also a program cochair of the European Conference on Computer Vision, ECCV 2006. He has received several awards. In 2002, he coauthored a paper, "Multiple Eigenspaces," which won the 29th Annual Pattern Recognition Society award. In 2004, he was awarded the title of Ambassador of Science of the Republic of Slovenia. He is a fellow of the IAPR and a member of the IEEE and the IEEE Computer Society.