What would you look like in Springfield? Linear Transformations between High-Dimensional Spaces

Ribana Roscher, Falko Schindler, Wolfgang Förstner

TR-IGG-P-2011-01

24. August 2011

Technical Report Nr. 1, 2011

Department of Photogrammetry
Institute of Geodesy and Geoinformation
University of Bonn

Available at
http://www.ipb.uni-bonn.de/papers/

What would you look like in Springfield? Linear Transformations between High-Dimensional Spaces

Ribana Roscher, Falko Schindler, Wolfgang Förstner

Zusammenfassung

High-dimensional data structures occur in many fields of computer vision and machine learning. Transformation between two high-dimensional spaces usually involves the determination of a large amount of parameters and requires much labeled data to be given. There is much interest in reducing dimensionality if a lower-dimensional structure is underlying the data points. We present a procedure to enable the determination of a low-dimensional, projective transformation between two data sets, making use of state-of-the-art dimensional reduction algorithms. We evaluate multiple algorithms during several experiments with different objectives. We demonstrate the use of this procedure for applications like classification and assignments between two given data sets. Our procedure is semi-supervised due to the fact that all labeled and unlabeled points are used for the dimensionality reduction, but only few them have to be labeled. Using test data we evaluate the quantitative and qualitative performance of different algorithms with respect to the classification and assignment task. We show that with these algorithms and our transformation approach high-dimensional data sets can be related to each other. Finally we can use this procedure to match real world facial images with cartoon images from Springfield, home town of the famous Simpsons.

1 Introduction

Many classification, recognition and assignment applications deal we large amounts of high-dimensional data as in the classification of images from databases, in the analysis of gene expression microarrays or time series. The data analysis can fail when no or only a few user annotation is given and the dimensionality is very large.

Dimensionality reduction, also called subspace learning, has emerged as a powerful technique in pattern recognition *e.g* for reducing computational effort and increasing classification performance. Applications of dimension

reductive methods are widely spread in the field of computer vision, namely concerning problems of detection, tracking, recognition, segmentation and reconstruction [1, 14, 15, 19]. Our goal is to provide a semi-supervised assignment strategy between data points living in two different high-dimensional spaces.

The problem of dimensionality reduction can be stated as follows: Given a set of N data points $\{\boldsymbol{\phi}_n\}$, $n = 1, \ldots, N$ in $\mathbb{R}^D$, we want to find lower dimensional data points $\{\boldsymbol{x}_n\}$ in $\mathbb{R}^d$ with $d \ll D$ so that $\boldsymbol{x}_n$ is an appropriate representation of $\boldsymbol{\phi}_n$ (compare one half of Figure 1). The data points $\{\boldsymbol{\phi}_n\}$ are collected in an $(d \times N)$-dimensional matrix $\boldsymbol{\Phi}$ and $\{\boldsymbol{x}_n\}$ in X, respectively.

Having points $\{\boldsymbol{x}_n\}$, $\{\boldsymbol{y}_n\}$ from two similar data sets, a unique projective transformation H from one space into another is to be determined. Therefore, a small amount of labeled points N_l can be used. Figure 1 shows these points and how both data sets can be related via the low-dimensional space.

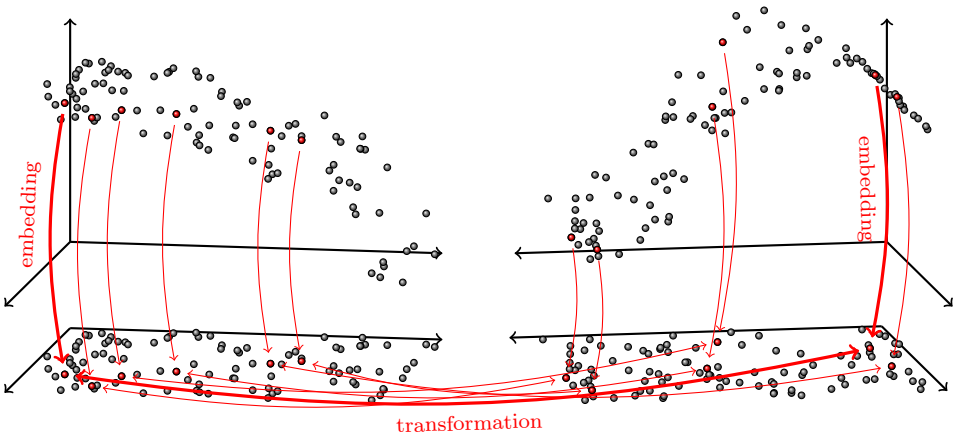


Abbildung 1: We apply dimensionality reduction on two similar data sets living in two different high-dimensional spaces, exploiting the structure of all available data. We estimate a unique, projective transformation between both reduces spaces using only few labeled data points. This way a new test point from one data set can be assigned to its nearest neighbor within the other set without labeling all data but in a semi-supervised manner.

We investigate and evaluate seven popular, linear and graph-based methods for dimensionality reduction in an extensive test framework. Since we assume to have only few labeled without geometric information, yet a large number of unlabeled data, we consider only unsupervised dimensionality reduction algorithms.

We relate both reduced data sets to each other with a linear, $(d + 1)$-dimensional transformation matrix. This projective matrix can be determined in case at least $d + 2$ common data points are available.

Using the dimensional reduction followed by a projective transformation we can relate every new test point to its nearest neighbor within the other data set using Euclidean distances, without the need of a complete set of labels and a full D-dimensional transformation matrix. The quality of these assignments is investigated depending on the dimensionality reduction algorithm and several parameters.

Our contribution is to show that using such algorithms and the presented transformation approach we can relate high-dimensional data sets to each other with a minimal amount of labeled data. This can be used for the assignment of data points from different spaces as well as for the classification of images.

In Section 2 we give an overview of the related work with special focus on spectral methods. Section 3 briefly illustrates representative linear and graph-based state-of-the-art spectral methods, which we analyze during our experiments. In Section 4 we explain how to relate two data sets using a projective transformation. Section 5 compares seven spectral methods in order to find preferably robust and low-dimensional structures in a semi-supervised manner. We evaluate the algorithms on gray-scale and colored images from the Aberdeen database, real photos from the FERET database and Simpsons cartoon images. In the last section we discuss our results and provide an outlook to future work.

2 Related work

The field of computer vision offers numerous feature selection algorithms and extraction methods. Spectral methods constitute a group specialized in reducing the dimensionality of data. We distinguish between two major types of algorithms:

1. Linear methods including Principal Component Analysis (PCA) [12], and Multidimensional Scaling (MDS) [6] and

2. Nonlinear methods including graph-based methods, *e.g* [3, 16, 21, 25, 27], and kernel methods, *e.g* [10, 23]. In contrast to linear methods the nonlinear methods perform better on complex nonlinear data structures as they occur in real world data sets.

In the last years further algorithms closely related to the latter have been developed focusing on acceleration, *e.g* [7, 28], and qualitative improvement [4, 8, 29].

Typical data sets used in the aforementioned methods are pictures of an object subject to changing illumination, angle, translation or other varying

characteristics. The data points defined by the vectorized pixel intensities of each image vary smoothly so that they define a manifold lying within a high-dimensional space.

For example, Roweis and Saul [22] demonstrate the Isomap algorithm employing it on face images with varying illumination conditions and angles, on hand images with natural hand movements and also on handwritten digits.

These originally unsupervised algorithms for dimensionality reduction can also be used in a semi-supervised manner if additional information exists, *e.g* must- and cannot-links between some data points. Several such methods have been proposed, including similarity adapting approaches, *e.g* [5, 9], or search based methods with user-provided constraints or labels, *e.g* [2, 26]. Given additional data these methods yield more reliable results than completely unsupervised approaches and even allow for user interaction.

We apply unsupervised dimensionality reducing algorithms on both labeled and unlabeled data and use labeled points to estimate the projective transformation between two low dimensional spaces. In this sense our approach can be seen as being semi-supervised.

3 Spectral Methods for Dimensionality Reduction

Spectral methods are a class of techniques used for dimensionality reduction. The reduction is done by detecting a low-dimensional structure in a higher-dimensional space by decomposing a specially constructed matrix, which is mostly a weighted graph of the initial data. Spectral methods are convex and therefore optimize an objective function globally.

In contrast to manifold learning, where some representation for the underlying manifold $f : f(\boldsymbol{\phi}) = 0$ is estimated, dimensionality reduction only considers the estimation of lower-dimensional data points $\{\boldsymbol{x}_n\}$ from the input data points $\{\boldsymbol{\phi}_n\}$. Consequently a transformation back into high-dimensional space is non-trivial and beyond the scope of our work. We stay with the output points and assign nearest neighbors after transforming one set of points into the other space.

3.1 Linear Methods

Generally, linear methods retrieve a structure of the lower dimensional data points $\{\boldsymbol{x}_n\}$ lying close to a linear affine *subspace* of the high-dimensional space. The methods yield d-dimensional data points $\boldsymbol{x}_n$, which are a linear

combination of the original D-dimensional data points $\boldsymbol{\phi}_n$:

$$\boldsymbol{x}_n = \boldsymbol{r}_1^* \boldsymbol{\phi}_{n,1} + \boldsymbol{r}_2^* \boldsymbol{\phi}_{n,2} + \ldots + \boldsymbol{r}_M^* \boldsymbol{\phi}_{n,M} = R^* \boldsymbol{\phi}_n \tag{1}$$

with R^* being the $(d \times D)$-dimensional matrix for the linear transformation with the star indicating the reduced dimensionality in contrast to R being a square matrix. For the combined point matrices we obtain $X = R^* \boldsymbol{\Phi}$.

We consider two state-of-the-art subspace methods: Principal Component Analysis (PCA) [12], and Metric Multidimensional Scaling (MDS) [6]. Since many years they are used widely in the field of pattern recognition.

PCA reduces the dimensionality by preserving the global covariance structure of all data points. We can compute the lower dimensional data points with (1) by mapping them onto the M basis vectors $\boldsymbol{r}$ with the largest eigenvalues s: $R^* = [\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_M]^\mathsf{T}$. The latter are derived from the eigen decomposition of the covariance matrix $\Sigma_{\boldsymbol{\phi},\boldsymbol{\phi}} = RS^2R^\mathsf{T}$.

MDS reduces the dimensionality by preserving the inner products between the data points by decomposing the Gram matrix $K : K_{nm} = \boldsymbol{\phi}_n \cdot \boldsymbol{\phi}_m$, having the same eigenvalues as the covariance matrix of the PCA up to a constant in the classical setup. Therefore, the output of classical MDS is identical to that of the PCA. Modern MDS algorithms use iterative methods, so that the points are better arranged.

The main drawback of both methods is that they retain large distances and do not consider the local distribution of the neighborhood around data points. Therefore important structures can be lost as can be seen in the Swiss roll data set [25].

3.2 Graph-based Methods

If the structure underlying the data is not affine, linear methods can fail. Graph-based methods can find this structure even if the data is lying within or close to a low dimensional manifold. The key aspect of these algorithms is to preserve local topological and geometrical properties.

These methods can by divided into three parts:

1. Construct a graph $\mathcal{G}$ with nodes representing the data points $\boldsymbol{\Phi}$ and edges defining relations between them. Each node is connected to all data points within a local ϵ-neighborhood or to its k-nearest neighbors.

2. A matrix W is derived from the graph $\mathcal{G}$ by choosing weights, *e.g* $w_{nm} = 0$ if there is no connection between points n and m and $w_{nm} = 1$ or some distance measure $w_{nm} = d(\boldsymbol{\phi}_n, \boldsymbol{\phi}_m)$ if there is one.

3. In the last step a matrix including the weights W is decomposed. The way of how to use W mainly makes up the difference between the algorithms.

We consider five representative state-of-the-art graph-based methods: Isometric Mapping (Isomap) [25], Locally Linear Embedding (LLE) [21, 22], Laplacian Eigenmaps [3], Diffusion Maps [16] and Local Tangent Space Alignment (LTSA) [29].

Isomap preserves pairwise distances between data points $\{\boldsymbol{\phi}_n\}$ along the estimated manifold $f : f(\boldsymbol{\phi}) = 0$. In principle the Isomap algorithm equals MDS, whereby the Euclidean distances are replaced by geodesic distances along the manifold. Due to the different distance measure Isomap is superior to linear methods in case of complex and non-linear structures. Although, Isomap may suffer from holes within the data structure and so called short-circuiting, *e.g* misleading connections to topologically separated points.

LLE preserves local linear structure of nearby data points. It builds up a weighted graph from k-nearest neighbors. After decomposing the matrix

$$M = (I - W)^{\mathsf{T}}(I - W) \tag{2}$$

with I being the identity matrix the largest eigenvector is discarded and the remaining ones yield the lower-dimensional data points. The neighborhood of every point is assumed to be planar. Despite of its good performance in a wide variety of applications LLE tends to cluster dense regions of the data and can hardly handle holes.

Laplacian Eigenmaps as well as Diffusion Maps preserve so called proximity relations: Nearby input data points $\{\boldsymbol{\phi}_n\}$ are projected to nearby output data points $\{\boldsymbol{x}_n\}$. They minimize the gradient norm in a least squares sense by decomposing the matrix

$$L = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}. \tag{3}$$

Using the $d + 1$ largest eigenvectors of the matrix yield the M-dimensional data points, whereby the largest eigenvector is discarded. The diagonal matrix D has elements $D_{nn} = \sum_m W_{nm}$. The matrix L is also called Graph Laplacian.

For Diffusion Maps the weight matrix W is constructed using the diffusion kernel:

$$W_{mn} = \exp(-|\boldsymbol{\phi}_n - \boldsymbol{\phi}_m|^2/\sigma^2) \tag{4}$$

with σ being a scale parameter, affecting the number of points having high weights in the graph. In contrast to all other graph-based methods the Diffusion Maps do not provide a parameter k defining the number of neighboring points. Another parameter is α controlling the graph normalization. Note that $\alpha = 0$ leads to L being the Graph Laplacian and $\alpha = 1$ yields the Laplace-Beltrami operator. We choose the latter for our experiments with a fixed scale parameter $\sigma = 10$.

Both algorithms suffer from similar drawbacks like LLE.

LTSA preserves the geometry within the tangent space at each data point. The method computes a graph from the k-nearest neighbors and approximates the local tangent space of each neighborhood. The local tangent space is aligned and embedded in a global coordinate system. As we can conclude from our experiments described in Section 5 LTSA as well as LLE cause high computational costs due to there complexity.

4 Transformations between Different Subspaces

Given two sets of data points written with homogeneous coordinate vectors $\mathbf{x}_n^\mathsf{T} = [\boldsymbol{x}_n^\mathsf{T}, 1], \mathbf{y}_n^\mathsf{T} = [\boldsymbol{y}_n^\mathsf{T}, 1]$ a linear, projective transformation $\mathsf{H} : \mathbf{y}_n = \mathsf{H}\mathbf{x}_n$ is to be determined. In terms of the combined coordinate matrices X and Y the homogeneous representation is $\mathsf{Y} = \mathsf{H}\mathsf{X}$.

Since the $(d{+}1)$-dimensional matrix H is homogeneous there are $(d{+}1)^2{-}1$ parameters to be determined. Each labeled pair of data points eliminates d parameters being the number of its Euklidean coordinates. Therefore, we need at least $N_l = ((d{+}1)^2{-}1)/(d{+}1) = d{+}2$ labeled points for determining a unique transformation H. We can multiply with the pseudo inverse of Y and obtain $\mathsf{H} = \mathsf{Y}\mathsf{X}^+$.

Figure 2 illustrates how the transformation can be determined in low-dimensional space with only $d + 2$ links.

5 Experiments

As depicted in Figure 3 we use several pairwise similar data sets in out experiments, including

Aberdeen contains 695 pairs of facial images from the Psychological Image Collection at Stirling[1], each pair containing a colored and a gray scale

[1] `http://pics.psych.stir.ac.uk/`

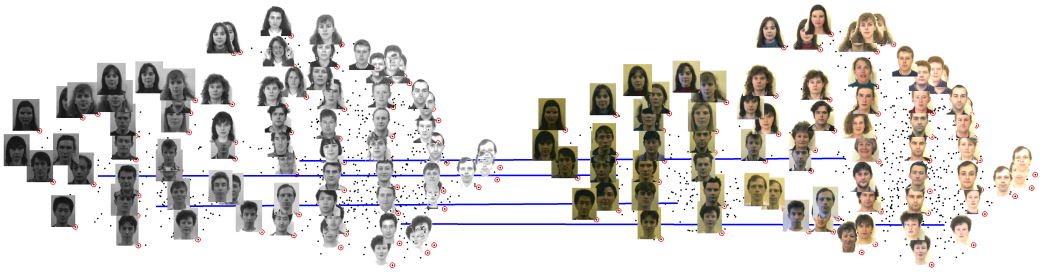


Abbildung 2: Assignment of a gray-scale and a colored version of images from the Aberdeen database. Both image spaces have been reduced to only two dimensions. The straight lines show how both low-dimensional spaces can be related with only $N_l = d + 2 = 4$ labeled points.

version. Goal of this first experiment is to assign one similar colored image to given gray-scale images.

Digits are 1900 handwritten digits [24] and 2940 digital ones. We use dimensionality reduction methods and the estimated transformation to classify handwritten images by finding corresponding digital versions.

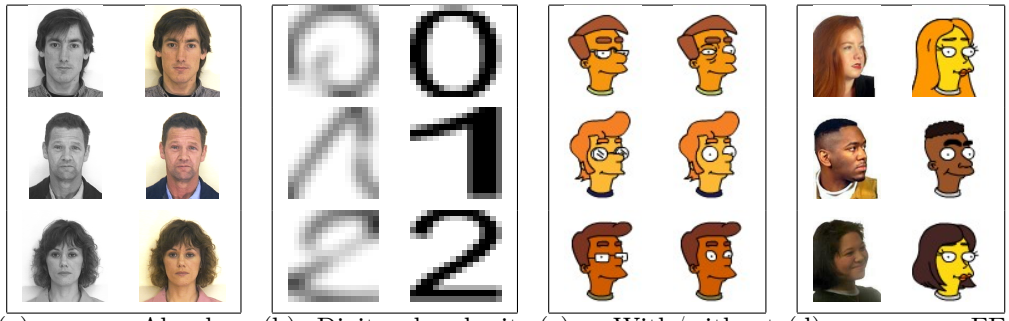
Glasses contain 1626 semi-automatically created Simpsons avatars[2], each pair containing one face with and without glasses. By searching for the nearest neighbor in the space of non-glass images we remove glasses from these cartoon faces.

FERET/Simpsons consists of 1577 images from the Color FERET face database [17, 18] and 789 selected Simpsons avatars. We "cartoonize" human faces from FERET by matching them with matching characters from Springfield, home town of the famous TV family, the Simpsons.

We reduce the dimensionality of both data sets to the same target dimension and compute the transformation between both subsets. The dimensionality reduction is done using all $N_l + N_t + N_u$ labeled, test, and unlabeled images, while the transformation estimation relies on N_l labeled data points only.

After determining the transformation between both low-dimensional spaces the test images from one subspace are transformed into the other subspace. There they are assigned to their nearest neighbors to yield an approximate relative position to other images within the high-dimensional space. The labels, *i.e* the unique ID, of each paired test images are compared to compute an accuracy measure between 0 and 100 %. In case of a classification task

[2] `http://www.simpsonsmovie.com/`



(a) Aberdeen gray/color (b) Digits handwritten/digital (c) With/without glasses (d) FERET/Simpsons

Abbildung 3: Example images for all data sets used in this paper. We relate grayscale images with their colored versions, recognize handwritten digits by mapping them on digitally generated counterparts, remove glasses from Simpsons avatars and human faces with a similar carton drawing.

like the classification of handwritten digits only the classes of a test image pair are compared, since the specific image ID is not relevant.

We compare seven state-of-the art subspace methods: PCA, MDS, Isomap, LLE, Laplacian Eigenmaps, Diffusion Maps and LTSA. We analyze the influence of changing parameters like the number of labeled points, the number of used neighbors k and the target dimension. The implementations of all graph-based algorithms are kindly provided by the authors. For the PCA we use a fast implementation from Mark Tygert[3] and the MDS implementation is an iterative version written by Michael Lee [13].

5.1 Assignment Accuracy of Cartoon Faces with and without Glasses

In our first experiment we choose two data sets, which equal in image content and feature dimension. We create 1626 pairs of randomly assembled Simpsons avatars with and without glasses. We run all experiments with an image size of 60×40 pixels, different target dimensions $d = \{2, \ldots, 80\}$, different neighborhood sizes $k = \{5, \ldots, 1600\}$ and a varying number of labeled points N_l.

Figure 4 shows that the accuracy for all methods but Laplacian Eigenmaps and Diffusion Maps significantly increases with the number of labeled data points. PCA, Isomap and MDS show the best results, whereby PCA

[3] `http://www.mathworks.de/matlabcentral/fileexchange/21524-principal-component-analysis`

improves mostly with increasing number of labeled points and Isomap outperforms all other methods with using the minimum number of labeled points $N_l = d + 2$. LTSA also improves significantly. Isomap yields better results at a high number of neighbors k. We observe that the accuracy increases with increasing number of neighbors. LTSA and PCA outperform all other methods with $N_l = 100$.

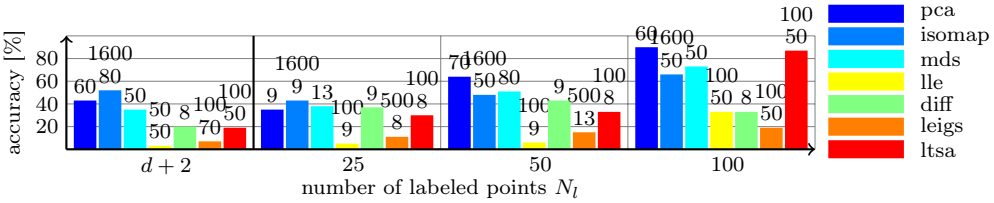


Abbildung 4: Number of labeled points N_l versus assignment accuracy of cartoon faces with and without glasses. For each algorithm the lowest error rate out of different neighborhood sizes k and dimensionalities d is plotted. The left part of the plot shows results with N_l depending on the dimensionality, while on the right side N_l is fixed for varying d. Above each bar there is the number of neighbors k (top) and the dimensionality d (below).

5.2 Assignment Accuracy of Gray-scale Images to Colored Images

In this experiment we compare the assignment accuracy of the algorithms for equal images: one set is colored, the other one is gray-scaled. Note that the dimensionality of colored images is three time the dimensionality of gray-scale images. We run the experiments with an image size of 60×40 pixels, different target dimensions $d = \{2, \ldots, 50\}$, different neighborhood sizes $k = \{5, \ldots, 656\}$ and several numbers of labeled points N_l.

Figure 5 shows that all algorithms increase the assignment accuracy with an increasing number of labeled points N_l. PCA, Isomap, MDS and LLE and LTSA perform best, whereby Isomap shows best results with a minimum number of labeled points $N_l = d + 2$. Both Diffusion Maps as well as the closely related Laplacian Eigenmaps show low assignment rates. Isomap, LLE and LTSA tend to achieve better results with a high number of neighbors k. The plot also shows that the dimension for best results seems to be $d = 8$.

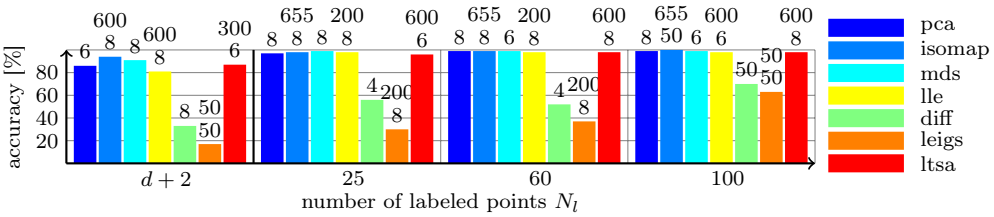


Abbildung 5: Results for the colored and gray-scale Aberdeen faces in analogy to Figure 4.

5.3 Classification Accuracy of Handwritten and Digital Digits

In this experiment we use 1900 handwritten digits [24] and 2940 digitally generated digits. Both data sets include gray-valued images of size 16×16. We use our transformation procedure to solve a classification task. Some of the handwritten and all of the digital digits are labeled with the number shown in the image. The idea is that images with the equal labels lie within similar regions in both subspaces. Given test images of handwritten digits can be assigned to labeled digital digits and classified.

Figure 6 shows that the classification accuracy increases rapidly with an increasing number of labeled points N_l. The Laplacian Eigenmaps outperform all other algorithms in nearly all cases. Again, Isomap needs much more neighbors than other algorithms. It can be seen that at a certain number of labeled points N_l the classification accuracy does not improve significantly anymore. In this experiment, LTSA and LLE show long running times. We quit there calculations with more than 100 neighbors after several hours. As Figure 7 illustrates most erroneous classifications arise from similarities between specific numbers, *e.g* 3, 8, 0.

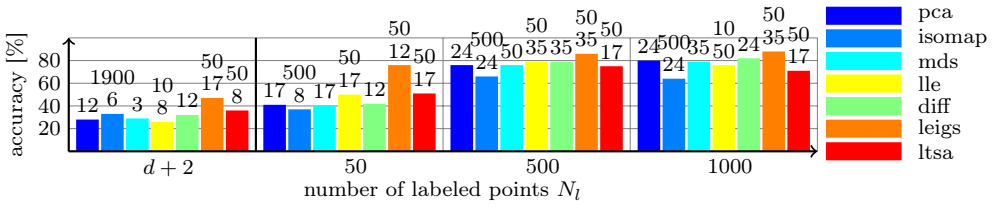
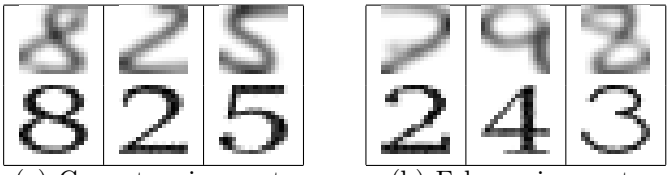


Abbildung 6: Results for the handwritten and digital digits in analogy to Figure 4.



(a) Correct assignments (b) False assignments

Abbildung 7: Example results for the assignment of handwritten and digital digits.

5.4 Assignment of Human Faces to Cartoon Avatars

The Color FERET database include colored images from many different persons, angles and illuminations. We only consider images with half right and mirrored images with half left angle to achieve a uniform view for all images. Furthermore, we aligned the resulting 1577 images with a free software from Huang *et al.* [11], set all background pixels to white, brightened up some dark images and adjusted their contrast. Despite this preprocessing step several images still show differences, even images of the same persons.

With the help of an online web-application we semi-automatically created and downloaded 789 selected Simpsons avatars with different race, hair, noses, eyes, mouths and glasses.

As in the other experiments we choose several target dimensions d, number of neighbors k and labeled points N_l. We manually labeled 80 images and assign one Simpsons avatar to every FERET test image.

Some examples from Isomap and Laplacian Eigenmaps can be seen in Figure 8. The qualitative results visually show good performance regarding to sex, hair color and skin.

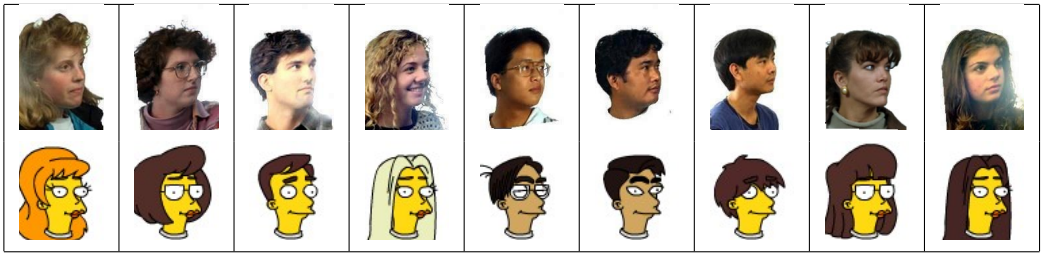


Abbildung 8: Example results for the assignment of FERET images and Simpsons avatars.

6 Discussion and Outlook

We propose a method to determine a transformation between similar high-dimensional data sets by reducing the dimensionality and estimate the transformation between two subspaces using a few labeled points. The dimensionality reduction is done with both labeled and unlabeled data such that our approach can be seen as semi-supervised.

We apply our approach to the "cartoonizationöf the FERET database and yield promising results. We could improve the results by choosing more consisting images with more distinctive labels. As can be seen from Figure 2 the location of one image within low-dimensional space clearly depends on image characteristics like background color and contrast, *i.e* properties being of no interest for the assignment. This application can be extended to the cartoonization of unseen images by aligning them and removing the background automatically with some segmentation algorithm, *e.g* GrabCut [20].

We show, that linear methods mostly perform as well as nonlinear methods if there are enough labeled data points. The nonlinear methods suffer from the aspect that unknown parameters like the number of neighbors have to be chosen carefully to achieve good results. However, if the parameters are chosen well, the nonlinear methods outperform the linear ones with a minimum number of labeled points $N_l = d + 2$. We show that in our application LLE and especially Isomap tend to achieve better results with a high number of neighbors $k \to N_l$. For the digit data set we observed LTSA and LLE not being practicable if the number of points N is high.

In future work the presented application can be extended to more general data sets to yield more robust assignment. The proposed framework can be applied to further fields of machine learning dealing with high-dimensional data.

Literatur

[1] F. R. Bach and M. I. Jordan. *Spectral Clustering for Speech Separation*, page 221. Wiley, 1 edition, 2009.

[2] S. Basu, A. Banerjee, and R. Mooney. Semi-Supervised Clustering by Seeding. In C. Sammut and A. G. Hoffmann, editors, *The Nineteenth International Conference on Machine Learning*, pages 19–26, Sydney, NSW, Australia, 2002. Morgan Kaufmann Publishers Inc.

[3] M. Belkin and P. Niyogi. Laplacian Eigenmaps for Dimensionality Re-

duction and Data Representation. *Neural Computation*, 15(6):1373–1396, 2003.

[4] H. Chang and D. Y. Yeung. Robust Locally Linear Embedding. *Pattern Recognition*, 39(6):1053–1065, 2006.

[5] O Chapelle and A Zien. Semi-Supervised Classification by Low Density Separation. In R. G. Cowell and Z. Ghahramani, editors, *Proceedings of the Tenth International Workshop on Artificial Intelligence and statistics*, volume 2005, Barbados, 2005. The Society for Artificial Intelligence and Statistics.

[6] T. F. Cox and M. A. Cox. Multidimensional Scaling. *Chapman & Hall*, 30, 1994.

[7] V. De Silva and J. B Tenenbaum. Global versus Local Methods in Nonlinear Dimensionality Reduction. In *Advances in Neural Information Processing Systems 15*, pages 721–728. MIT Press, 2003.

[8] D. L. Donoho and C. Grimes. Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data. In *Proceedings of the National Academy of Sciences*, volume 100, pages 5591–5596. National Acad Sciences, 2003.

[9] B. Fischer, V. Roth, and J. M. Buhmann. Clustering With the Connectivity Kernel. In *Advances in Neural Information Processing Systems 16: Proceedings of the 2003 Conference*, page 89. The MIT Press, 2004.

[10] J. Ham, D. D. Lee, S. Mika, and B. Schölkopf. A Kernel View of the Dimensionality Reduction of Manifolds. In *Proceedings of the Twenty-First International Conference on Machine Learning*, page 47. ACM, 2004.

[11] G.B. Huang, V. Jain, and E. Learned-Miller. Unsupervised Joint Alignment of Complex Images. In *IEEE International Conference on Computer Vision*, pages 1–8. Citeseer, October 2007. ISBN 978-1-4244-1630-1. doi: 10.1109/ICCV.2007.4408858.

[12] I. T. Jolliffe. *Principal Component Analysis*. Springer, 2nd edition, 2002.

[13] M. Lee. Algorithms for Representing Similarity Data, 1999.

[14] A Mittal, A. Monnet, and N. Paragios. Scene Modeling and Change Detection in Dynamic Scenes: A Subspace Approach. *Computer Vision and Image Understanding*, 113(1):63–79, January 2009. ISSN 10773142.

[15] H. Murase. Moving Object Recognition in Eigenspace Representation: Gait Analysis and Lip Reading. *Pattern Recognition Letters*, 17(2):155–162, February 1996. ISSN 01678655. doi: 10.1016/0167-8655(95)00109-3.

[16] B. Nadler, S. Lafon, and R. R. Coifman. Diffusion Maps, Spectral Clustering and Reaction Coordinates of Dynamical Systems. *Applied and Computational Harmonic Analysis*, 21(1):113–127, 2006.

[17] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The FERET Evaluation Methodology for Face Recognition Algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000.

[18] P.J. Phillips, H. Wechsler, J. Huang, and P.J. Rauss. The FERET Database and Evaluation Procedure for Face Recognition Algorithms. *Image and Vision Computing*, 16(5):295–306, 1998.

[19] S.R. Rao, R. Tron, R. Vidal, and Y. Ma. Motion segmentation via robust subspace separation in the presence of outlying, incomplete, or corrupted trajectories. In *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008*, volume 37, page 1â8, June 2008. doi: 10.1109/TPAMI.2005.123.

[20] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive Foreground Extraction Using Iterated Graph Cuts. *ACM SIGGRAPH 2004 Papers*, 2004.

[21] S. T. Roweis and L. K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 2000.

[22] L. K. Saul and S. T. Roweis. Think Globally, Fit Locally: Unsupervised Learning of Low Dimensional mManifolds. *Journal of Machine Learning Research*, 4:119–155, 2003.

[23] B. Schölkopf, A. Smola, and K.R. Müller. *Kernel Principal Component Analysis*. MIT Press, Cambridge, MA, 1999.

[24] A. K. Seewald. DigitsâA dataset for Handwritten Digit Recognition TR-2005-27, 2005.

[25] J. B. Tenenbaum, V. Silva, and J. C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500): 2319, 2000.

[26] K. Wagstaff and C. Cardie. Clustering with Instance-Level Constraints. In *Proceedings of Seventeenth National Conference on Artificial Intelligence*, pages 1103–1110, Austin, Texas, 2000. AAAI Press/The MIT Press.

[27] K. Q. Weinberger and L. K. Saul. Unsupervised Learning of Image Manifolds by Semidefinite Programming. *International Journal of Computer Vision*, 70(1):77–90, 2006.

[28] K. Q. Weinberger, B. D. Packer, and L. K. Saul. Nonlinear Dimensionality Reduction by Semidefinite Programming and Kernel Matrix Factorization. In R. Cowell and Z. Ghahramani, editors, *Proceedings of the Tenth International Workshop on Artificial Intelligence and statistics*, pages 381–388, Barbados, 2005. The Society for Artificial Intelligence and Statistics.

[29] Z. Zhang and H. Zha. Principal Manifolds and Nonlinear Dimension Reduction via Local Tangent Space Alignment. *SIAM Journal of Scientific Computing*, 2004.