Video sequences association for people re-identification across multiple non-overlapping cameras

N. Truongcong, Cyrille Achard, L. Khoudour, L. Douadi

Lecture Notes in Computer Science, numéro N5716, pp179-189
Proceedings of ICIAP 2009, 15th International Conference on Image Analysis and Processing, September 2009, Vietri sul Mare, Italy

DOI:10.1007/978-3-642-04146-4_21
http://www.springer.com/series/558

Video sequences association for people re-identification across multiple non-overlapping cameras

D-N. Truong Cong[1], C. Achard[2], L. Khoudour[1], L. Douadi[1]

[1] French National Institute for Transport and Safety Research (INRETS),
20 rue Elise Reclus, 59650 Villeneuve d'Ascq, France
[2] UPMC Univ Paris 06, Institute of Intelligent Systems and Robotics (ISIR)
Case Courrier 252, 3 rue Galile, 94200 IVRY SUR SEINE, France

Abstract. This paper presents a solution of the appearance-based people re-identification problem in a surveillance system including multiple cameras with different fields of vision. We first utilize different color-based features, combined with several illuminant invariant normalizations in order to characterize the silhouettes in static frames. A graph-based approach which is capable of learning the global structure of the manifold and preserving the properties of the original data in a lower dimensional representation is then introduced to reduce the effective working space and to realize the comparison of the video sequences. The global system was tested on a real data set collected by two cameras installed on board a train. The experimental results show that the combination of color-based features, invariant normalization procedures and the graph-based approach leads to very satisfactory results.

1 Introduction

In recent years, public security has been facing an increasing demand from the general public as well as from governments. An important part of the efforts to prevent the threats to security is the ever-increasing use of video surveillance cameras throughout the network, in order to monitor and detect incidents without delay. Existing surveillance systems rely on human observation of video streams for high-level classification and recognition. The typically large number of cameras makes this solution inefficient and in many cases unfeasible. Although the basic imaging technologies for simple surveillance are available today, the reliable deployment of them in a large network is still ongoing research.

In this paper, we tackle the appearance-based people re-identification problem in a surveillance system including multiple cameras with different fields of vision. The video sequences capturing moving people are analyzed and compared in order to re-establish a match of the same person over different camera views located at different physical sites. In most cases, such a system relies on building an appearance-based model that depends on several factors, such as illumination conditions, different camera angles and pose changes. Thus, building an ideal appearance model is still a challenge.

A significant amount of research has been carried out in the field of appearance-based person recognition. Kettnaker and Zabih [1] exploit the similarity of views of the person, as well as plausibility of transition time from one camera to another. Nakajima

et al. [2] present a system which can recognize full-body people in indoor environments by using multi-class SVMs that were trained on color-based and shaped-based features extracted from the silhouette. Javed et al. [3] use various features based on space-time (entry/exit locations, velocity, travel time) and appearance (color histogram). A probabilistic framework is developed to identify best matches. Bird et al. [4] detect loitering individuals by matching pedestrians intermittently spotted in the camera field of view over a long time. Snapshots of pedestrians are extracted and divided into thin horizontal slices. The feature vector is based on color in each slice and Linear Discriminant Analysis is used to reduce the dimension. Gheissari et al. [5] propose a temporal signature which is invariant to the position of the body and the dynamic appearance of clothing within a video shot. Wang et al. [6] represent objects using histograms of oriented gradients that incorporate detailed spatial distribution of the color of objects across different body parts. Yang et al. [7] propose an appearance model constructed by kernel density estimation. A key-frame selection and matching technique was presented in order to represent the information contained in video sequences and then to compare them.

The research presented in this paper is the development of a multi-camera system installed on board trains that is able to track people moving through different sites. Different color-based features combined with invariant normalizations are first used to characterize the silhouettes in static frames. A graph-based approach is then introduced to reduce the effective working space and to realize the comparison of the video sequences.

The organization of the article is as follows. Section 2 describes how the invariant signature of a detected silhouette is generated. In Section 3, after a few theoretical reminders on the graph-based method for dimensionality reduction, we explain how to adapt this latter to our problem. Section 4 presents global results on the performance of our system on a database of given facts. Finally, in Section 5, the conclusion and important short-term perspectives are given.

2 Invariant signature extraction

2.1 Color-based feature extraction

The first step in our system consists in extracting from each frame a robust signature characterizing the passage of a person. To do this, a detection of moving objects (silhouettes in our case), by using a background subtraction algorithm [8], combined with morphological operators is first carried out. Let us assume now that each person's silhouette is located in all the frames of a video sequence. Since the appearance of people is dominated by their clothes, color features are suitable for their description.

The most widely used feature for describing the color of objects is the *color histograms* [3]. Given a color image I of N pixels, the histogram is produced first by discretization of the colors in the image into M bins $b = \{1, ..., M\}$, and counting the number of occurrences per bin:

$$n_b = \sum_{k=1}^{N} \delta_{kb} \tag{1}$$

where $\delta_{kb} = 1$ if the k^{th} pixel falls in bin b and $\delta_{kb} = 0$ otherwise. The similarity between two histograms can be computed using histogram intersection. Although histograms are robust to deformable shapes, they cannot discriminate between appearances that are the same in color distribution, but different in color structure, since they discard all spatial information.

A limited amount of spatial information of histograms can be retained by using *spatiograms* [9] that are a generalization of histograms, including higher order spatial moments. For example, the second-order spatiogram contains, for each histogram bin, the spatial mean and covariance:

$$\mu_b = \sum_{k=1}^{N} \mathbf{x}_k \delta_{kb} \, , \; \Sigma_b = \sum_{k=1}^{N} \left(\mathbf{x}_k - \mu_b\right) \left(\mathbf{x}_k - \mu_b\right)^T \delta_{kb} \tag{2}$$

where $\mathbf{x}_k = [x, y]$ is the spatial position of pixel k. To compare two spatiograms (S, S'), the following similarity measure is used:

$$\rho\left(S, S'\right) = \sum_{b=1}^{M} \psi_b \rho_n \left(n_b, n'_b\right) \tag{3}$$

where $\rho_n \left(n_b, n'_b\right)$ is the similarity between histogram bins and ψ_b is the spatial similarity measure, given by:

$$\psi_b = \eta \exp\left\{-\frac{1}{2} \left(\mu_b - \mu'_b\right)^T \hat{\Sigma}_b^{-1} \left(\mu_b - \mu'_b\right)\right\} \tag{4}$$

where η is the Gaussian normalization term and $\hat{\Sigma}_b^{-1} = \left(\Sigma_b^{-1} + \left(\Sigma'_b\right)^{-1}\right)$.

Another approach for building appearance models is the *color/path-length feature* [7] that includes some spatial information: each pixel inside the silhouette is represented by a feature vector $(\mathbf{v}, l)$, where $\mathbf{v}$ is the color value and l is the length between an anchor point (the top of the head) and the pixel. The distribution of $p\left(\mathbf{v}, l\right)$ is estimated with a 4D histogram.

2.2 Invariant normalization

Since the color acquired by cameras is heavily dependent on several factors, such as the surface reflectance, illuminant color, lighting geometry, response of the sensor…, a color normalization procedure has to be carried out in order to obtain invariant signatures. Many methods have been proposed in literature [10–12] and we tested most of them. In this paper, we only present the three invariances that lead to better results:

- Greyworld normalization [13] is derived from the RGB space by dividing the pixel value by the average of the image (or in the area corresponding to the moving person in our case) for each channel:

$$I_k^* = \frac{I_k}{mean\left(I_k\right)} \tag{5}$$

where I_k is the color value of channel k.

- Normalization using histogram equalization [11] is based on the assumption that the rank ordering of sensor responses is preserved across a change in imaging illuminations. The *rank measure* for level i and channel k is obtained with:

$$M_k(i) = \sum_{u=0}^{i} H_k(u) \Big/ \sum_{u=0}^{Nb} H_k(u) \tag{6}$$

where Nb is the number of quantization steps and $H_k(\cdot)$ is the histogram for channel k.

- Affine normalization is defined by:

$$I_k^* = \frac{I_k - \text{mean}(I_k)}{\text{std}\,(I_k)} \tag{7}$$

Thus, the color normalization is applied inside the silhouette of each person before computing its color-based signature. A comparative study of the different color-based features combined with the invariant normalization procedures is presented in Section 4.

3 Dimensionality reduction for categorization

Dimensionality reduction is an important procedure employed in various high dimensional data analysis problems. It can be performed by keeping only the most important dimensions, i.e. the ones that hold the most information for the task, and/or by projecting some dimensions onto others. A representation of the data in lower dimensions can be advantageous for further processing of the data, such as classification, visualization, data compression, etc.

In recent years, a large number of nonlinear techniques for dimensionality reduction have been proposed, such as Locally Linear Embedding [14], Isomap [15], Laplacian Eigenmaps [16], Diffusion Maps [17]. These techniques can deal with complex nonlinear data by preserving global and/or local properties of the original data in the lower dimensional representation and therefore constitute traditional linear techniques.

In this paper, we only focus on Graph-based methods for nonlinear dimensionality reduction. Sometimes called Diffusion Maps, Laplacian Eigenmaps or Spectral Analysis [18], these manifold-learning techniques preserve the local proximity between data points by first constructing a graph representation for the underlying manifold with vertices and edges. The vertices represent the data points, and the edges connecting the vertices represent the similarities between adjacent nodes. After representing the graph with a matrix, the spectral properties of this matrix are used to embed the data points into a lower dimensional space, and gain insight into the geometry of the dataset.

3.1 Mathematical formulation

Given a set of m frames $\{I_1, I_2, \ldots, I_m\}$, this set is associated to a complete neighborhood graph $G = (V, E)$ where each frame I_i corresponds to a vertex v_i in this graph. Two vertices corresponding to two frames I_i and I_j are connected by an edge

that is weighted by $W_{ij} = K(I_i, I_j) = exp\left(-\frac{d(I_i,I_j)^2}{\sigma^2}\right)$, where $d(I_i, I_j)$ is the distance between two signatures extracted from these two frames, and σ is chosen as $\sigma = mean\left[d(I_i, I_j)\right],\ \forall i, j = 1, \ldots, m\ \ (i \neq j)$. The first step of dimensionality reduction consists in searching for a new representation $\{\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_m}\}$ with $\mathbf{y}_i \in \mathbb{R}^m$ obtained by minimizing the cost function:

$$\phi = \sum_{ij} \|\mathbf{y}_i - \mathbf{y}_j\|_2 W_{ij} \tag{8}$$

Let D denote the diagonal matrix with elements $D_{ii} = \sum_j W_{ij}$ and L denote the unnormalized Laplacian defined by $L = D - W$. The cost function can be reformulated as:

$$\phi = \sum_{ij} \|\mathbf{y}_i - \mathbf{y}_j\|_2 W_{ij} = 2\mathtt{Tr}\left(\mathbf{Y}^T L \mathbf{Y}\right) \ \mathtt{with}\ \ \mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_m] \tag{9}$$

Hence, minimizing the cost function ϕ is proportional to minimizing $\mathtt{Tr}\left(\mathbf{Y}^T L \mathbf{Y}\right)$. Dimensionality reduction is obtained by solving the generalized eigenvector problem

$$Ly = \lambda D y \tag{10}$$

for the d lowest non-zero eigenvalues.

Figure 1 presents an example of the results obtained by the graph-based approach for nonlinear dimensionality reduction. In order to make the representation easier to read, in this example, we use only the first two lowest eigenvectors to create the new coordinate system. The input of this test is a set of 120 frames belonging to 6 video sequences (20 frames per sequence). On the left-hand diagram, the frames are illustrated by points, while the points are directly illustrated by the corresponding silhouettes on the right-hand one. The first two sequences that are shown by the red and green points on the left-hand diagram belong to the same person captured by two cameras with different fields of vision. The other points correspond to the people similarly or differently dressed. One can notice that the new visualization of the frame sets in 2D space preserves almost all their original characteristics: frames belonging to one video sequence are represented by the neighbor points, while the space among the clusters varies according to the similarity of the color-based appearances of the silhouettes. The clusters corresponding to the two sequences of the same person strongly overlap, while the cluster belonging to person P3 who dresses very differently is well-separated from the others.

Hence, the representation of the frame set in the new coordinate system shows that, by evaluating how separate the clusters are, we can measure the similarities among the video sequences. In the following section, we describe how to apply it to our problem of people re-identification.

3.2 Implementation for people re-identification

As mentioned in the introduction, the objective of this framework consists in re-identifying a person who has appeared in the field of one camera (e.g. camera 1) and then reappears

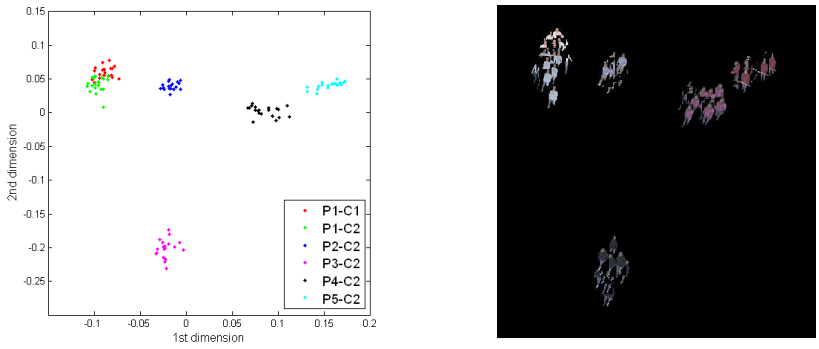


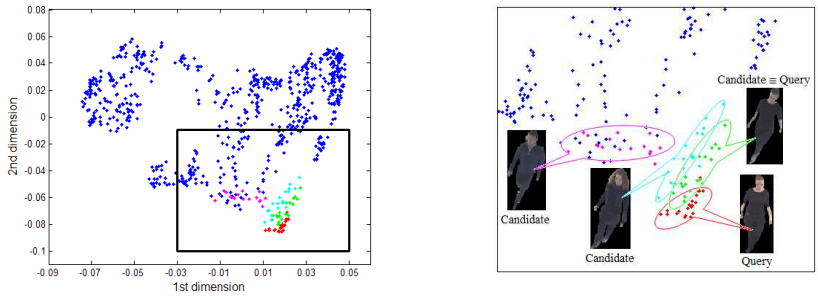
Fig. 1. Example obtained by the graph-based approach for nonlinear dimensionality reduction.

in front of another camera (e.g. camera 2). Thus, a set of m frames $\{I_1, I_2, \ldots, I_m\}$ belonging to p passages in front of camera 1 and the query passage in front of camera 2, is considered for the re-identification (let us notice that a passage is characterized by the concatenation of several frames of the sequence denoted by n). Thus, $m = n * (p + 1)$. By applying the graph-based method for nonlinear dimensionality reduction, we obtain the new coordinate system by considering the d lowest eigenvectors. In our current problem, we use the 20 lowest eigenvectors to create the new coordinate system. The dimensionality reduction operator can be defined as $h : I_i \rightarrow u_i = [y_1(i), \ldots, y_{20}(i)]$ where $y_k(i)$ is the i^{th} coordinate of eigenvector $\mathbf{y}_k$.

Since each passage is represented by the signatures of n frames, the barycentre of the n points obtained by projecting the n frames into the new coordinate system is calculated. The distance between two barycentres is considered as the dissimilarity measure between two corresponding people. The larger the distance, the more dissimilar the two people.

Hence, the dissimilarities between the query person detected in front of camera 2 and each candidate person of camera 1 are calculated and then classified in increasing order. Then, normally and in a perfect system, if the same person has been detected by the two cameras, the lowest distance between the barycentres should lead to the correct re-identification.

Figure 2 presents an example of the results obtained by such a procedure. In this example, we use only the first two principal components. In figure 2a, the query passage is shown by the red dots, while the candidates are represented by other color dots. Figure 2b is an illustration of the most interesting portion of figure 2a (surrounded by a square) enlarged to highlight more details. In figure 2b, the red dots correspond to the set of signatures belonging to the query passage; the green, cyan and yellow dots correspond respectively to the set of frames that have the nearest barycentres, in increasing order, when compared to the query barycentre. The silhouettes corresponding to these four sets of signatures are added in figure 2b. We notice that the silhouettes related to the query set and the nearest candidate set correspond to the same person, or, in other words, we get a true re-identification in this case.



(a) (b)

Fig. 2. Visualization of the set of signatures in the 2D space obtained by applying the graph-based dimensionality reduction method.

4 Experimental results

Our algorithms were entirely evaluated on real data sets containing video sequences of 35 people collected by two cameras installed on board a train at two different locations: one in the corridor and one in the cabin. This dataset is very difficult, since these two cameras are set up with different angles and the acquisition of the video is influenced by many factors, such as fast illumination variations, reflections, vibrations. Figure 3 illustrates an example of the dataset. For each passage in front of a camera, we extract 20 frames regularly spaced in which people are entirely viewed.

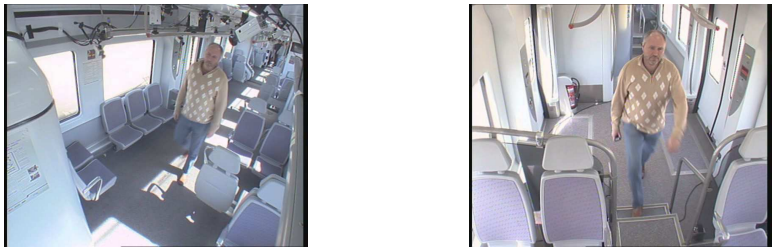


Fig. 3. Illustrations of the real dataset representing the same person in two different locations: in the cabin (left) and in the corridor (right).

In our experimentations, we calculate three types of signatures mentioned in Section 2: color histograms and spatiograms with 8 bins for each color channel, color/path-length descriptor with 8 bins per color channel and 8 bins for the path-length feature. For each silhouette, the illuminant invariant procedure is applied before extracting the signatures. Then, as described in section 3, for each query passage in front of one camera, the distances (i.e the dissimilarities) between the query person and each of the candidate people of the other camera are calculated. Distances are then classified in increasing order and the probability of correct re-identification at rank k is calculated. This leads to a Cumulative Match Characteristic (CMC) curve that illustrates the performance of our system.

Figure 4, which is divided into four parts according to the illuminant invariant, represents the CMC curves obtained by combining three color-based signatures with the graph-based method for dimensionality reduction. The rates of re-identification at the top rank are very satisfying: the best rate of 97% is obtained by combining spatiograms, normalization using histogram equalization and the graph-based method for sequence comparison. We note that the performance of color histograms is the worst. The others lead to better results thanks to the additional spatial information. The invariant normalizations have actually improved the results compared to the RGB space: the re-identification rate obtained by using spatiograms increases from 51% for RGB space to 97% for the histogram equalization.

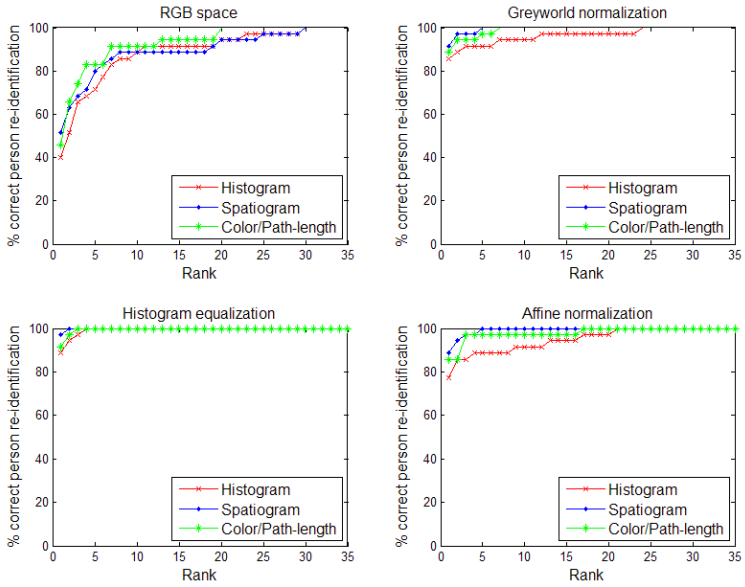


Fig. 4. CMC curves corresponding to four color spaces obtained by using the color-based signature coupled with the graph-based approach for comparing sequences.

Figure 5 shows an example of the top five matching sequences for several query passages. The query passages are shown in the left column, while the remaining columns present the top matches ordered from left to right. The red box highlights the candidate sequence corresponding to the same person of the query. In this figure, the two cases of the first and third rows correspond to a true re-identification, while the second row falls in a false re-identification (the correct match is not the nearest sequence).

5 Conclusion and perspectives

In this paper, we have presented a novel approach for people re-identification in a surveillance system including multiple cameras with different fields of vision. Our approach relies on the graph-based technique for dimensionality reduction which is ca-

Fig. 5. Example of the top five matching sequences for several query passages.

pable of learning the global structure of the manifold and preserving the properties of the original data in the low-dimensional representation. The first step of our system consists in extracting a feature that describes the person in each selected frame of each passage. Three color-based descriptors (histograms, spatiograms and color/path-length) combined with several invariant normalization algorithms are utilized for this step. Since the passage of a person needs to be characterized by several frames, a large quantity of data has to be processed. Thus, the graph-based method for dimensionality reduction is applied to reduce the effective working space and realize the comparison of video sequences.

The global system was tested on a real data set collected by two cameras installed on board a train under real difficult conditions (fast illumination variations, reflections, vibrations, etc). The experimental results show that the combination of color-based features, invariant normalization procedures and the graph-based approach leads to very satisfactory results: 97% for the true re-identification rate at the top rank.

In order to further improve the performance of our system, the appearance-based features need to add more temporal and spatial information in order to be further discriminating among different people and to be unifying in order to make coherent classes with all the features belonging to the same person. The other features, such as camera transition time, moving direction of the individual, biometrics features (face, gait) should also be considered in order to improve the performance of the re-identification system, especially in the more challenging scenarios (multiple passages in front of cameras, many people wearing same color clothes, occlusion, partial detection, etc). More extensive evaluation also needs to be carried out. A good occasion will be to test them, more intensively, through BOSS European project. On-board automatic video surveillance is a challenge due to the difficulties in dealing with fast illumination variations,

reflections, vibrations, high people density and static/dynamic occlusions that perturb actual video interpretation tools.

References

1. Kettnaker, V., Zabih, R.: Bayesian multi-camera surveillance. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Volume 2. (1999)
2. Nakajima, C., Pontil, M., Heisele, M., Poggio, T.: Full body person recognition system. Pattern Recognition **36**(9) (2003) 1997–2006
3. Javed, O., Rasheed, Z., Shafique, K., Shah, M.: Tracking across multiple cameras with disjoint views. In: Ninth IEEE International Conference on Computer Vision. (2003)
4. Bird, N., Masoud, O., Papanikolopoulos, N., Isaacs, A.: Detection of loitering individuals in public transportation areas. IEEE Transactions on Intelligent Transportation Systems **6**(2) (2005) 167–177
5. Gheissari, N., Sebastian, T., Hartley, R.: Person reidentification using spatiotemporal appearance. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, IEEE Computer Society (2006) 1528–1535
6. Wang, X., Doretto, G., Sebastian, T., Rittscher, J., Tu, P.: Shape and appearance context modeling. In: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on. (2007) 1–8
7. Yu, Y., Harwood, D., Yoon, K., Davis, L.: Human appearance modeling for matching across video sequences. Machine Vision and Applications **18**(3) (2007) 139–149
8. Kim, K., Chalidabhongse, T., Harwood, D., Davis, L.: Background modeling and subtraction by codebook construction. In: International Conference on Image Processing, ICIP'04. Volume 5. (2004)
9. Birchfield, S., Rangarajan, S.: Spatiograms versus histograms for region-based tracking. IEEE Computer Society Conference on Computer Vision and Pattern Recognition **2** (2005) 1158–1163
10. Gevers, T., Stokman, H.: Robust histogram construction from color invariants for object recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence (2004) 113–118
11. Finlayson, G., Hordley, S., Schaefer, G., Yun Tian, G.: Illuminant and device invariant colour using histogram equalisation. Pattern Recognition **38**(2) (2005) 179–190
12. Madden, C., Piccardi, M., Zuffi, S.: Comparison of Techniques for Mitigating the Effects of Illumination Variations on the Appearance of Human Targets. Volume 4842 of Lecture Notes in Computer Science. Springer (2007)
13. Buchsbaum, G.: A spatial processor model for object color perception. Journal of the Franklin Institute **310**(1) (1980) 1–26
14. Roweis, S., Saul, L.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**(5500) (2000) 2323–2326
15. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science **290**(5500) (2000) 2319–2323
16. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation **15**(6) (2003) 1373–1396
17. Nadler, B., Lafon, S., Coifman, R.R., Kevrekidis, I.G.: Diffusion maps, spectral clustering and eigenfunctions of fokker-planck operators. In: Advances in Neural Information Processing Systems. (2005) 955–962
18. von Luxburg, U.: A tutorial on spectral clustering. Statistics and Computing **17**(4) (2007) 395–416