Video Analysis Via Nonlinear Dimensionality Reduction

Alvaro Pardo*

DIE, Facultad de Ingeniería y Tecnologías, Universidad Católica del Uruguay
apardo@ucu.edu.uy

Abstract. In this work we present an application of nonlinear dimensionality reduction techniques for video analysis. We review several methods for dimensionality reduction and then concentrate on the study of Diffusion Maps. First we show how diffusion maps can be applied to video analysis. For that end we study how to select the values of the parameters involved. This is crucial as a bad parameter selection produces misleading results. Using color histograms as features we present several results on how to use diffusion maps for video analysis.

Keywords: video analysis, dimensionality reduction.

1 Introduction

Most of the available techniques for video analysis begin reducing the amount of information via feature extraction. Usually this means a strong reduction in the amount of data contained in a video sequence. Recently, due to the increase of memory and processing power in actual computers, some algorithms for video analysis are based on a more detailed description of the video using a big number of features [7,10]. Some of these methods can be rooted to image analysis method that use all the pixels and visualize the image as a point in a high dimensional space [7]. Other methods for video analysis recently proposed in the literature use pixel-wise histograms to describe video segments [6]. These methods keep a lot of information from the beginning of the processing. Although, it has been shown that this is beneficial, novel methods for high dimensional data analysis must be developed.

Although the idea of high dimensional data analysis is not new, recently several authors presented new results in this direction. One of the main problems in the context of high dimensional data analysis is dimensionality reduction. The two main goals of this step are: visualization and extraction of smaller set of meaningful and useful coordinates. In what follows we present a brief overview of existing methods for dimensionality reduction.

The most popular method for dimensionality reduction is Principal Component Analysis (PCA). PCA finds the basis of a projection space (of smaller dimension) minimizing the square reconstruction error. It is well known that the

* Supported by Proyecto PDT-S/C/OP/46/18. A. Pardo is on leave from Facultad de Ingeniería, Universidad de la República.

L. Rueda, D. Mery, and J. Kittler (Eds.): CIARP 2007, LNCS 4756, pp. 152–161, 2007.

subspace which produces the minimum reconstruction error is the subspace with maximum variance. Therefore, this method intends to preserve the covariance structure [9]. PCA has two advantages. First, the projection is performed with a *linear transformation* which is extremely easy to apply. Second, any new vector can be easily projected. Unfortunately, not all spaces are linear and a linear combination of basis vectors will not produce good results. In the same category we can find Independent Component Analysis (ICA) which also uses a linear projection and Multi Dimensional Scaling [9].

The main drawback of the previous methods is that they are not capable of dealing with nonlinear spaces[1]. For this reason, recently several authors presented nonlinear methods for dimensionality reduction. Some of them are graph based methods. Basically, the idea is to discover the underlying structure of the data constructing a graph which joins data points with a given criteria. ISOMAP [11] is one example of these techniques. Other methods such as Locally Linear Embedding (LLE) [8], Laplacian and Hessian Eigenmaps [1,3], and others [5] minimize the reconstruction error using a local linear expansion. That is, each sample is linearly reconstructed using nearby samples. In this way this method overcomes the linear limitation. Diffusion Maps [5,2] provide a unified vision of previous spectral methods in a unified framework. Also, this method includes a natural notion of scale and distance. In next section we review diffusion maps. The weakness of these methods is that is difficult to project a new sample in the obtained projected space. The embedding is given by the data and there is no general method to obtain the projection. In [4] the authors propose a solution to solve the extension to new data points.

Before concluding this section we review some works that apply nonlinear dimensionality reduction for video analysis. In [7] Pless uses Isomap to explore video sequences. The results are for simple sequences. The work is similar to ours but it does not include the notion of scale given by diffusion maps. We also present results with general sequences with transitions. In [10] uses LLE to discover periodicity in video sequences.

The outline of the paper is the following. In next section we present a detailed review of diffusion maps. In section 3 we analyze how to select the appropriate diffusion map parameter values. Then, in section 4 we present several examples of diffusion maps applied to video analysis. Finally in section 5 we discuss our results and present our main conclusions

2 Review of Diffusion Maps

In this section we review Diffusion Maps (DM) following [5]. Let $\Omega = \{x_1, ..., x_n\}$ be a set of data points of dimension N and $d(x_i, x_j)$ a distance between data points. The idea behind DM is to construct a graph with each data point x_i being a vertex and $w(x_i, x_j)$ a weight function between vertices. In what follows we restrict ourselves to $w(x_i, x_j) = \exp(d_\Omega(x_i, x_j)^2)/\sigma^2)$. This graph intends to reflect the knowledge of the local geometry of Ω. Once we have the definition of

[1] There are some extensions of PCA and other methods to deal with this problem.

the graph a Markov random walk can be defined over it. If $d(x) = \sum_{z \in \Omega} w(x,z)$ is the degree of node x the quantity $p_1(x,y) = \frac{w(x,y)}{d(x)}$ can be interpreted as the probability of transition from x to y. The 1 means that this transition is made in one step and therefore reflects first order information of the graph structure. Let the matrix P be the one with entries $p_1(x,y)$. Considering powers, P^t, information over larger neighborhoods can be captured. In this way $p_t(x,y)$ means the probability of transition form x to y in t steps. As t increases P^t captures more global information and this enables us to view t as a scale parameter.

If the graph is connected it can be shown that: $\lim_{t \to \infty} p_t(x,y) = \phi_0(y) = \frac{d(y)}{\sum_{z \in \Omega} d(z)}$ where $\phi_0(x)$ is the stationary distribution. Based on the above elements the distance between points in Ω can be computed as the distance between its corresponding distributions p_t. Thus, the diffusion distance, $D_t(x,y)$ is defined as:

$$D_t^2(x,y) = ||p_t(x,.) - p_t(y,.)||_{1/\phi_0}^2 = \sum_{z \in \Omega} \frac{(p_t(x,z) - p_t(y,z))^2}{\phi_0(z)}.$$

Observe that the distance includes the normalization by $\phi_0(z)$ which is used to decrease the influence of points with small densities. The main result for diffusion distance is the following. The diffusion distance can be expressed as:

$$D_t^2(x,y) = \sum_{i=1}^{n-1} \lambda_i^{2t} (\psi_i(x) - \psi_i(y))^2, \tag{1}$$

where λ_i and ψ_i are the eigenvalues and eigenvectors of P ($P\psi_i = \lambda_i \psi_i$). It can be shown that $1 = |\lambda_0| \geq |\lambda_1| \geq ... \geq |\lambda_{n-1}|$ and $\psi_0 = 1$. Due to the ordering of eigenvalues the diffusion distance can be approximated taking only the first coordinates. The number of retained terms depends on the desired precision and on t. The diffusion map is defined as:

$$\Psi_t : x \to \left(\lambda_1^t \psi_1(x), \lambda_2^t \psi_2(x), ..., \lambda_{M(t)}^t \psi_{M(t)}(x)\right)^t, \tag{2}$$

where $M(t)$ is the number of retained terms. The mapping projects the graph information to a lower dimensional space.

3 Application of Diffusion Maps

DM not only project the data points to a lower dimensional space, but also provide a notion on scale t and precision $M(t)$. This means that we obtain a lower set of coordinates with an associated significance score. In this section we show how to use these ideas for video analysis.

First we must discuss the representation of each image in the video sequence and the associated distance, d_Ω. In this work we will use a representation based on histograms[2]. We describe each frame with its histogram $h_i(q)$. For color

[2] We are currently investigating a representation based on pixels.

frames we will take $x_i = \{h_i^R(q), h_i^G(q), h_i^B(q)\}$ a concatenation of the histograms of each color channel. The distance in the original space ,$d_\Omega(x_i, x_j)$, will be the L_2 distance[3]. In following paragraphs we address the specification of the DM parameters: σ and $M(t)$.

How can we select σ? The value of σ must be carefully selected since it determines the final graph and P. This parameter affects the weights, $w(x, y)$, and with them the connection between nodes. If σ is set too big it may cause the graph to be fully connected while a too small value may produce a completely disconnected one. At the end of the day its value is reflected in the graph topology which is in turn what we expect to obtain as a natural description of the data.

To set the value of σ we assume that each point (frame), x_i, must be connected with at least two other points (typically nearby frames). We set the minimum weight for farthest points with distance d_{max} as θ so:

$$\sigma = \frac{d_{max}}{\sqrt{\log(1/\theta)}}.$$

In all the experiments we use $\theta = 0.1$.

How can we decide the value of $M(t)$? It is clear form equation (1) that $M(t)$ determines the precision of the distance approximation. Since are ordered and less or equal than one, we know that to achieve a certain level of approximation we must retain the first coordinates. From equation (2) we conclude that the diffusion map gives a parametrization of the data in a lower dimensional space. Furthermore, the scale of the dimensionality reduction is given by t and the decay of the eigenvalues.

How many data points are needed? In the case of video we may encounter restrictions on the amount of data points available with respect to the dimensionality of the feature space. On one hand DM are insensitive to points density and therefore permit to recover intrinsic data properties [4]. On the other hand, they are more resilient to the number of samples comparing with other existing methods such as [8]. We will confirm this in the experiments where we have in some cases more features than samples. We will have three histograms with 256 bins producing a feature vector of dimension 768 while some of the video sequences have around 300 frames.

Do we need to include temporal information? Traditionally, video analysis is strongly linked to temporal relations. In fact, most of existing methods study the distance between frames at different times[4]. This turns to produce an aperture problem since we observe the data across a given temporal window. Diffusion Maps, and other of the methods reviewed above, enable us to link frames without taking into account small temporal windows. At the end of the day, the method discovers the relevant coordinates within the data and sorts them according to its relevance. This interesting feature comes with the expense

[3] Other distance can be used.

[4] Shot detection is a classical example.

of bigger data sets. In the experiments we will show that there is no need to explicitly include temporal information.

4 Experiments

In this section we present a set of experiments of video analysis using the ideas discussed in previous sections. In the two first experiments we show the results for video sequences which contain a smooth transition. Later we will show some results on abrupt changes which are easier to detect.

In the first experiment, see Figure 1 we processed a sequence with a wipe which starts at frame 120 and ends at frame 165. If we observe the first coordinate of the diffusion map at Figure 1 we can see that it correctly detects this smooth transition. From frames 1 to 119 and from frames 166 to 259 this coordinates has little variation. This means that when looking a coarse description these two sets are disjoint. Between frames 120 and 165 we observe a gradual transition between both sets. Therefore, in this case the first coordinate correctly captures the essence of the video sequence. If we observe the remaining coordinates up to the fourth one, we confirm that in finer scales this transition can also be detected. In Figure 1-(b) we depict the video trajectory along the first three

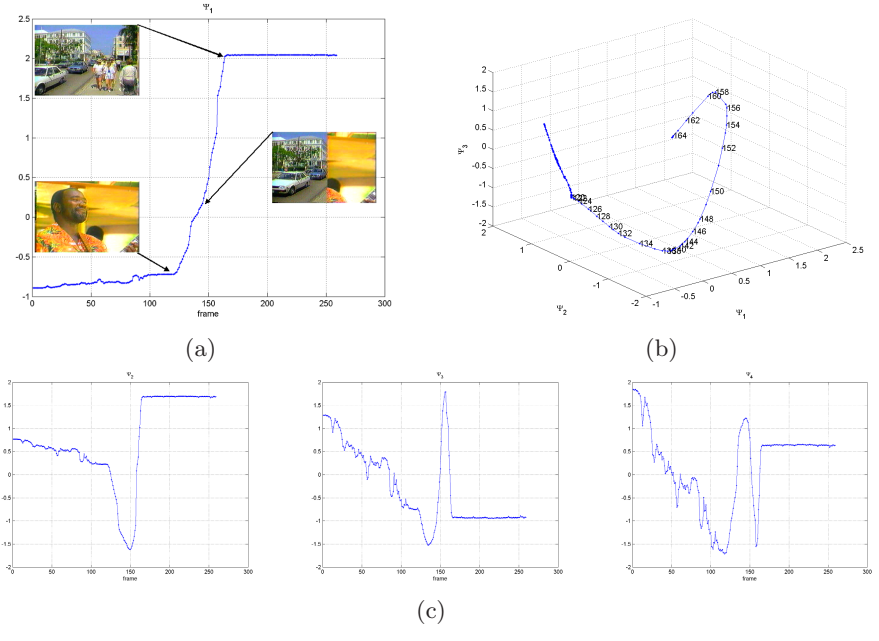


Fig. 1. (a) First coordinate showing clearly showing the structure of the sequence. (b) Video trajectory in the first three coordinates. As we can see the beginning and end of the sequence produces two clusters while the wipe-transition produces a trajectory between them. (c) Remaining coordinates.

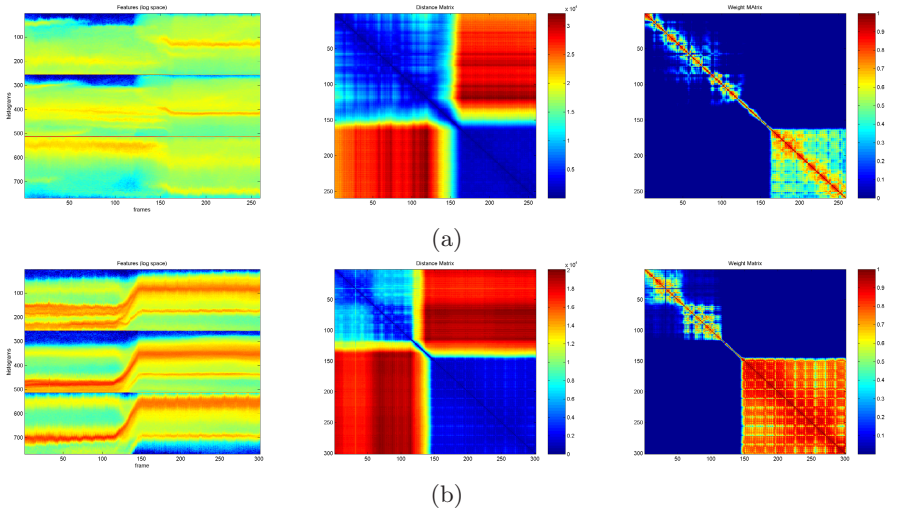


Fig. 2. (a) Features along frames, Distance Matrix and Weight Matrix for sequence in Figure 1. (b) The same for sequence in Figure 3. Observe the video structure.

coordinates. As we can see, there is a transition which expands form frames 120 to 165. To complete the information for this sequence we show the histograms evolution along frames, the Distance and Weight matrices (see Figure 2).

The second experiment presents the results for a dissolve transition. These are the most difficult transitions to detect. As we can see in Figure 3 in this case we can successfully detect the transition. The dissolve starts at frame 115 and ends in frame 145. It is interesting to note the stability of the first coordinate at the beginning and end of the sequences. This shows that the first coordinates classifies, at coarse level, all these frames in the same cluster. This behavior is repeated in the other coordinates. Obviously, as we increase the coordinate number, its corresponding eigenvalues decreases and with it its importance. Finally we note a peak in coordinate Ψ_{10} at frame 31. This is due to an error on the video as can be seen in Figure 3-(b). Therefore, at finer scales we are able to detect such small discrepancies between frames. As we did in the first experiment we show other complementary information in Figure 2.

For our third experiment we used sequence with 1561 frames and mainly abrupt transitions. Observing Figure 4 we see that once again the first components successfully summarize the characteristic of the sequence. If we look at finer scale, in coordinates Ψ_8 to Ψ_{10}, we can see a gradual changes within a shot obtained at coarser scales. This gradual changes are caused by a panning (see Ψ_10 in Figure 4).

To further evaluate the power of discrimination of the method we tested the algorithm with a small sequence with only smooth transitions. In this case, the first coordinate clearly detects a shot from frames 175 to 200, however, it is difficult to declare other shots while looking only the first coordinate. On the

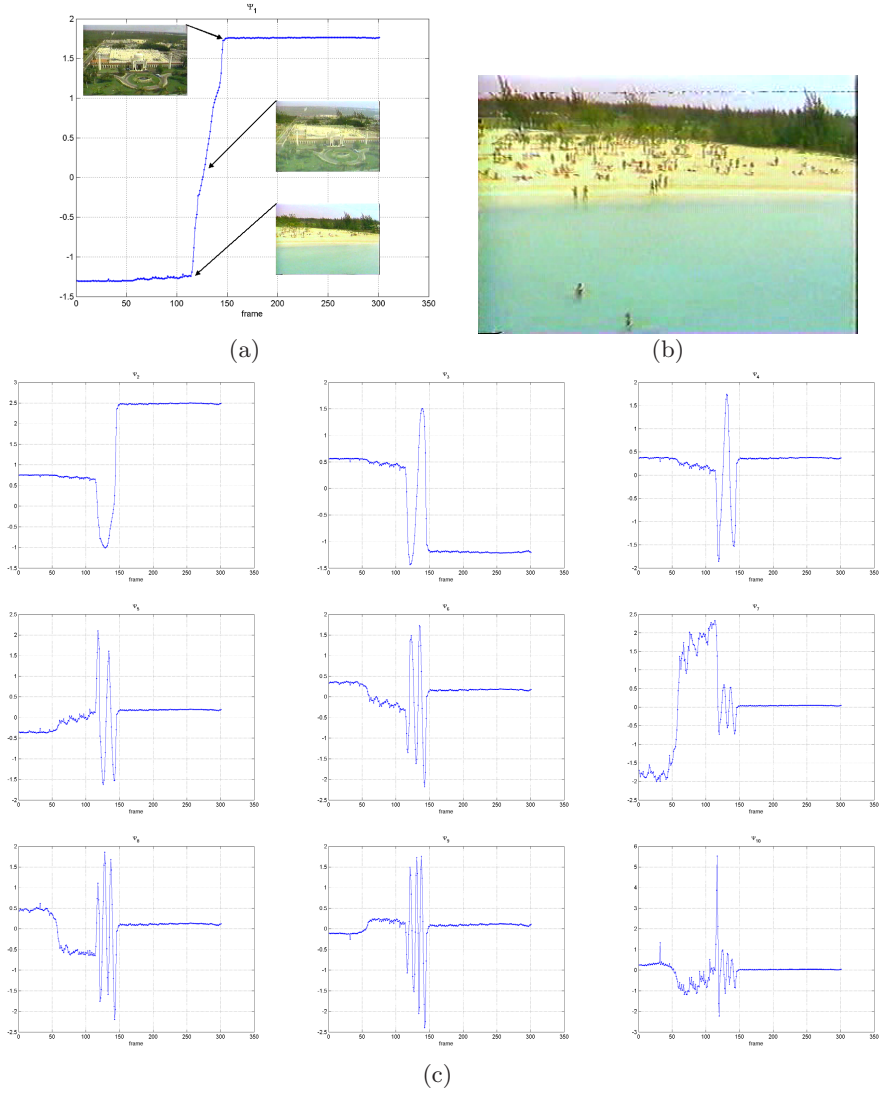
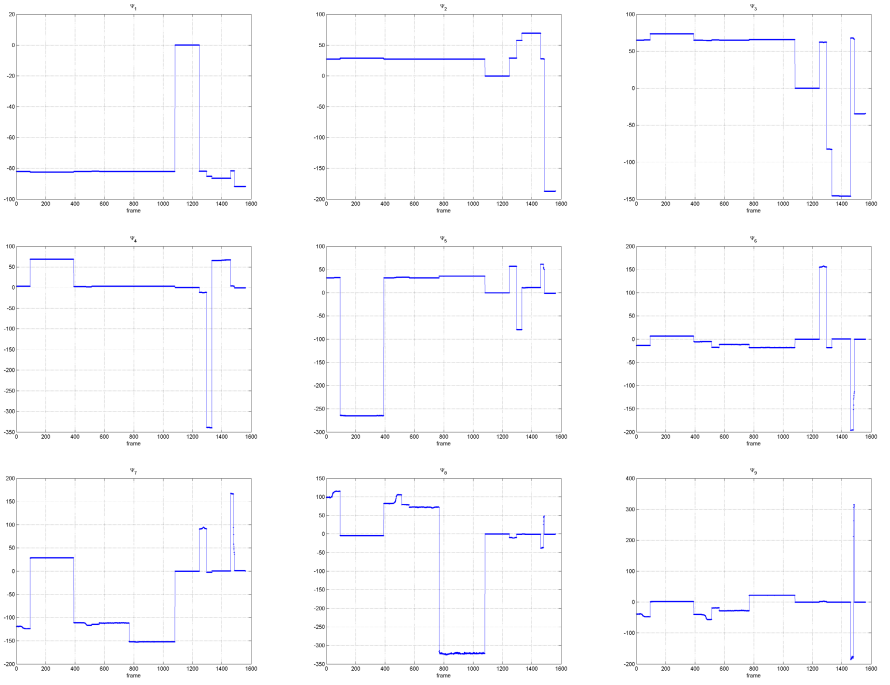
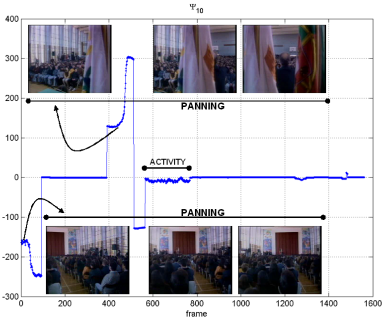


Fig. 3. (a) First coordinate showing clearly showing the structure of the sequence. (b) Frame with error visible at finer scales (c) Remaining coordinates.

other hand, if we observe Ψ_2 we can see a block of frames from frames 45 to 125. However, there is still a big variation within this block which indicates other cluster at finer scale. This is can be confirmed observing Ψ_4, Ψ_5 and Ψ_7. Hence, we conclude that the method effectively detects the structure of the video sequence. However, to do so we must observe several coordinates to discriminate at finer scales.



(a)



(b)

Fig. 4. (a) First nine coordinates. (b) Details at finer scales showing pannings and some detailed activity within a shot that it is not visible at coarse levels.

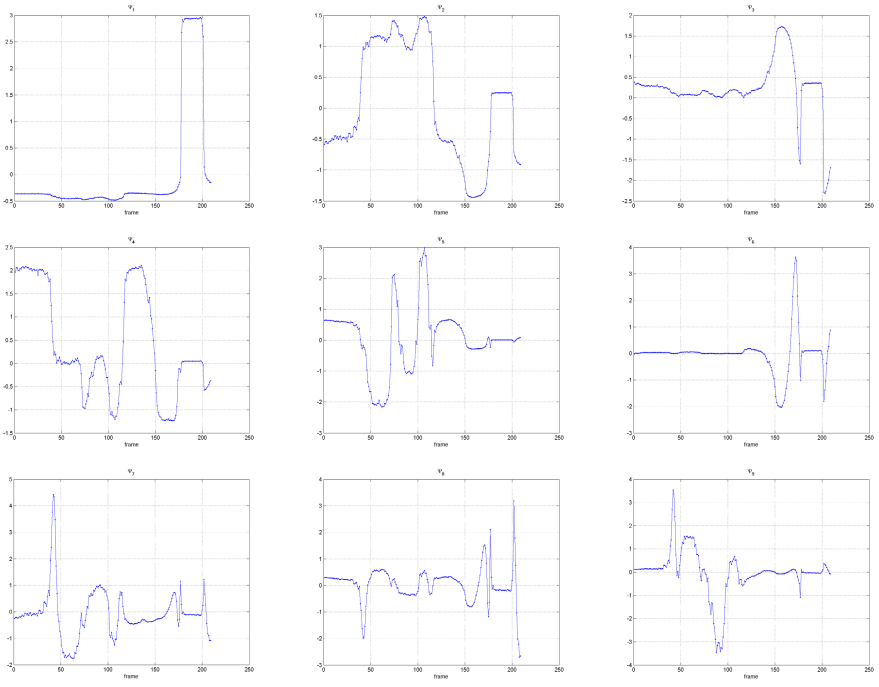


Fig. 5. First nine coordinates for the fourth example

5 Discussion and Conclusions

In this work we study the application of DM to video analysis. We showed how this method can be successfully applied. We presented estimations for the method parameters and confirmed this in the experiments. We showed how the coordinates obtained compress the information of the sequence structure in few coordinates. Although in several cases the information is compressed in the first few components, this depends on the sequences, and in some cases we will need to explore finer scales. This was confirmed with our last experiments. Therefore, although this are mainly preliminary results, they are very promising. We are currently testing other descriptions and a more exhaustive evaluation of the results and their comparison against other methods.

References

1. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput. 15(6), 1373–1396 (2003)
2. Coifman, R., Lafon, S.: Diffuion maps. Applied and Computational Harmonic Analysis 21, 5–30 (2006)

3. Donoho, D., Grimes, C.: Hessian eigenmaps: locally linear embedding techniques for high dimensional data. Proc. of National Academy of Sciences 100(10), 5591–5596 (2003)
4. Lafon, S., Keller, Y., Coifman, R.: Data fusion and multicue data matching by diffusion maps. IEEE Trans. Pattern Anal. Mach. Intell. 28(11), 1784–1797
5. Lafon, S., Lee, A.B.: Diffusion maps and coarse-graining: a unified framework for dimensionality reduction, graph partitioning, and data set parameterization. IEEE Trans. on Pattern Anal. and Mach. Intell. 28(9), 1393–1403 (2006)
6. Pardo, A.: Pixel-wise histograms for visual segment description and applications. In: Martínez-Trinidad, J.F., Carrasco Ochoa, J.A., Kittler, J. (eds.) CIARP 2006. LNCS, vol. 4225, pp. 873–882. Springer, Heidelberg (2006)
7. Pless, R.: Image spaces and video trajectories: Using isomap to explore video sequences. In: ICCV, pp. 1433–1440 (2003)
8. Roweis, S., Saul, L.: Nonlinear dimensionality reduction by locally linear embedding. Science 290, 2323–2326 (2000)
9. Saul, L., Weinberger, K., Sha, F., Ham, J., Lee, D.: Spectral methods for dimensionality reduction. In: Schoelkopf, B., Chapelle, O., Zien, A. (eds.) Semisupervised Learning, MIT Press, Cambridge (2006)
10. Stich, T., Magnor, M.: Keyframe Animation from Video. In: ICIP 2006, pp. 2713–2716 (2006)
11. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science 290, 2319–2323 (2000)