Regularized Neighborhood Component Analysis

Zhirong Yang and Jorma Laaksonen

Laboratory of Computer and Information Science *
Helsinki University of Technology
P.O. Box 5400, FI-02015 TKK, Espoo, Finland
{zhirong.yang, jorma.laaksonen}@tkk.fi

Abstract. Discriminative feature extraction is one of the fundamental problems in pattern recognition and signal processing. It was recently proposed that maximizing the class prediction by neighboring samples in the transformed space is an effective objective for learning a low-dimensional linear embedding of labeled data. The associated methods, Neighborhood Component Analysis (NCA) and Relevant Component Analysis (RCA), have been proven to be useful preprocessing techniques for discriminative information visualization and classification. We point out here that NCA and RCA are prone to overfitting and therefore regularization is required. NCA and RCA's failure for high-dimensional data is demonstrated in this paper by experiments in facial image processing. We also propose to incorporate a Gaussian prior into the NCA objective and obtain the Regularized Neighborhood Component Analysis (RNCA). The empirical results show that the generalization can be significantly enhanced by using the proposed regularization method.

1 Introduction

Discriminant Analysis (DA) is one of the central problems in pattern recognition and signal processing. The supervised training data set consists of pairs $(\mathbf{x}_j, c_j)$, $j = 1, \ldots, n$, where $\mathbf{x}_j \in \mathbb{R}^m$ is the primary data, and the auxiliary data c_j takes categorical values. DA seeks for a transformation $\mathbf{f} : \mathbb{R}^m \to \mathbb{R}^r$ (usually $r \ll m$) such that $\mathbf{y}_j = \mathbf{f}(\mathbf{x}_j)$ encodes only the relevant information with respect to c_j. Here relevance or discrimination can be measured by the expectation of predictive probability $E\{p(c|\mathbf{y})\}$. Because in real applications only finite data are available, one has to estimate $p(c|\mathbf{y})$ according to a certain density model, which leads to different DA algorithms.

Fisher's *Linear Discriminant Analysis* (LDA) [3] is a classical method for this task. LDA maximizes the trace quotient of between-class scatter against within-class scatter and can be solved by *Singular Value Decomposition* (SVD). LDA is attractive for its simplicity. Nevertheless, each class in LDA is modeled by a single Gaussian distribution and all classes share a same covariance, which

* Supported by the Academy of Finland in the projects *Neural methods in information retrieval based on automatic content analysis and relevance feedback* and *Finnish Centre of Excellence in Adaptive Informatics Research*.

restricts the applicability of LDA. For example, the output dimensions of LDA cannot be more than the number of classes minus one. Furthermore, LDA is prone to overfitting because no complexity control is involved.

Recently Goldberger et al. [6] presented an algorithm that learns a Mahalanobis distance measure. Their method, called *Neighborhood Component Analysis* (NCA), maximizes the approximated discrimination $\sum_j p(c_j|\mathbf{y}_j)$. NCA is able to learn a transformation matrix such that the *Nearest-Neighbor* (NN) classifier performs well in the linear transformed space. Peltonen and Kaski [8] proposed a tightly connected technique, *Relevant Component Analysis* (RCA), where the objective is to maximize $\sum_j \log p(c_j|\mathbf{y}_j)$ and the transformation matrix is constrained to be orthonormal. Both NCA and RCA can handle complicated class distributions and output arbitrary number of dimensions. They have been applied to several data sets, where both methods seem to generalize well [6, 8].

However, we argue that these two methods are not free of overfitting in very high-dimensional spaces and complexity control is therefore required for the transformation matrix. To improve the generalization, we propose to incorporate a Gaussian prior to the NCA objective and obtain a novel method called *Regularized Neighborhood Component Analysis* (RNCA). In this paper, several empirical examples in facial image processing are presented, where the dimensionality of data is much higher than those used in [6, 8]. It turns out that both NCA and RCA behave poorly in our generalization tests, but the overfitting problems can be significantly overcome by using RNCA.

The remaining of the paper is organized as follows. We briefly review the principles of NCA and RCA in Section 2. Next we describe the motivation of regularizing the transformation matrix and present the Regularized NCA in Section 3. In Section 4 we demonstrate the experimental results, both qualitative and quantitative, on facial image processing. Finally the conclusions are drawn in Section 5.

2 Neighborhood Component Analysis

Neighborhood Component Analysis (NCA) [6] learns an $r \times m$ matrix $\mathbf{A}$ by which the primary data $\mathbf{x}_i$ are transformed into a lower-dimensional space. The objective is to maximize the *Leave-One-Out* (LOO) performance of nearest neighbor classification. NCA measures the performance based on "soft" neighbor assignments in the transformed space:

$$p_{ij} = \frac{\exp\left(-\|\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_j\|^2\right)}{\sum_{k \neq i} \exp\left(-\|\mathbf{A}\mathbf{x}_i - \mathbf{A}\mathbf{x}_k\|^2\right)} \quad , \quad p_{ii} = 0. \tag{1}$$

Denote n the number of samples and

$$p_i = \sum_{j:c_i=c_j} p_{ij}. \tag{2}$$

The objective of NCA can then be expressed as maximization of

$$\mathcal{J}_{\mathrm{NCA}}(\mathbf{A}) = \sum_{i=1}^{n} \sum_{j:c_i=c_j} p_{ij} = \sum_{i=1}^{n} p_i. \tag{3}$$

The NCA learning algorithm is based on the gradient

$$\frac{\partial \mathcal{J}_{\mathrm{NCA}}(\mathbf{A})}{\partial \mathbf{A}} = 2\mathbf{A} \sum_{i=1}^{n} \left(p_i \sum_{k=1}^{n} p_{ik} \mathbf{x}_{ik} \mathbf{x}_{ik}^T - \sum_{j:c_i=c_j} p_{ij} \mathbf{x}_{ij} \mathbf{x}_{ij}^T \right) \tag{4}$$

and employs an optimizer such as conjugate gradients.

A method closely related to NCA is the *Relevant Component Analysis* (RCA) proposed by Peltonen and Kaski [8]. RCA maximizes the sum of log-probability, i.e. the discriminative information

$$\mathcal{J}_{\mathrm{RCA}}(\mathbf{W}) = \sum_{i=1}^{n} \log \left(\sum_{j:c_i=c_j} p_{ij}^{\mathrm{RCA}} \right) = \sum_{i=1}^{n} \log p_i^{\mathrm{RCA}}. \tag{5}$$

Different from NCA, the transformation matrix $\mathbf{W}$ in RCA is restricted to be orthogonal. That is, the rows of $\mathbf{W}$ form an orthonormal basis. Due to this constraint, RCA requires an additional parameter $\beta > 0$ to control the smoothness of the "soft" assignment:

$$p_{ij}^{\mathrm{RCA}} = \frac{\exp\left(-\beta \|\mathbf{W}\mathbf{x}_i - \mathbf{W}\mathbf{x}_j\|^2\right)}{\sum_{k \neq i} \exp\left(-\beta \|\mathbf{W}\mathbf{x}_i - \mathbf{W}\mathbf{x}_k\|^2\right)} \quad , \quad p_{ii}^{\mathrm{RCA}} = 0. \tag{6}$$

The optimization of RCA is based on a stochastic gradient of (5) with respect to the parameters in Givens rotation [7] of $\mathbf{W}$.

3 Regularized NCA

Goldberger et al. claimed that they had never observed an "overtraining" effect when using NCA [6]. In the presented empirical results [6, 8], it seems that NCA and RCA generalize well on several different data sets. On the other hand, the objectives of these two methods do not involve any explicit complexity control on the matrix elements. One is therefore tempted to ask whether they really are free of overlearning.

NCA and RCA do have some implicit complexity control mechanisms on the "soft" assignments or density estimation in order to avoid overfitting. The formulation of p_{ij} leaves out the i-th training sample, which improves the generalization of NCA (although it might cause divide-by-zero errors for outliers). The β parameter in RCA also controls the smoothness of the predictive probability estimation. Smaller β usually leads to coarser estimation but better generalization. All in all, these two strategies seem to work well for low-dimensional data sets.

However, as shown in the next section by experiments on high-dimensional facial images, NCA and RCA behave poorly beyond the training sets. Actually, given a training data set, the number of parameters to be learned by a DA algorithm is r times of that in one-dimensional projection methods such as the Support Vector Machines (SVMs) [2]. That is, the DA problem is more ill-posed and regularization is hence necessary. This motivates us to impose regularization on DA algorithms.

For many parameterized machine learning algorithms, techniques of regularization [5, 1] have demonstrated great success for improving the generalization performance in ill-posed problems. The overlearning effect caused by a small training data set and a large number of parameters to be learned can be significantly reduced by attaching a penalty term behind the original objective. Viewed from Bayesian inference, regularization actually incorporates a certain prior on the parameters themselves, and the learning maximizes the posterior

$$p(\mathbf{A}|\{\mathbf{x}_j, c_j\}) \propto p(\{\mathbf{x}_j, c_j\}|\mathbf{A})p(\mathbf{A}). \tag{7}$$

We choose the decomposable Gaussian prior

$$p(\mathbf{A}) = \prod_{k=1}^{r}\prod_{l=1}^{m} p(A_{kl}) = \prod_{k=1}^{r}\prod_{l=1}^{m} \frac{\sqrt{\lambda}}{\pi} \exp(-\lambda A_{kl}^2) \tag{8}$$

for regularization because the first-order derivative of its logarithm is simple and exists throughout the parameter sapce, which facilitates gradient-based optimization. If we model

$$p(\{\mathbf{x}_j, c_j\}|\mathbf{A}) \propto \exp\left(\sum_{i=1}^{n} p_i\right) \tag{9}$$

the regularized NCA objective becomes:

$$\begin{aligned} \text{Maximize} \quad \mathcal{J}_{\text{RNCA}}(\mathbf{A}) &= \log p(\mathbf{A}|\{\mathbf{x}_j, c_j\}) && (10)\\ &= \log p(\{\mathbf{x}_j, c_j\}|\mathbf{A}) + \log p(\mathbf{A}) + \text{constant} && (11)\\ &= \sum_{i=1}^{n} p_i - \lambda\|\mathbf{A}\|_F^2 + \text{constant}. && (12) \end{aligned}$$

Here λ acts as a non-negative trade-off parameter and $\|\mathbf{A}\|_F^2$ notates the Frobenius matrix norm $\sum_{k=1}^{r}\sum_{l=1}^{m} A_{kl}^2$. For one-dimensional subspace analysis, $y = \mathbf{a}^T\mathbf{x}$, the matrix norm reduces to $\mathbf{a}^T\mathbf{a}$, i.e. the large margin regularizer used in SVMs. We call the new algorithm *Regularized Neighborhood Component Analysis* (RNCA). It is equivalent to NCA when $\lambda = 0$. Note that the Gaussian prior is not applicable to RCA optimization because the Frobenius norm of an orthonormal matrix is a constant r. The estimation of a suitable λ in (12) is a further problem that could presumably be solved by Bayesian methods. In this paper, we simply try different values for λ empirically.

There exist other priors, e.g. the Laplacian prior [11], on linear transformation parameters. Some of them are reported to produce better results for particular learning problems, but here it is difficult to obtain convenient optimization algorithms when combining these priors with the NCA or RCA objective.

4 Experiments

4.1 Data

Several empirical results of NCA and RCA have been provided in [6, 8]. However, the data used in these experiments have been low-dimensional relative to the number of training samples. The highest dimensionality is 560 in [6] and 76 in [8]. In addition, both training and testing data in their visualization and classification experiments have been selected from the same database. Here we present the learning results of NCA, RCA and RNCA on much higher-dimensional data. Furthermore, we performed the tests on a database obtained from another source, which will certainly demonstrate the generalization powers of the compared methods better.

We have used the FERET database of facial images [9] as the training data set. After face segmentation, 2,409 frontal images (poses "fa" and "fb") of 867 subjects were stored in the database for the experiments. We obtained the coordinates of the eyes from the ground truth data of FERET collection, with which we calibrated the head rotation so that all faces are upright. Afterwards, all face boxes were normalized to the size of 32×32, with fixed locations for the left eye (26,9) and the right eye (7,9). Each image were reshaped to a 1024-dimensional vector by column-wise concatenation. The testing data set we used is the UND database (collection B) [4], which contains 33,247 frontal facial images of 491 subjects. We applied the same preprocessing procedure to the UND images as to the FERET database.

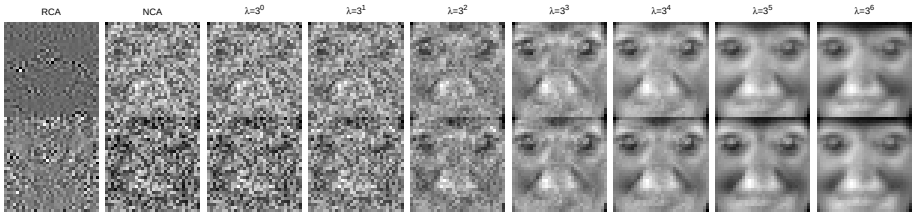
We compared the DA methods in two problems where the extracted features were used to discriminate the *gender* of a subject and whether she or he is wearing *glasses*. Table 1 shows the statistics of the classes.

Table 1. Images (subjects) of the experimented classes

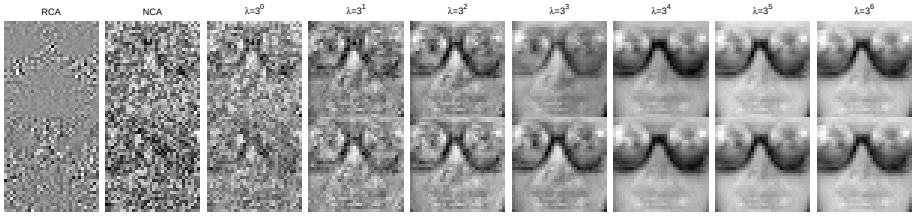
	gender		*glasses*	
	Male	Female	Yes	No
FERET	1495 (501)	914 (366)	262 (126)	2147 (834)
UND	2524 (63)	855 (19)	2601 (149)	30538 (482)

4.2 Visualizing the Transformation Matrix

It is intuitive to inspect the elements in the trained transformation matrix **A** before performing quantitative evaluation. Each row of the transformation matrix



(a) gender



(b) glasses

Fig. 1. The rows of transformation matrices are plotted as filter images.

acts as a linear filter and can be displayed like a *filter image*. If the transformation matrix works well for a given DA problem, it is expected to find some semantic connections between the filter images and our common prior knowledge of the considered classes.

Figure 1 shows the resulting filter images for the *glasses* and *gender* DA problems with a two-dimensional transformed space, i.e. $r = 2$. In both cases, RCA has stuck in a local optimum, where some pixel pairs of high contrast can be found in the plotted images. These pixel pairs look like Gabor wavelets, but they are too small to represent any semantically relevant visual patterns of gender or glasses. Such overfitting effect could be relieved by reducing the image size, as employed in [6], but much useful visual information would be lost during downsampling. The filter images trained by NCA are composed of nearly random pixels, from which it is difficult to perceive any visual patterns.

NCA is a special case of RNCA with $\lambda = 0$ in (12). Next we increased λ in the power scale of three and run RNCA with these different values of λ. The results are shown in the third to ninth columns in Figure 1. We can see that the facial patterns become clearer with higher λ values. However, too large λ's cause underfitting—all filter vectors lie in the straight line passing the two class means and thus the filter images look identical. A proper tradeoff value for λ should therefore occur in between. In the following experiments we chose to use $\lambda = 3^3$ for the *gender* case and $\lambda = 3^2$ for *glasses*. Careful readers can in these cases perceive the small differences between the displayed filter images.

4.3 Visualizing the Transformed Distributions

NCA, as well as RCA, is able to extract more than one discriminative component for two-class DA problems, which allows plotting the 2-D transformed data.

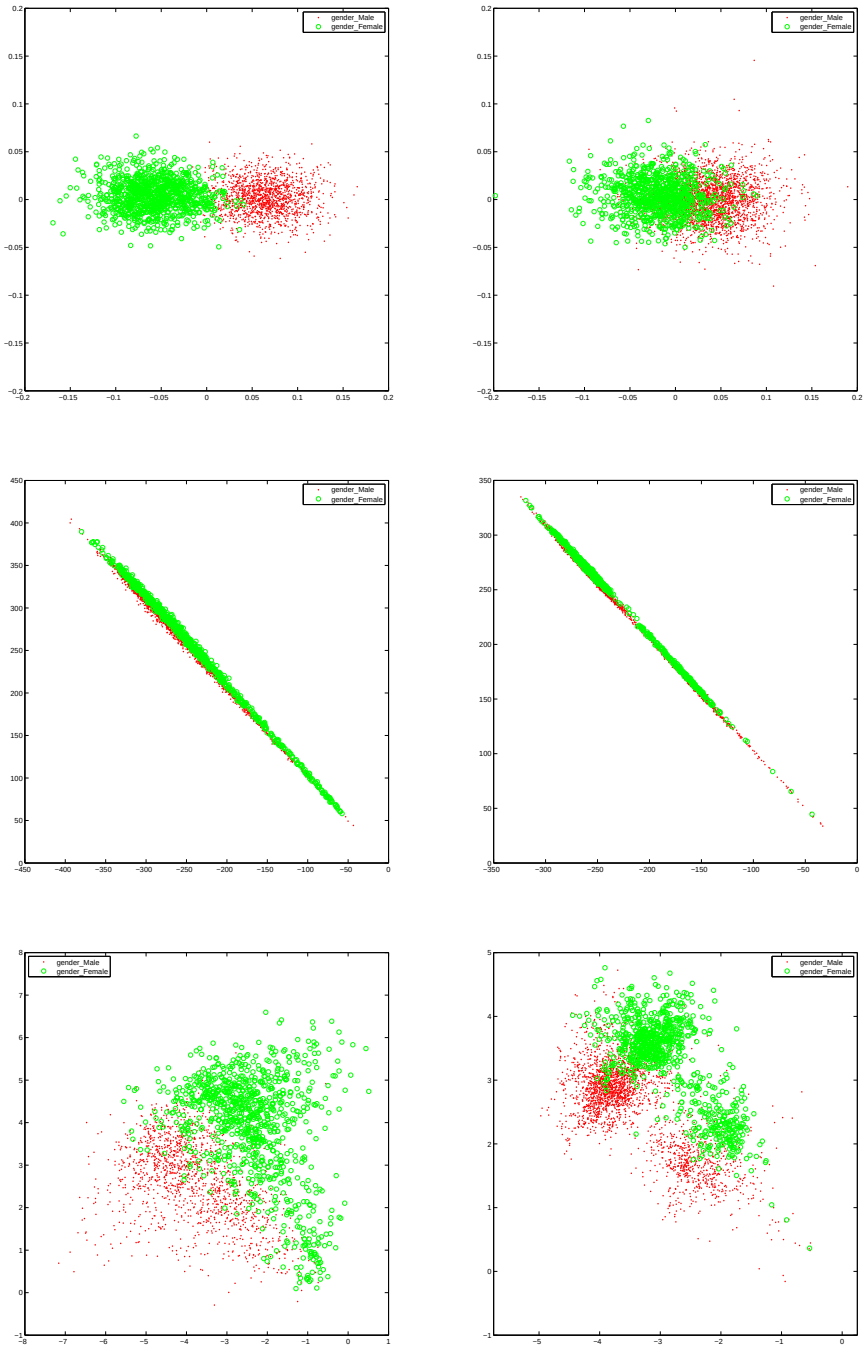


Fig. 2. 2-D transformed *gender* training data (left) and testing data (right). Rows from top to bottom are results of RCA, NCA and RNCA ($\lambda = 3^3$), respectively.

Table 2. Numbers of classification errors (false positives, false negatives) using $r = 2$ components.

	gender		*glasses*	
	training	testing	training	testing
RCA	46, 49	380, 376	2, 2	2140, 2101
NCA	84, 94	249, 248	15, 1	2113, 2141
RNCA	124, 121	201, 211	64, 67	2002, 2051

RNCA inherits the same property from NCA and can hence also be used for visualization. In this section we illustrate a qualitative comparison of NCA (3), RCA (5) and RNCA (12). Due to the space limit, only the plots of *gender* are displayed. Similar results can be obtained for the *glasses* case.

The training and testing results are shown in Figure 2. RCA starts from an orthonormalized LDA result and tries to improve it [8], but in this case the initial solution already separates data well and hence RCA does only little change. On the other hand, it can be seen in the right column that the *gender* classes are heavily mixed for the testing data.

NCA learns a transformed space in which both training data and testing data mainly distribute around a straight line, representing the *boundary direction*. The two classes are slightly separated in a *discriminative direction* nearly orthogonal to the boundary for the training data. However, such discrimination can barely be seen from the transformed testing data, where the two classes are heavily overlapped. Moreover, the 2-D Euclidean metric would not perform properly in the transformed space because of the presence of a dominant direction.

By contrast, one can easily see the almost separated classes in both training and testing cases with RNCA. Although the separation is not as clear as that of RCA for the training data, it is relatively better preserved for the testing data. That is, the overfitting effect is much alleviated by employing our regularization technique. The neighborhood based on the 2-D Euclidean metric should be more suitable for the RNCA results because the scales in the boundary and discriminative directions have become comparable.

4.4 Classification results

One of the most important applications of discriminative features is for classification. The classification results therefore serve as a quantitative comparison measure of different DA methods. Following the conventional terms in binary classification, we specify the prediction of *gender_Male* and *glasses_Yes* as positive and their counterparts as negative.

Table 2 shows the Nearest-Neighbor classification error counts when using the DA results with $r = 2$. A pair of numbers are shown in each table entry, the first for false positives and the second for false negatives. Although RNCA is not as good as RCA and NCA in classifying the training data, it performs best for the testing data. This conforms to the qualitative results demonstrated in the previous section.

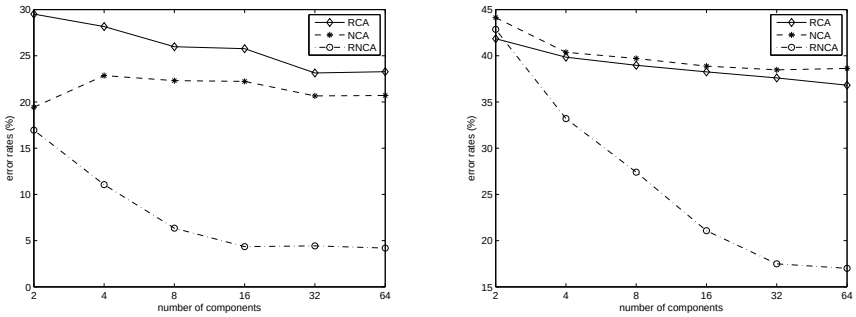


Fig. 3. Nearest-Neighbor classification error rates of the compared methods for *gender* (left) and *glasses* (right) with different numbers of r components.

The classification accuracy of RNCA can be further improved by increasing the number of components. Figure 3 illustrates the classification error rates on the testing data with different values of r. Here we have taken the average of the false positive and false negative error rates. RNCA achieves its best performance for classifying *gender* when $r \geq 16$. More components for the *glasses* case are required because there exist more different eyeglasses styles in the UND database than those in FERET [4, 9]. By contrast, RCA and NCA benefit only little from the additional components. Although high dimensionality of the transformed subspace brings more expressive power, RCA and NCA suffer from severe overfitting without regularization on the additional parameters.

The nearest neighbor classifier based on the RNCA results can even outperform the well-known Support Vector Machines (SVMs) [2]. The best classification accuracies we obtained with RNCA+NN are 95.3% for *gender* and 82.5% for *glasses*, while SVMs with linear kernel achieve only 90.8% and 78.2%, respectively. All the DA methods discussed in this paper, as well as SVM, can be generalized to non-linear cases by adopting the kernel trick [10]. However, more efforts are required to tune the additional parameter involved in the kernel, which is beyond the scope of this paper.

5 Conclusions

Two existing discriminant analysis methods, NCA and RCA, have gained success with low-dimensional data. In this paper we have pointed out that they are prone to overfitting with high-dimensional facial image data. We also proposed regularizing the neighborhood component analysis by imposing a Gaussian prior on the transformation matrix. Experimental results confirm our statement and show that the Regularized NCA becomes more robust in extracting discriminative features.

Moreover, we demonstrated that more than one component exists for two-class discriminant analysis problems. Unlike SVM and other algorithms dedi-

cated for classification, our RNCA method can be applied to many other applications, for instance, preprocessing of discriminative feature visualization and creation of Discriminative Self-Organizing Maps [12].

Similar to other linear subspace methods, RNCA is readily extendable to non-linear versions. The nonlinear discriminative components can be obtained by mapping the primary data to a higher-dimensional space with appropriate kernels. This will be a topic of our future work.

References

1. Zhe Chen and Simon Haykin. On different facets of regularization theory. *Neural Computation*, 14:2791–2846, 2002.
2. N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, 2000.
3. R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7, 1963.
4. P. J. Flynn, K. W. Bowyer, and P. J. Phillips. Assessment of time dependency in face recognition: An initial study. *Audio- and Video-Based Biometric Person Authentication*, pages 44–51, 2003.
5. Federico Girosi, Michael Jones, and Tomaso Poggio. Regularization theory and neural networks architectures. *Neural Computation*, 7(2):219–269, 1995.
6. J. Goldberger, S. T. Roweis, G. E. Hinton, and R. Salakhutdinov. Neighbourhood components analysis. In *NIPS 2004*, 2004.
7. Gene H. Golub and Charles F. van Loan. *Matrix Computations*. The Johns Hopkins University Press, 2 edition, 1989.
8. J. Peltonen and S. Kaski. Discriminative components of data. *IEEE Transactions on Neural Networks*, 16(1):68–83, 2005.
9. P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The FERET evaluation methodology for face recognition algorithms. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22:1090–1104, October 2000.
10. B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.
11. R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, 58(1):267–288, 1996.
12. Zhirong Yang and Jorma Laaksonen. Partial relevance in interactive facial image retrieval. In *Proceedings of 3rd International Conference on Advances in Pattern Recognition (ICAPR 2005)*, pages 216–225, Bath, UK, August 2005.