Villegas, M.; Paredes Palacios, R. (2013). On improving robustness of LDA and SRDA by using tangent vectors. Pattern Recognition Letters. 34(9):1094-1100. doi:10.1016/j.patrec.2013.03.001.

On Improving Robustness of LDA and SRDA by Using Tangent Vectors

Mauricio Villegas* and Roberto Paredes

Institut Tecnològic d'Informàtica
Universitat Politècnica de València
Camí de Vera s/n, 46022 València, Spain
{mvillegas,rparedes}@iti.upv.es

Abstract

In the area of pattern recognition, it is common for few training samples to be available with respect to the dimensionality of the representation space; this is known as the *curse of dimensionality*. This problem can be alleviated by using a dimensionality reduction approach, which overcomes the curse relatively well. Moreover, supervised dimensionality reduction techniques generally provide better recognition performance; however, several of these tend to suffer from the curse when applied directly to high-dimensional spaces. We propose to overcome this problem by incorporating additional information to supervised subspace learning techniques using what is known as *tangent vectors*. This additional information accounts for the possible differences that the sample data can suffer. In fact, this can be seen as a way to model the unseen data and make better use of the scarce training samples. In this paper, methods for incorporating tangent vector information are described for one classical technique (LDA) and one state-of-the-art technique (SRDA). Experimental results confirm that this additional information improves performance and robustness to known transformations.

Keywords:
Subspace Learning, Dimensionality Reduction, Tangent Vectors, LDA, SRDA

*Corresponding author, Tel: (+34) 963877235, Fax: (+34) 963877239

1. Introduction

In the area of pattern recognition, it is common for few training samples to be available with respect to the dimensionality of the representation space; this is known as the *curse of dimensionality* (Bellman, 1961). To handle this problem, it has become popular to use dimensionality reduction (also known as subspace learning) as a preprocessing step. However, several dimensionality reduction techniques also struggle due to the lack of samples, or in other words, they are also affected by the curse. In these cases, a tandem strategy is often used by applying a more robust technique as an initial step. This strategy, though less useful from the discriminative point of view, reduces the dimensionality down to a more appropriate size for the subsequent *discriminative* dimensionality reduction. The most well-known tandem is PCA+LDA (Yang and Yang, 2003; Yang et al., 2005), i.e., where Principal Component Analysis is performed over the original representation space and afterwards Linear Discriminant Analysis (Fukunaga, 1990) is applied. Note that PCA is an unsupervised technique, whereas LDA is supervised, which is crucial since the use of a supervised technique generally helps to boost the recognition performance considerably.

The motivation for this paper was to improve supervised subspace learning techniques so that they are able to cope with scarce data in high-dimensional feature spaces. Even though a tandem strategy overcomes the curse of dimensionality for the less robust supervised subspace learning techniques, it would clearly be more desirable for these techniques to work well in high-dimensional spaces, up to the point of not necessarily requiring a previous dimensionality reduction. This goal is addressed in this paper by considering the known transformations that a sample can exhibit which do not modify the class membership. In fact, we can consider that these known transformations model the unseen samples (as if increasing the training set), thereby overcoming the curse of dimensionality. Consider for instance the rotations and displacements of facial images due to imperfect alignments. Even though these variations are expected to appear, it is known that they do not change the identity of the person

appearing in the image. One method to account for the possible combinations of these base transformations is the *tangent distance* (Simard et al., 1993); however, it is only applicable to distance-based classifiers. In this work only the *tangent vectors* are used as a way to obtain more information from the training set, without imposing any restrictions on the classifier. Related to this paper, Schölkopf et al. (1997) and Mika et al. (1999) use the tangent vectors to improve Support Vector kernels and make them somewhat invariant to the tangent vector transformations.

The paper addresses two supervised techniques: the first is the classic LDA (including the PCA+LDA variant), and the second is the state-of-the-art Spectral Regression Discriminant Analysis (SRDA) (Cai et al., 2008; Chen et al., 2009). In the literature, there are many other methods that could be considered (see for instance Burges (2005) and van der Maaten and Postma (2009) for a review of some of them). Nevertheless, the techniques we have chosen are known to perform well and illustrate an idea which could be applied to other methods in future works.

The contributions of the paper are the following. First, we reformulate LDA so that it is expressed in terms of the covariance matrix, which can be better estimated by using tangent vectors (see Section 3.1). This modification helps to overcome the singularity problems that LDA has when there are few training samples, improves recognition performance, and also increases the robustness of the learned subspace to known transformations. Second, we present a method to incorporate the tangent vector information in SRDA that keeps the characteristic of being solvable by systems of linear equations, thus continuing to be efficient for learning (see Section 3.2). Also, the recognition performance improves and the robustness of the learned subspaces to known transformations increases. Finally, in Section 4, we present empirical results that confirm the benefits when using the proposed modifications.

2. Preliminaries and Overview of the Tangent Vectors

Suppose we have a point $\boldsymbol{x} \in \mathbb{R}^D$ generated from an underlying distribution, and that the possible transformations or manifold of $\boldsymbol{x}$ is given by $\hat{\boldsymbol{t}}(\boldsymbol{x}, \boldsymbol{\alpha})$, a function which depends on a parameter vector $\boldsymbol{\alpha} \in \mathbb{R}^L$ with the characteristic that $\hat{\boldsymbol{t}}(\boldsymbol{x}, \boldsymbol{0}) = \boldsymbol{x}$. The dimensionality of $\boldsymbol{\alpha}$ is essentially

the degrees of freedom of possible variations that $\boldsymbol{x}$ can have. In real applications, the manifold $\hat{\boldsymbol{t}}(\boldsymbol{x}, \boldsymbol{\alpha})$ is highly non-linear; however, for values close to $\boldsymbol{\alpha} = \boldsymbol{0}$, it can be reasonable to approximate it by a linear subspace. This can also be interpreted as representing the manifold by its Taylor series expansion evaluated at $\boldsymbol{\alpha} = \boldsymbol{0}$, and discarding the second and higher order terms (Simard et al., 1998), i.e.,

$$\hat{\boldsymbol{t}}(\boldsymbol{x}, \boldsymbol{\alpha}) = \hat{\boldsymbol{t}}(\boldsymbol{x}, \boldsymbol{0}) + \sum_{l=0}^{L} \alpha_l \frac{\partial \hat{\boldsymbol{t}}(\boldsymbol{x}, \boldsymbol{\alpha})}{\partial \alpha_l} + \dots \Bigg|_{\boldsymbol{\alpha}=\boldsymbol{0}} \tag{1}$$

$$\approx \boldsymbol{t}(\boldsymbol{x}, \boldsymbol{\alpha}) = \boldsymbol{x} + \sum_{l=0}^{L} \alpha_l \boldsymbol{v}_l \ . \tag{2}$$

The partial derivatives $\boldsymbol{v}_l = \partial \hat{\boldsymbol{t}} / \partial \alpha_l$ are known as the *tangent vectors*, since they are tangent to the transformation manifold $\hat{\boldsymbol{t}}$ at point $\boldsymbol{x}$.

The concept of the tangent vector approximation is illustrated in Figure 1 for a single direction of variability. As can be observed, the approximation can be quite good for small values of $\|\boldsymbol{\alpha}\|$; however, as the norm $\|\boldsymbol{\alpha}\|$ increases, the deviation from the true manifold $\hat{\boldsymbol{t}}$ is expected to increase.

When comparing two points, as a similarity measure between them, it would be ideal to use the minimum distance between their respective transformation manifolds. As an approximation to this, one can use the minimum distance between the subspaces spanned by the tangent vectors (Simard et al., 1998), which is known as the *tangent distance* (TD). The *single-sided tangent distance* considers only one of the tangent subspaces and has the advantage of being more efficient to compute (Dahmen et al., 2001). From a classification perspective, the tangent subspace can either be for the reference (RTD) or the observation (OTD).

2.1. Estimation of Tangent Vectors

There are several methods to estimate the tangent vectors, although, unfortunately, there is no general way to estimate them for every task. The most intuitive method is to use the difference between the sample and its transformation as tangent vectors. However, this method can only be used if it is possible to generate a transformation of a sample. The most well-known method of estimating tangent vectors is the one proposed

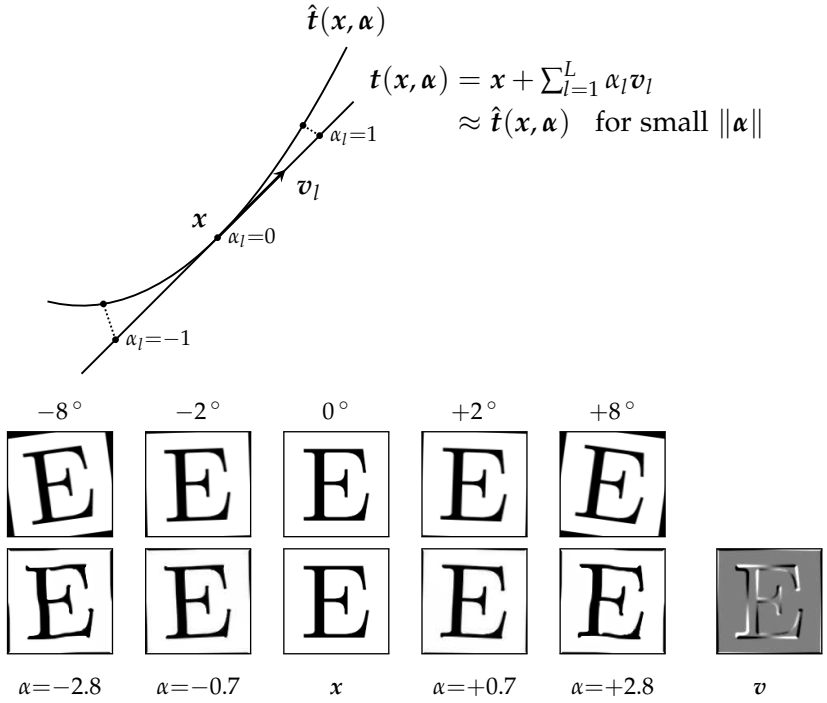

Figure 1: Top: An illustration of the linear approximation of transformations by means of tangent vectors. Bottom: An example of an image rotated at various angles and the corresponding rotation approximations using a tangent vector.

by Simard et al. (1998). This method is only applicable to image based problems, having been employed successfully to model the following: scaling, rotation, vertical and horizontal translation, parallel and diagonal hyperbolic transformations, and trace thickening.

There are other methods that try to estimate the tangent vectors from the training set, instead of adding some prior knowledge. One method of this type is presented in Keysers et al. (2004), which is based on maximum likelihood estimation. Another method is to use the difference between a sample and its nearest neighbors from the same class as tangent vectors.

The methods of Simard and the nearest neighbors were used in the experiments. However, as discussed in Section 3 and empirically observed, the latter is less useful since it does not provide as much additional information and it does not help to overcome the singularity problems.

3. Tangent Vectors in Subspace Learning

3.1. Tangent Vectors in LDA

The objective of LDA is that the obtained subspace should discriminate the classes well. To this end, LDA simultaneously maximizes the distances between the class centers (between-class scatter matrix) and minimizes the distances within each class (within-class scatter matrix). It is straightforward to reformulate LDA so that it is stated in terms of the covariance matrix $\mathbf{\Sigma_x}$ and a normalized between-class scatter matrix $\mathbf{\Sigma}_\mu$. The objective function is then

$$\hat{\boldsymbol{B}} = \arg\max_{\boldsymbol{B}} \frac{\mathrm{Tr}(\boldsymbol{B}^\mathsf{T}\mathbf{\Sigma}_\mu\boldsymbol{B})}{\mathrm{Tr}(\boldsymbol{B}^\mathsf{T}\mathbf{\Sigma_x}\boldsymbol{B})} \ . \tag{3}$$

The solution of the LDA objective (3) is the following generalized eigenvalue decomposition

$$\mathbf{\Sigma}_\mu\boldsymbol{B} = \mathbf{\Sigma_x}\boldsymbol{B}\mathbf{\Lambda} \ , \tag{4}$$

with $\mathbf{\Lambda}$ being a diagonal matrix of generalized eigenvalues and the columns of $\boldsymbol{B}$ being the generalized eigenvectors.

By having a solution of LDA in terms of the covariance matrix, for a given dataset $\mathcal{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, we are able to use a better empirical estimation for $\mathbf{\Sigma_x}$ that considers tangent vectors (Keysers et al., 2004) given by

$$\mathbf{\Sigma}_{\mathcal{X}\cup\mathcal{V}} = \mathbf{\Sigma}_\mathcal{X} + \mathbf{\Sigma}_\mathcal{V} \tag{5}$$

$$= \mathbf{\Sigma}_\mathcal{X} + \frac{1}{L}\sum_{l=1}^{L}\frac{\gamma_l^2}{|\mathcal{V}_l|}\sum_{\forall \boldsymbol{v}_l\in\mathcal{V}_l}\boldsymbol{v}_l\boldsymbol{v}_l^\mathsf{T} \ , \tag{6}$$

where $\mathbf{\Sigma}_\mathcal{X}$ is the usual estimation of the covariance matrix and $\mathcal{V}_l$ is the set that includes all of the tangent vectors of type l. The weights $\gamma_1, \ldots, \gamma_L$ account for the importance of each tangent vector type depending on the distributions $p(\alpha_1), \ldots, p(\alpha_L)$; however, in practice, these are parameters that need to be estimated. A good rule of thumb is to set them equal for all tangent types and to sample them so that $\mathrm{Tr}(\mathbf{\Sigma}_\mathcal{V})$ is a certain fraction of $\mathrm{Tr}(\mathbf{\Sigma}_\mathcal{X})$. This improved version of LDA will be referred to as TLDA.[1]

[1] Matlab/Octave implementations available at http://mvillegas.info/research

A better estimation for $\mathbf{\Sigma}_{\boldsymbol{\mu}}$ is not obtained since it is reasonable to assume that the distributions $p(\alpha_1), \ldots, p(\alpha_L)$ are symmetric, a case in which the tangent vectors cancel out. Intuitively, this makes sense because the tangent vectors give information about how a sample might vary and not about the other classes.

One of the shorthands of classical LDA is that if the dimensionality of the vectors is higher than the number of training samples, then $\mathbf{\Sigma}_{\mathcal{X}}$ is singular and therefore there is no unique solution to the generalized eigenvalue problem. With the new estimation of $\mathbf{\Sigma}_{\mathbf{x}}$, if the tangent vectors are linearly independent from the training samples (which is the case for Simard's method), the rank goes from N for $\mathbf{\Sigma}_{\mathcal{X}}$ to $N(L+1)$ for $\mathbf{\Sigma}_{\mathcal{X}\cup\mathcal{V}}$, which can drastically reduce the number of samples that are necessary to avoid singularities. Thus, for LDA, the estimation (6) of $\mathbf{\Sigma}_{\mathbf{x}}$ provides for LDA a way to take better advantage of the limited training data, normally avoiding a singularity and making the solution more robust to the known class-invariant transformations. This is observed in the experimental results.

A popular method to overcome the singularity problem in classical LDA is to previously reduce dimensionality by means of PCA. Since PCA is also based on the estimation of the covariance matrix, it can also be improved by including tangent vector information, i.e., TPCA. Even though TLDA overcomes the singularity problem, TLDA combined with a previous reduction by TPCA gives very good performance as can be observed in the experiments section.

The computational complexity of the method basically depends on the tangent vector estimation technique. In the case of Simard's method, the computation of the tangent vectors are $\mathcal{O}(KDN)$ with $K \ll D$ being the size of the convolution kernel. In comparison to the complexity of the eigenvalue decomposition, this additional cost is not significant.

3.2. Tangent Vectors in SRDA

SRDA is similar to LDA. In fact, the starting optimization criterion is the same as LDA and, thus, the solution is given by Equation (4). However, to make the computation fast, SRDA avoids the need to do an eigenvalue decomposition. Observe that Equation (4) can be rewritten as

$$\bar{\boldsymbol{X}}\boldsymbol{W}\bar{\boldsymbol{X}}^{\mathsf{T}}\boldsymbol{B} = \bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^{\mathsf{T}}\boldsymbol{B}\boldsymbol{\Lambda}\,, \tag{7}$$

where $\bar{\boldsymbol{X}} \in \mathbb{R}^{D \times N}$ is the centered data matrix ordered by classes, i.e., $\bar{\boldsymbol{X}} = [\boldsymbol{x}_{1,1} - \boldsymbol{\mu}, \ldots, \boldsymbol{x}_{C,N_C} - \boldsymbol{\mu}]$, and $\boldsymbol{W} \in \mathbb{R}^{N \times N}$ is a matrix composed mostly of zeros, having the matrices $\boldsymbol{W}_c \in \mathbb{R}^{N_c \times N_c}$ for $c = 1, \ldots, C$ along the diagonal, in which the elements of $\boldsymbol{W}_c$ are all equal to N_c^{-1}, with N_c being the number of samples of class c.

As presented in Cai et al. (2008), the problem in Equation (4) can be tackled by solving the system of linear equations $\boldsymbol{Y} = \bar{\boldsymbol{X}}^\mathsf{T}\boldsymbol{B}$, where $\boldsymbol{Y} \in \mathbb{R}^{N \times C-1}$ is a matrix whose columns are $C-1$ eigenvectors of $\boldsymbol{W}$ which are orthogonal to the vector of ones. Matrix $\boldsymbol{W}$ has exactly C nonzero eigenvalues all equal to one, and obtaining a set of eigenvectors is trivial. In fact, for each class c, there is an eigenvector with elements equal to $N_c^{-1/2}$ in the positions of the samples of class c and zeros for all other elements. Therefore, finding $\boldsymbol{Y}$ reduces to orthogonalize these trivial eigenvectors with respect to the vector of ones. However, the system $\boldsymbol{Y} = \bar{\boldsymbol{X}}^\mathsf{T}\boldsymbol{B}$ may not have a solution, so it was proposed to find the $\boldsymbol{B}$ that best fits the equation in a regularized least squares sense. The optimization function of SRDA is then given by

$$\hat{\boldsymbol{B}} = \arg\min_{\boldsymbol{B}} \mathrm{Tr}\left[(\bar{\boldsymbol{X}}^\mathsf{T}\boldsymbol{B} - \boldsymbol{Y})^\mathsf{T}(\bar{\boldsymbol{X}}^\mathsf{T}\boldsymbol{B} - \boldsymbol{Y}) + \rho\boldsymbol{B}^\mathsf{T}\boldsymbol{B}\right], \tag{8}$$

where ρ is the regularization parameter.

Incorporating the tangent vectors into SRDA cannot be done in the same way as for LDA (Section 3.1), since SRDA does not directly use the covariance matrix. Furthermore, the modification should be done in such a way that the problem can still be solved as systems of linear equations, thus retaining the advantages of SRDA. Using matrix notation, Equation (6) can be written as

$$\boldsymbol{\Sigma}_{\mathcal{X}\cup\mathcal{V}} = \frac{1}{|\mathcal{X}|}\bar{\boldsymbol{X}}\bar{\boldsymbol{X}}^\mathsf{T} + \frac{\gamma^2}{|\mathcal{V}|}\boldsymbol{V}_\mathcal{X}\boldsymbol{V}_\mathcal{X}^\mathsf{T}, \tag{9}$$

where the columns of matrix $\boldsymbol{V}_\mathcal{X} \in \mathbb{R}^{D \times NL}$ are all the tangent vectors of the training samples in $\mathcal{X}$. With a few manipulations, it can be shown that, when including the tangent vectors, the generalized eigenvalue problem can be expressed as

$$\boldsymbol{Z}\boldsymbol{W}'\boldsymbol{Z}^\mathsf{T}\boldsymbol{B} = \boldsymbol{Z}\boldsymbol{Z}^\mathsf{T}\boldsymbol{B}\boldsymbol{\Lambda}, \tag{10}$$

where $\boldsymbol{Z} = [\bar{\boldsymbol{X}}\ \gamma \boldsymbol{V}_{\mathcal{X}}]$, and the new weight matrix is the same as $\boldsymbol{W}$, but $\boldsymbol{W}'$ is padded with NL rows and columns of zeros at the bottom and on the right. The eigenvalues of $\boldsymbol{W}'$ are also only C all equal to one, and the eigenvectors are the same as for $\boldsymbol{W}$, but padded with NL zeros at the bottom. The final optimization function for SRDA including the tangent vector information is then given by

$$\hat{\boldsymbol{B}} = \arg\min_{\boldsymbol{B}} \mathrm{Tr}\left[(\boldsymbol{Z}^\mathsf{T}\boldsymbol{B} - \boldsymbol{Y}')^\mathsf{T}(\boldsymbol{Z}^\mathsf{T}\boldsymbol{B} - \boldsymbol{Y}') + \rho \boldsymbol{B}^\mathsf{T}\boldsymbol{B}\right], \tag{11}$$

where the columns of matrix $\boldsymbol{Y}'$ are $C-1$ eigenvectors of $\boldsymbol{W}'$ which are orthogonal to the vector composed of ones for the first N elements and zeros for the remaining NL.

The optimization function for the Tangent Vector SRDA (TSRDA)[1] is basically the same as for the original SRDA. The computational complexity increases by the computation of the tangent vectors, and the linear systems increase in number of equations, but with the same number of unknowns. Thus, it continues to be a very efficient algorithm.

4. Experiments

The proposed approaches have been assessed with three facial image recognition tasks: gender recognition, expression recognition, and identification.[2] Although only image problems were considered, the same approach can be applied in other tasks as long as the known transformations can be represented as tangent vectors. The gender dataset that we used is composed of 1892 samples of 32×40 pixel images and the objective for this dataset is to determine the gender, male or female (Villegas and Paredes, 2011). The expressions dataset is composed of 1929 samples of 32×32 pixel images obtained from the databases: Cohn-Kanade (Kanade et al., 2000) (using the emotion labels from Buenaposada et al. (2008)), AR (Martinez and Benavente, 1998), and JAFFE (Lyons et al., Dec 1999). The objective for this dataset is to classify the facial expression as one of 7 possibilities: a neutral expression, a person screaming, or one of the six basic emotions (Donato et al., Oct 1999). The performances for

[2]Datasets available at `http://mvillegas.info/research`

both datasets were estimated using a 5-fold cross-validation. Finally, for face identification, the dataset and evaluation protocol were the same as in Zhao et al. (2007), although with an image size of 32×40. A special characteristic of this dataset is that the 116 subjects used for the subspace learning are different from the 200 subjects used in the testing phase (10 images per subject). Therefore, the goal for this dataset is to learn a subspace that is suitable for face identification in general.

Since the objective of the experimentation was to compare the subspace learning methods with and without the tangent vector information, we kept the feature set and classifier constant. A k-NN classifier with the Euclidean distance in the learned subspace was used on all tasks; however, for the face identification task, the tangent distances were also tested, applying the same learned projection on the tangents. This shows that the tangent distances can also be applied in a discriminative subspace, which can further improve the recognition results.

The tangent vectors were estimated using the method from Simard et al. (1998) for modeling horizontal and vertical translation, rotation, and scaling. Tangent vectors estimated by nearest neighbors were also used, which illustrates the behavior of the algorithms with tangent vectors that are linearly dependent on the training set. To indicate which tangent types were employed for learning, in the results, a sub-index is added to the method acronym showing "hvrs" for Simard's method and "kθ" for $\theta = \{1, 2, 4, 8, 16, 32\}$ nearest neighbors. The γ parameter was set to be the same for all tangent types and was varied as explained in Section 3.1. The intermediate PCA/TPCA and the final subspace dimensionalities were varied between 1 and 256 and the best result for each technique is the one presented. The ρ for SRDA was varied and adjusted like the other parameters, and did not rely on an automatic method (Chen et al., 2009).

As a reference and to show that supervised subspace learning methods generally perform poorly when applied directly to high-dimensional feature vectors, results for other linear and supervised techniques are included: namely, Marginal Fisher Analysis (MFA) (Yan et al., 2007), Locality Sensitive Discriminant Analysis (LSDA) (Cai et al., 2007), Supervised Locality Preserving Projections (SLPP) (Zheng et al., 2007), and Nonparametric Discriminant Analysis (NDA) (Bressan and Vitrià, 2003). The implementations used for MFA, LSDA, SLPP, and SRDA were written by

D. Cai and for NDA was written by J. Vitrià.

4.1. Recognition Performance

The results of the experiments for each dataset are presented in tables 1, 2, and 3, respectively. Included are the estimated classification error rates, the subspace dimensionalities, and the training and testing times. Note that the training times agree with the complexity analysis in Section 3; however, the computation of the tangents was not optimized, unlike the eigenvalue decomposition. Thus, the execution time of the proposed methods could be significantly improved, especially considering that it is easily parallelizable.

As can be observed in the tables, in several cases the proposed methods show a statistically significant improvement when using Simard's tangents; in the rest of the cases, the performance remains the same. For the gender dataset (Table 1), no results are shown for LDA or $\text{TLDA}_{\text{k}\theta}$ because the covariance matrix is singular. The proposed $\text{TLDA}_{\text{hvrs}}$ was not affected by this singularity problem and, except for SRDA, obtains a significantly better performance than all baseline techniques. The performance of TLDA improves further when using a previous step of TPCA, even when using tangent vectors obtained by nearest neighbors. Still, the best result was when Simard's tangents were used. In the expressions dataset (Table 2), the improvement of $\text{TLDA}_{\text{hvrs}}$ with respect to LDA was enormous, up to the point of being statistically the same as with PCA preprocessing. In the case of SRDA, an improvement was also observed when using tangents, although with a lower confidence level. Note that for both TLDA and TSRDA, there was no further improvement with PCA preprocessing; nevertheless, this is not necessarily a bad thing. Ideally, the supervised dimensionality reduction should be done in the original feature space so that no discriminative information is discarded. If the method works well in the original space, it is not expected to get better results when using PCA preprocessing. In the identification dataset (Table 3), very competitive recognition rates were achieved, particularly with TSRDA, even though a simple pixel based representation was used in comparison to the Local Binary Pattern (LBP), which is known to be a better representation for face recognition (Ahonen et al., 2006). For this dataset, the improvement comparing SRDA and TSRDA was statistically

significant; however, the results for TLDA remained the same. Note that the use of the tangent distance improved the performance further, confirming that the tangent vectors are also useful in the learned subspace.

Regarding the combination of different types of tangents (i.e., hvrs+kθ), we did not observe that the performance improved further when more tangent vector types were used. The reason for this may have been setting the γ parameter to be the same for all tangent vectors. The problem of using different factors for each tangent type and automatically obtaining them is a topic that needs further research.

4.2. Robustness to Known Transformations

To analyze the robustness to the transformations considered during learning, figures 2, 3, and 4 present plots of the relative improvement in recognition of the proposed methods. These plots compare the methods with or without the tangents as the test samples are artificially transformed. All the graphs have the same y-axis range in order to make it easier to compare them with each other.

For the gender dataset, only LDA shows marginal improvement. However, it must be taken into account that gender recognition is a two-class problem and the target space dimensionality is only one, so it is a challenging task. In contrast, for the other two datasets (expressions and identification), a much greater improvement is observed. As the transformations of the images grow, when the tangents are used (TPCA+TLDA$_{\text{hvrs}}$ and TSRDA$_{\text{hvrs}}$), each time the relative improvement is greater than without them (PCA+LDA and SRDA). This shows that the subspaces learned with the tangents are more robust to these transformations. This is observed even though the Euclidean distance is employed, which is not invariant to these transformations. This effect can be explained by the fact that without the tangents, those directions of variability are not taken into account and might be removed by the learned projection. Interestingly, a better transformation invariance was obtained for LDA than for SRDA on the three datasets.

5. Conclusions

In this paper, two supervised dimensionality reduction techniques, Linear Discriminant Analysis (LDA) and the state-of-the-art Spectral Re-

gression Discriminant Analysis (SRDA), have been analyzed and modifications that use the tangent vector information have been proposed for them.

Experiments were conducted using three facial image recognition tasks: gender, expressions, and identification. Depending on the dataset, the proposed methods, TLDA and TSRDA, provided a significantly better recognition performance, and in no case did we observe a degradation. In the case of TLDA, the singularity of the covariance matrix due to the limited number of training samples was avoided. Another result was that the subspaces learned tended to be more robust to the transformations that were used during learning. Furthermore, an additional gain was obtained by using the same dimensionality reduction for the tangents and by employing a tangent distance.

The parameters $\gamma_1, \ldots, \gamma_L$, which weight the importance that is given to each of the tangent types (e.g. rotation, nearest neighbor, etc.), were set to be the equal, i.e., $\gamma_1 = \gamma_2 = \ldots = \gamma_L$. This seemed to prevent further improvements when combining several tangent types. The γ parameters could be set to be different for each tangent type and adjusted using cross-validation or a Markov Chain Monte Carlo. By doing this, better results would be expected; however, the cost of the algorithm could increase dramatically. A future direction for research is to develop a better method to find adequate values for these parameters. Another interesting idea to explore would be to use an optimization criterion based on the single-sided tangent distance, which can eliminate the need to estimate such parameters.

The same ideas presented here could be used to improve other existing subspace learning methods. Interesting examples are the works by Yang et al. (2011) and Villegas and Paredes (2011), which try to learn the subspace while also considering what the subsequent classifier will be. Another example would be the family of methods based on preserving the neighborhood such as the recent work of Gui et al. (2012).

Another important direction for future research is the development of methods for estimating tangent vectors for other tasks so that these techniques can be used in more applications, not only the ones related to image recognition. We plan to apply this approach to the bag-of-words representation to model word-count variability.

Table 1: Results for the face gender recognition dataset.

Approach	**Error Rate (%)** [95% conf. int.][†]		**Dim.**	**Tr. time [s]**	**Test time [ms]**
Orig. Space	19.6	[17.9 − 21.4]	1280	-	0.880
PCA	17.7	[16.0 − 19.4]	64	9.13	0.356
MFA	35.7	[33.6 − 37.9]	2	6.21	0.323
LSDA	35.7	[33.6 − 37.9]	2	8.81	0.321
SLPP	34.0	[31.9 − 36.2]	1	5.28	0.328
NDA	29.6	[27.6 − 31.7]	24	75.00	0.389
PCA+LDA	16.2	[14.6 − 18.0]	1	9.21	0.330
SRDA	10.4	[9.1 − 11.9]	1	0.57	0.347
TPCA+TLDA$_{\text{k16}}$	11.3	[10.0 − 12.8]	1	18.75	0.353
TPCA+TLDA$_{\text{hvrs+k4}}$	11.0	[9.7 − 12.5]	1	15.49	0.387
TPCA+TLDA$_{\text{hvrs}}$	10.6	[9.3 − 12.1]	1	12.66	0.380
TPCA+TSRDA$_{\text{hvrs}}$	11.0	[9.7 − 12.5]	1	11.88	0.349
TPCA+TSRDA$_{\text{k8}}$	10.8	[9.5 − 12.5]	1	14.50	0.329
TPCA+TSRDA$_{\text{hvrs+k8}}$	10.8	[9.5 − 12.3]	1	17.30	0.332
TLDA$_{\text{hvrs+k4}}$	15.3	[13.7 − 17.0]	1	54.87	0.357
TLDA$_{\text{hvrs}}$	13.3[‡]	[11.8 − 14.9]	1	53.48	0.335
TSRDA$_{\text{hvrs}}$	10.5	[9.2 − 12.0]	1	2.45	0.322
TSRDA$_{\text{k8}}$	10.3	[9.0 − 11.8]	1	4.56	0.332
TSRDA$_{\text{hvrs+k8}}$	10.2	[8.9 − 11.6]	1	7.10	0.328

[†]Wilson interval estimation.

[‡]TLDA$_{\text{hvrs}}$ better than all baselines (except SRDA) for a confidence level of 99%.

Table 2: Results for the 8 facial expressions dataset.

Approach	**Error Rate (%)** [95% conf. int.][†]		**Dim.**	**Tr. time [s]**	**Test time [ms]**
Orig. Space	35.6	[33.5 − 37.8]	1024	-	0.543
PCA	30.6	[28.5 − 32.6]	16	6.82	0.115
LSDA	80.7	[78.9 − 82.4]	1	4.64	0.124
SLPP	30.2	[28.2 − 32.3]	7	3.01	0.124
MFA	63.7	[61.6 − 65.7]	16	2.71	0.123
NDA	50.9	[48.6 − 53.1]	64	55.53	0.160
LDA	30.2[‡]	[28.2 − 32.3]	7	39.98	0.118
SRDA	20.0[§]	[18.2 − 21.8]	7	0.46	0.116
TPCA+TLDA$_{\text{hvrs}}$	20.9	[19.1 − 22.7]	7	9.21	0.120
TPCA+TLDA$_{\text{k16}}$	22.1	[20.2 − 23.9]	7	14.48	0.125
TPCA+TLDA$_{\text{hvrs+k8}}$	20.4	[18.6 − 22.2]	7	13.24	0.146
TPCA+TSRDA$_{\text{hvrs}}$	20.6	[18.8 − 22.4]	7	9.80	0.122
TPCA+TSRDA$_{\text{k2}}$	21.6	[19.7 − 23.4]	7	8.63	0.119
TPCA+TSRDA$_{\text{hvrs+k4}}$	19.8	[18.0 − 21.5]	7	11.12	0.114
TLDA$_{\text{hvrs}}$	21.4[‡]	[19.6 − 23.2]	7	40.55	0.127
TLDA$_{\text{k4}}$	28.5	[26.5 − 30.6]	7	41.71	0.126
TLDA$_{\text{hvrs+k4}}$	20.6[‡]	[18.8 − 22.4]	7	44.16	0.117
TSRDA$_{\text{hvrs}}$	18.4[§]	[16.6 − 20.1]	7	1.84	0.110
TSRDA$_{\text{k8}}$	19.7	[17.9 − 21.5]	7	3.53	0.136
TSRDA$_{\text{hvrs+k4}}$	19.5	[17.7 − 21.3]	7	3.94	0.134

[†]Wilson interval estimation.

[‡]TLDA$_{\text{hvrs}}$ better than LDA for a confidence level of 99%.

[§]TSRDA$_{\text{hvrs}}$ better than SRDA for a confidence level of 90%.

Table 3: Results for the face identification dataset.

Approach	Dist.	Error Rate (%) [95% conf. int.][†]		Dim.	Tr. time [s]	Test time [ms]
Laplacianfaces	Euc.	15.0	[13.5 – 16.6]	100	Unk.[‡]	Unk.[‡]
L-Fisherfaces	Euc.	9.5	[8.3 – 10.9]	140	Unk.[‡]	Unk.[‡]
LBP + Dual LLD	χ^2	7.4	[6.3 – 8.6]	500	Unk.[‡]	Unk.[‡]
Orig. Space	Euc.	29.3	[27.3 – 31.3]	1280	-	0.811
PCA	Euc.	31.2	[29.2 – 33.3]	256	7.49	0.196
MFA	Euc.	53.2	[51.0 – 55.4]	16	3.26	0.118
LSDA	Euc.	86.9	[85.4 – 88.3]	16	10.16	0.126
SLPP	Euc.	86.3	[84.7 – 87.7]	16	9.94	0.145
NDA	Euc.	14.9	[13.4 – 16.5]	115	26.80	0.159
PCA+LDA	Euc.	6.1	[5.1 – 7.2]	64	7.87	0.137
SRDA	Euc.	7.3[§]	[6.2 – 8.5]	115	0.89	0.145
TPCA+TLDA$_{\text{hvrs}}$	Euc.	6.3	[5.3 – 7.5]	64	13.41	0.162
TPCA+TLDA$_{\text{hvrs}}$	TD	6.1	[5.1 – 7.2]	64	13.41	95.5
TPCA+TLDA$_{\text{hvrs}}$	OTD	5.9	[4.9 – 7.0]	64	13.41	1.088
TPCA+TLDA$_{\text{hvrs}}$	RTD	5.9	[4.9 – 7.0]	64	13.41	0.762
TSRDA$_{\text{hvrs}}$	Euc.	6.9[♯]	[5.9 – 8.1]	115	2.76	0.191
TSRDA$_{\text{hvrs}}$	TD	5.1[§♯]	[4.2 – 6.2]	115	2.76	101.5
TSRDA$_{\text{hvrs}}$	OTD	6.3	[5.3 – 7.5]	115	2.76	2.274
TSRDA$_{\text{hvrs}}$	RTD	6.3	[5.3 – 7.5]	115	2.76	1.672

[†]Wilson interval estimation.
[‡]Result obtained from Zhao et al. (2007), execution times are not available.
[§]TSRDA$_{\text{hvrs,TD}}$ better than SRDA$_{\text{Euc.}}$ for a confidence level of 95%.
[♯]TSRDA$_{\text{hvrs,TD}}$ better than TSRDA$_{\text{hvrs,Euc.}}$ for a confidence level of 95%.

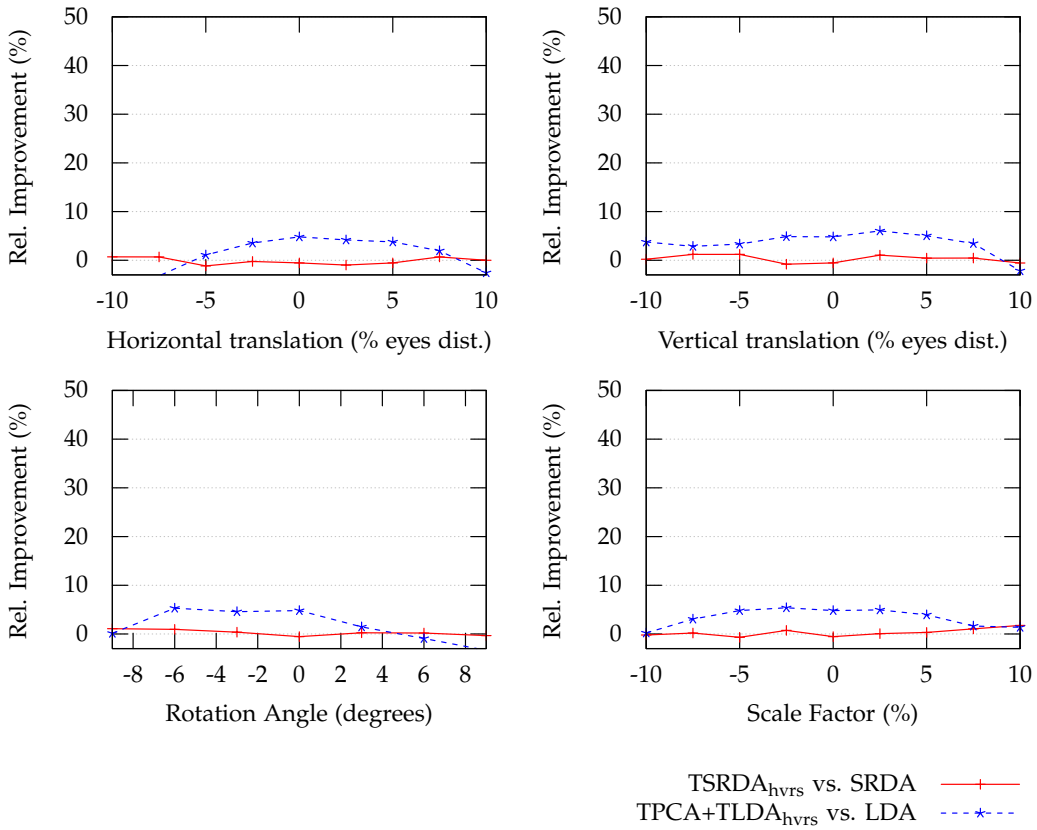

Figure 2: Relative improvement for face gender recognition when varying for the test samples: horizontal/vertical translation, angle of rotation, and scaling factor. Classification is done with the Euclidean distance.

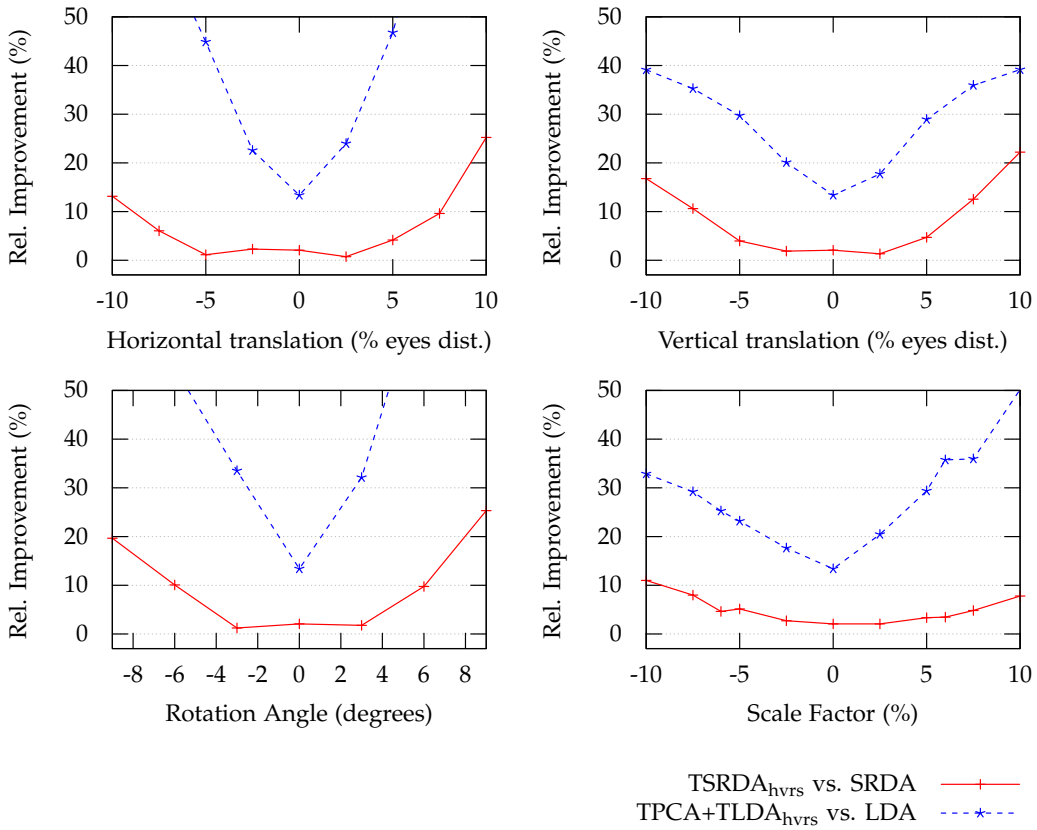

Figure 3: Relative improvement for face expression recognition when varying for the test samples: horizontal/vertical translation, angle of rotation, and scaling factor. Classification is done with the Euclidean distance.

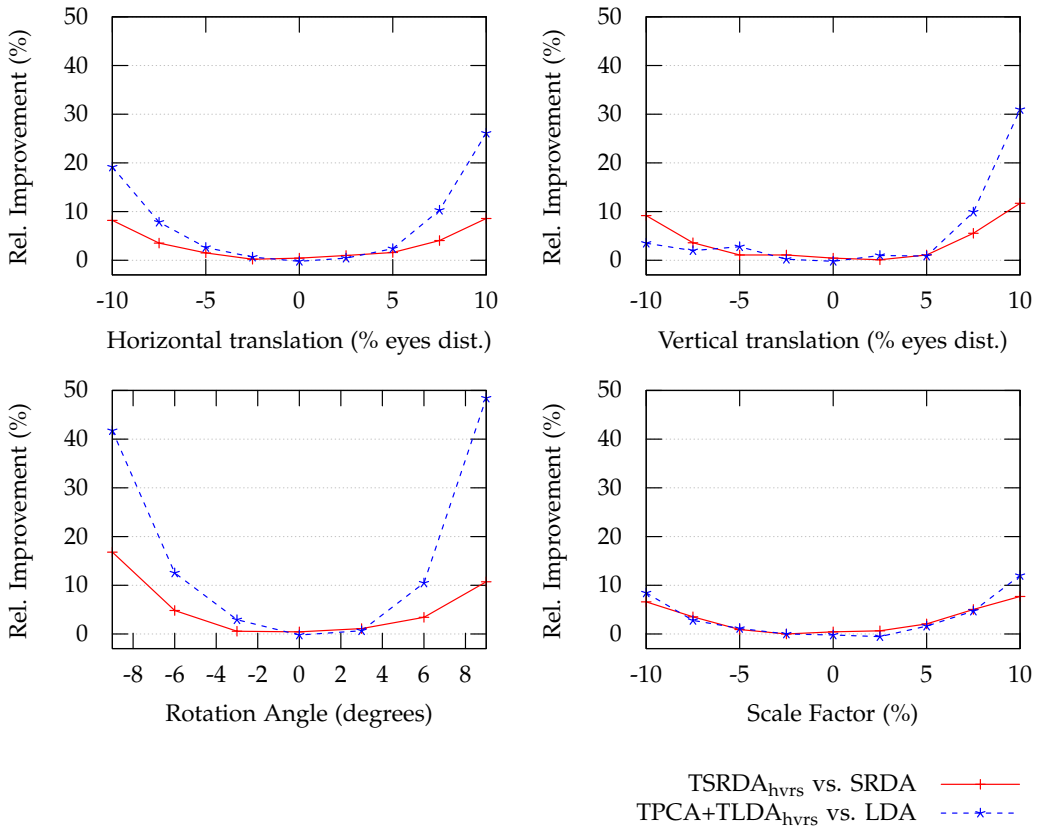

Figure 4: Relative improvement for face identification when varying for the test samples: horizontal/vertical translation, angle of rotation, and scaling factor. Classification is done with the Euclidean distance.

Acknowledgments

Work partially supported through the EU 7th Framework Programme grant tranScriptorium (Ref: 600707), by the Spanish MEC under the STraDA research project (TIN2012-37475-C02-01) and by the Generalitat Valenciana under grant Prometeo/2009/014.

References

Ahonen, T., Hadid, A., Pietikainen, M., 2006. Face description with local binary patterns: Application to face recognition. IEEE Trans. Pattern Anal. Mach. Intell. 28, 2037–2041. doi:10.1109/TPAMI.2006.244. 11

Bellman, R., 1961. Adaptive Control Processes: A Guided Tour. Princeton University Press. 2

Bressan, M., Vitrià, J., 2003. Nonparametric discriminant analysis and nearest neighbor classification. Pattern Recognit. Lett. 24, 2743–2749. doi:10.1016/S0167-8655(03)00117-X. 10

Buenaposada, J., M. Muñoz, E., Baumela, L., 2008. Recognising facial expressions in video sequences. Pattern Anal. Appl. 11, 101–116. doi:10.1007/s10044-007-0084-8. 9

Burges, C.J.C., 2005. Geometric methods for feature extraction and dimensional reduction - a guided tour, in: The Data Mining and Knowledge Discovery Handbook. Springer, pp. 59–92. 3

Cai, D., He, X., Han, J., 2008. Srda: An efficient algorithm for large-scale discriminant analysis. IEEE Trans. on Knowl. and Data Eng. 20, 1–12. 3, 8

Cai, D., He, X., Hu, Y., Han, J., Huang, T., 2007. Learning a spatially smooth subspace for face recognition, in: Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on, pp. 1–7. doi:10.1109/CVPR.2007.383054. 10

Chen, W., Shan, C., De Haan, G., 2009. Optimal regularization parameter estimation for spectral regression discriminant analysis. IEEE Trans.

Cir. and Sys. for Video Technol. 19, 1921–1926. doi:10.1109/TCSVT.2009.2026953. 3, 10

Dahmen, J., Keysers, D., Ney, H., Güld, M.O., 2001. Statistical image object recognition using mixture densities. J. Math. Imaging Vis. 14, 285–296. doi:10.1023/A:1011242314266. 4

Donato, G., Bartlett, M., Hager, J., Ekman, P., Sejnowski, T., Oct 1999. Classifying facial actions. Transactions on Pattern Analysis and Machine Intelligence 21, 974–989. doi:10.1109/34.799905. 9

Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition. 2nd ed., Academic Press. 2

Gui, J., Sun, Z., Jia, W., Hu, R., Lei, Y., Ji, S., 2012. Discriminant sparse neighborhood preserving embedding for face recognition. Pattern Recogn. 45, 2884–2893. doi:10.1016/j.patcog.2012.02.005. 13

Kanade, T., Tian, Y., Cohn, J.F., 2000. Comprehensive database for facial expression analysis, in: FG '00: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition 2000, IEEE Computer Society, Washington, DC, USA. pp. 46–53. 9

Keysers, D., Macherey, W., Ney, H., Dahmen, J., 2004. Adaptation in statistical pattern recognition using tangent vectors. IEEE Trans. Pattern Anal. Mach. Intell. 26, 269–274. doi:10.1109/TPAMI.2004.1262198. 5, 6

Lyons, M., Budynek, J., Akamatsu, S., Dec 1999. Automatic classification of single facial images. IEEE Trans. Pattern Anal. Mach. Intell. 21, 1357–1362. doi:10.1109/34.817413. 9

van der Maaten, L., Postma, E., 2009. Dimensionality Reduction: A Comparative Review. Technical Report 2009âĂŞ005. Tilburg University. 3

Martinez, A., Benavente, R., 1998. The AR Face Database. CVC Technical Report #24. 9

Mika, S., Rätsch, G., Weston, J., Schölkopf, B., Smola, A.J., Müller, K.R., 1999. Invariant feature extraction and classification in kernel spaces, in:

Advances in Neural Information Processing Systems 12, [NIPS Conference, Denver, Colorado, USA, November 29 - December 4, 1999], The MIT Press. pp. 526–532. 3

Schölkopf, B., Simard, P., Smola, A.J., Vapnik, V., 1997. Prior knowledge in support vector kernels, in: Advances in Neural Information Processing Systems 10, [NIPS Conference, Denver, Colorado, USA, 1997], The MIT Press. pp. 640–646. 3

Simard, P., LeCun, Y., Denker, J., 1993. Efficient pattern recognition using a new transformation distance, in: Advances in Neural Information Processing Systems 5, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. pp. 50–58. 3

Simard, P., LeCun, Y., Denker, J.S., Victorri, B., 1998. Transformation invariance in pattern recognition-tangent distance and tangent propagation, in: Neural Networks: Tricks of the Trade, this book is an outgrowth of a 1996 NIPS workshop, Springer-Verlag, London, UK. pp. 239–27. 4, 5, 10

Villegas, M., Paredes, R., 2011. Dimensionality reduction by minimizing nearest-neighbor classification error. Pattern Recognit. Lett. 32, 633–639. doi:10.1016/j.patrec.2010.12.002. 9, 13

Yan, S., Xu, D., Zhang, B., Zhang, H.J., Yang, Q., Lin, S., 2007. Graph embedding and extensions: A general framework for dimensionality reduction. IEEE Trans. Pattern Anal. Mach. Intell. 29, 40–51. doi:10.1109/TPAMI.2007.250598. 10

Yang, J., Frangi, A., Yang, J., Zhang, D., Jin, Z., 2005. Kpca plus lda: a complete kernel fisher discriminant framework for feature extraction and recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on 27, 230–244. 2

Yang, J., Yang, J.y., 2003. Why can LDA be performed in PCA transformed space? Pattern Recogn. 36, 563–566. doi:10.1016/S0031-3203(02)00048-1. 2

Yang, J., Zhang, L., Yang, J.y., Zhang, D., 2011. From classifiers to discriminators: A nearest neighbor rule induced discriminant analysis. Pattern Recogn. 44, 1387–1402. doi:10.1016/j.patcog.2011.01.009. 13

Zhao, D., Lin, Z., Xiao, R., Tang, X., 2007. Linear Laplacian Discrimination for Feature Extraction, in: Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on, pp. 1–7. 10, 16

Zheng, Z., Yang, F., Tan, W., Jia, J., Yang, J., 2007. Gabor feature-based face recognition using supervised locality preserving projection. Signal Processing 87, 2473–2483. doi:10.1016/j.sigpro.2007.03.006. Special Section: Total Least Squares and Errors-in-Variables Modeling. 10