Proc. Neural Information Processing (ICONIP 2007), Kitakyushu, LNCS 4984, pp.653-662, Springer 2008

# Walking Appearance Manifolds without Falling Off

Nils Einecke[1], Julian Eggert[2], Sven Hellbach[1], and Edgar Körner[2]

[1] Technical University of Ilmenau
Department of Neuroinformatics and Cognitive Robotics
98684 Ilmenau, Germany
[2] Honda Research Institute Europe GmbH
Carl-Legien-Str.30, 63073 Offenbach/Main, Germany

**Abstract.** Having a good description of an object's appearance is crucial for good object tracking. However, modeling the whole appearance of an object is difficult because of the high dimensional and nonlinear character of the appearance. To tackle the first problem we apply nonlinear dimensionality reduction approaches on multiple views of an object in order to extract the appearance manifold of the object and to embed it into a lower dimensional space. The change of the appearance of the object over time then corresponds to a walk on the manifold, with view prediction reducing to a prediction of the next step on the manifold. An inherent problem here is to constrain the prediction to the embedded manifold. In this paper, we show an approach towards solving this problem by applying a special mapping which guarantees that low dimensional points are mapped only to high dimensional points lying on the appearance manifold.

## 1 Introduction

One focus of the current research in computer vision is to find a way to represent the appearance of objects. Attempts of full 3D modeling of an object's 3D shape turned out to be not reasonable as it is computationally intensive and learning or generating appropriate models is laborious. According to the viewer-centered theory [1,2,3] the human brain stores multiple views of an object in order to be able to recognize the object from various view points. For example in [4] an approach is introduced that uses multiple views of objects to model their appearance. Thereto the desired object is tracked and at each time step the pose of the object is estimated and a view is inserted into the model of appearance if it holds new or better information. Unfortunately, this is very time consuming as this approach works directly with the high dimensional views.

Actually, the different views of an object are samples of the appearance manifold of the object. This manifold is a nonlinear subspace in the space of all possible appearances (appearance space) where all the views of this particular object are located. In general, the appearance manifold has a much lower dimensionality than the appearance space it is embedded in. Non-Linear Dimensionality Reduction (NLDR) algorithms, like Locally Linear Embedding (LLE)

M. Ishikawa et al. (Eds.): ICONIP 2007, Part I, LNCS 4984, pp. 653–662, 2008.

[5], Isometric Feature Mapping (Isomap) [6] or Local Tangent Space Alignment (LTSA) [7], can embed a manifold into lower dimensional spaces (embedding space) by means of a sufficient number of samples of the manifold. Elgammal and Lee [8] use embedded appearance manifolds for 3D body pose estimation of humans based on silhouettes of persons and LLE. Pose estimation is realized via a RBF[1]-motivated mapping from the visual input to the embedded manifold and from there to the pose space. Note that they model the embedded manifold with cubic splines in order to be able to project points mapped into the embedding space onto the manifold. Lim et al. [9] follow a similar approach but in contrast to Elgammal and Lee they use the model of the embedded manifold to predict the next appearance. Actually, both approaches are limited to one-dimensional manifolds as the views were sampled during motion sequences and the modeling of the manifold is based on the available time information. In [10] Lui et al. use an aligned mixture of linear subspace models to generate the embedding of the appearance manifold which does not dependent on additional time information. Using a Dynamic Bayesian Network they infer the next position in the embedding space and based on this the position and scale parameters.

The approach of Lui et al. is able to handle manifolds with more than one dimension but the prediction process is not constrained to the structure of the manifold. This, however, is very important for predictions over a larger time span as without this constraint the prediction would tend to leave the manifold, leading to awkward views when projected back to the appearance space or to wrong pose parameter estimates. Unfortunately, this constraining is quite difficult because of the highly nonlinear shape of the manifold. In the work presented here, we do not attempt to tackle this problem directly. Instead, we just use a simple non-constrained linear predictor in the low dimensional embedding space and leave the work of imposing the manifold constraint to the mapping procedure between the low dimensional embedding space and the high dimensional appearance space.

The rest of this paper is organized as follows. In Sect. 2 we show what kind of objects we used to investigate our approach and we discuss the shape of appearance manifolds of rigid objects in the light of our way of sampling views. Section 3 introduces our approach for mapping between the spaces which guarantees to map only to points lying on the manifold and its embedding. Then Sect. 4 provides the workflow of our view prediction approach. In Sect. 5 we describe the experiments conducted for analyzing our view prediction approach and present the results. Finally, Sect. 6 summarizes this paper and outlines future work.

## 2 Appearance Manifolds and Embedding

All possible views of an object together form the so-called appearance manifold. By embedding such a manifold in a low dimensional space one gets a low dimensional equivalent of this manifold. If one is able to correctly map between the spaces one can work efficiently in the low dimensional space and project the

[1] Radial Basis Function.

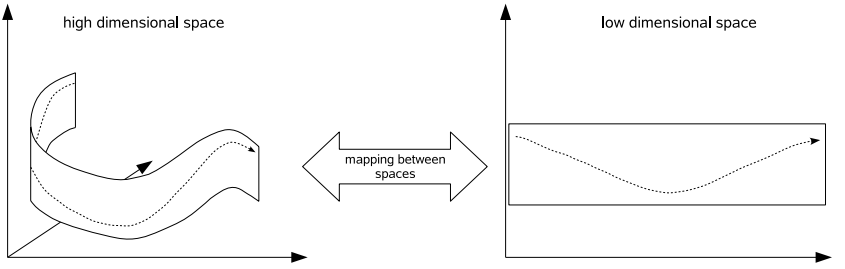


**Fig. 1.** A trajectory on a two-dimensional band-like manifold. On the left we see the actual manifold and on the right its two-dimensional embedding.

results to the high dimensional space. For example series of views exhibit as a trajectory on the appearance manifold. Mapping such a trajectory into the low dimensional space eases the processing as the trajectory's dimensionality and nonlinearity is reduced. Figure 1 shows a simple band-like manifold resident in the three-dimensional space and its embedding in the two-dimensional space.

We used the POV-Ray[2] tool for generating views of virtual objects[3] (see Fig. 2). This way we are able to verify our approach under ideal conditions and, for now, we do not have to deal with problems like segmenting the object from the background. A view of an object is described mainly by the orientation parameters of the object. These could for example comprise: scaling, rotation about the three axis of the three dimensional space, deformation and translation. However, we will concentrate only on the rotation here. While tracking deformation would considerably blow up the complexity of the problem, scaling can be handled by a resolution pyramid. Furthermore, it makes sense to use views which are centered because this could be dealt with by a preprocessing step, like a translational tracker. So we are left with a three-dimensional parameter space spanned by the three rotation angles. In addition, sampling views over all three angles is not feasible as this would lead to a too large number of views. Therefore we decided to sample views by varying only 2 axes. Unfortunately, experiments have shown that, in general, the views sampled varying 2 axes are not embeddable in a non-pervasive manner in a low dimensional (three-dimensional) space. Hence we reconsidered to rotate the objects full 360° only about one axis.

We sampled views every 5° while rotating the object 360° about its vertical axis (y-axis) and tilting it from −45° to +45° about its horizontal axis (x-axis). Each 360° rotation for itself leads to a cyclic trajectory in the appearance space. As these trajectories are neighboring, all views together form a cylindric appearance manifold. This can be seen exemplarily at the embedding of the views of the bee and the bird in Fig. 3. For embedding the appearance manifold into a low dimensional space we use the Isomap approach because comparisons of LLE [5], LTSA [7] and Isomap [6] have shown that Isomap is most appropriate for this purpose.

[2] POV-Ray is a freely available tool for rendering 3D scenes.
[3] The objects we used are templates from http://objects.povworld.org


**Fig. 2.** The objects used for analyzing our view prediction approach

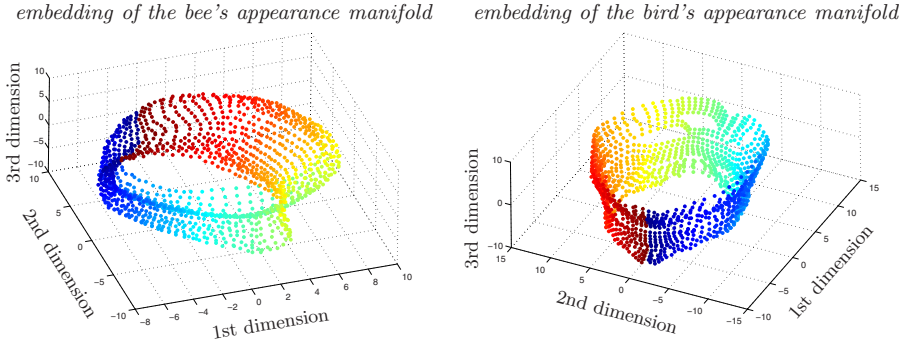


**Fig. 3.** Three-dimensional embeddings of the views of the bee (left) and the bird (right) generated with Isomap. Views were sampled in an area of 360° vertical and from −45° to +45° horizontal. The colors encode the rotation angle about the vertical axis from blue 0° to red 360°. As each full rotation about the vertical axis exhibits as a cyclic trajectory in the appearance space and since all cyclic trajectories are neighboring, the embedding of the views leads to a cylinder-like structure (appearance manifold).

## 3 Mapping between the Spaces

We prefer not to predict views directly in the high dimensional appearance space but on the low dimensional embedding of the appearance manifold. Two problems arise. First, most NLDR algorithms do not yield a function for mapping between appearance space and embedding space, and second, it is difficult to ensure that the prediction does not leave the manifold. In order to actually ensure that the prediction is done only along the manifold one has to constrain the prediction with the nonlinear shape of the manifold. This, however, is very problematic because appearance manifolds often exhibit highly nonlinear and wavy shapes. Take for example a simple linear prediction. Such a prediction is quite likely to predict positions that do not lie on the manifold as can be seen in Fig. 4 a).

Leaving the manifold in the low dimensional space means also leaving the appearance manifold, i.e. for a point in the low dimensional space which is not lying on the embedded manifold there is simple no valid corresponding view of the object. Usual interpolation methods cannot handle this problem. They just try to find an appropriate counterpart but in doing so they are not directly constrained to the appearance manifold. This means that the views they map those points to are no valid views of the object and often show heavy distortions.

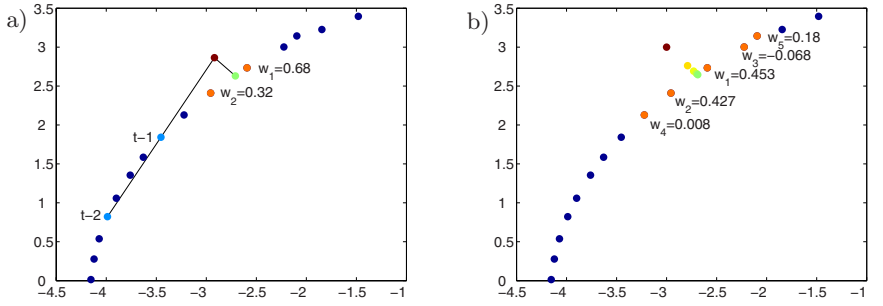


**Fig. 4.** These two figures show a subsection of a one-dimensional cyclic manifold. a) A linear prediction using the last two positions (light blue) on the manifold leads to a point (red) not belonging to the manifold. Reconstructing this point by convex combination of its nearest neighbors (orange) projects it back to the manifold. b) Reconstruction using the LLE idea does not ensure positive weights. However, iterative repetition of the reconstruction (yellow-to-green points) makes the weights converge to positive values. The reconstruction weights after 4 iterations are displayed.

A possible way out of this dilemma is the reconstruction idea upon which LLE [11] is based. What we want to do is to map between two structures whereas one is a manifold in a high dimensional space and the other its embedding in a low dimensional space. By assuming a local linearity (which is a fundamental assumption of most NLDR algorithms anyway) it is possible to calculate reconstruction weights for a point on one of these structures that accounts for both spaces, i.e. it is possible to calculate the reconstruction weights for a point in either of the two spaces and by means of these the counterpart of this point in the other space can be reconstructed. The weights in the appearance space are calculated via minimizing the following energy function

$$E(\mathbf{w}_i) = \left| \mathbf{x}_i - \sum_{j \in N_i} w_i^j \cdot \mathbf{x}_j \right|^2 \quad \text{with} \sum_{j \in N_i} w_i^j = 1, \tag{1}$$

where $\mathbf{x}$ is a $D$-dimensional point in the appearance space, $\mathbf{x}_i$ is the point to reconstruct, $N_i = \{j | \mathbf{x}_j$ is a $k$-NearestNeighbor of $\mathbf{x}_i\}$ and $\mathbf{w}_i$ is the vector holding the reconstruction weights. After the weights are determined the counterpart $\mathbf{y}_i$ of $\mathbf{x}_i$ in the embedding space can be calculated by

$$\mathbf{y}_i = \sum_{j \in N_i} w_i^j \cdot \mathbf{y}_j, \tag{2}$$

with the $\mathbf{y}_j$'s being the $d$-dimensional embedding counterparts of the $\mathbf{x}_j$'s. Naturally $d < D$ but in general $d \ll D$. Reconstructing a $\mathbf{x}_i$ from a $\mathbf{y}_i$ works in an analogous way. The neighbors $N_i$ of a data point are chosen only among those data points whose mapping is known, namely the data points that were used for the nonlinear dimensionality reduction.

If one demands the reconstruction weights to be larger than zero and summing up to one, then the reconstructed points always lie on the manifold. The reason is that this corresponds to a convex combination whose result is constrained to lie in the convex hull of the support points. Together with the local linear assumption this leads to reconstruction results where the reconstructed points always lie on the manifold. So even if a point beyond the manifold is predicted the mapping by reconstruction ensures that only valid views of the object are generated because it inherently projects the point onto the manifold. This can be seen in Fig. 4 a).

In [11] it has been shown that the energy function (1) can be rewritten as a system of linear equations. This enables to directly calculate the weights using matrix operations. Although the calculated weights are constrained to sum up to one they are not constrained to be positive. This is a problem as it violates the convex combination criteria and hence it is not ensured that a reconstructed point lies on the manifold. However, an iterative repetition of the reconstruction, i.e. reconstructing the reconstructed point, projects the reconstructed point onto the manifold. During this process the weights converge to positive values. Figure 4 b) depicts an example.

## 4 View Prediction

Embedding a set of views of an object into a low dimensional space leads to tuples ($\mathbf{x}_i$,$\mathbf{y}_i$) of views $\mathbf{x}_i$ in the appearance space and their low dimensional counterparts $\mathbf{y}_i$. With this representation of the object's appearance the process of view prediction is as follows:

1) At each time step $t$ the current view $\mathbf{x}_t$ of the object is provided e.g. from a tracking or a detection stage.
2) Determine the $k$ nearest-Neighbors among the represented views.
   $N_t = \{i | \mathbf{x}_i \text{ is a } k\text{-NearestNeighbor of } \mathbf{x}_t\}$
3) Calculate the reconstruction weights $\mathbf{w}_t$ in the appearance space.
   $\hat{\mathbf{w}}_t = \arg\min_{\mathbf{w}_t} \left| \mathbf{x}_t - \sum_{i \in N_t} w_t^i \cdot \mathbf{x}_i \right|^2 , \quad \sum_{i \in N_t} w_t^i = 1$
4) Calculate the mapping to the embedding space by reconstructing the low dimensional counterpart of view $\mathbf{x}_t$.
   $\mathbf{y}_t = \sum_{i \in N_t} \hat{w}_t^i \cdot \mathbf{y}_i$
5) Predict the next position in the low dimensional embedding space, e.g. using the last two views.
   $\mathbf{y}_{t-1}, \mathbf{y}_t \rightarrow \mathbf{y}_{t+1}^{\text{pred}}$
6) Determine the reconstruction weights $\mathbf{w}_a$ in the embedding space by iterative reconstruction.
   Set $\mathbf{y}_a = \mathbf{y}_{t+1}^{\text{pred}}$ and repeat the following steps:
   (i) $N_a = \{i | \mathbf{y}_i \text{ is a } k\text{-NearestNeighbor of } \mathbf{y}_a\}$
   (ii) $\hat{\mathbf{w}}_a = \arg\min_{\mathbf{w}_a} \left| \mathbf{y}_a - \sum_{i \in N_a} w_a^i \cdot \mathbf{y}_i \right|^2 , \quad \sum_{i \in N_a} w_a^i = 1$
   (iii) $\mathbf{y}_a = \sum_{i \in N_a} \hat{w}_a^i \cdot \mathbf{y}_i$

7) Map back to the appearance space.
$\mathbf{x}_{t+1}^{\text{pred}} = \sum_{i \in N_a} \hat{w}_a^i \cdot \mathbf{x}_i$

As explained in the last section, the iterative reconstruction assures that only valid object views are generated. We denote this procedure *embedding view prediction*.

## 5 Experiments

In order to analyze the *embedding view predictor* we conducted some experiments where we compared this view predictor with two view predictors working directly in the high dimensional appearance space.

The first predictor predicts linearly the next view directly in the appearance space from the last two views. In general, this predicted view will lie beyond the manifold of the views. In order to be comparable to the embedding view predictor, the nearest neighbor of the linearly predicted view is determined and returned as the actual predicted view. We denote this view predictor the *nearest neighbor view predictor*.

The second view predictor works like our embedding view predictor but in contrast to this it works directly in the high dimensional appearance space. This means that it linearly predicts views in the appearance space and projects the predicted views onto the appearance manifold using the iterative reconstruction idea. We denote this view predictor the *iterative reconstruction view predictor*.

To validate our view prediction we generated two trajectories in the appearance space for each object. The trajectories are depicted exemplarily with the views of the bird in Fig. 5. It can be seen that the views of the trajectory do not correspond to already represented views in the set of sampled views as introduced in Sect. 2.

The tests we conducted surveyed only the prediction ability of the embedding view predictor compared to the other two view predictors. The view predictors had to predict the views along the discussed trajectories. Thereto each view is predicted using its two predecessors in the trajectory. The predicted views are compared with the actual next views by means of a sum of absolute difference.

Figure 6 shows the prediction error of the three view predictors applied on the two trajectories rotate and whirl (see Fig. 5) of the bee and the bird. It can be observed that the prediction in the low dimensional space is comparable to the predictors operating directly in the high dimensional appearance space. In general, the embedding view predictor is even slightly better. Sometimes, however, it tends to predict views with a large error which appear as single high peaks in the error curve. A closer look revealed that this may be due to topological defects of the embedded appearance manifolds. The strong peaks occur more often when predicting the bird than the bee and indeed the embedding of the bird's appearance manifold is more distorted than that of the bee (see Fig. 3).

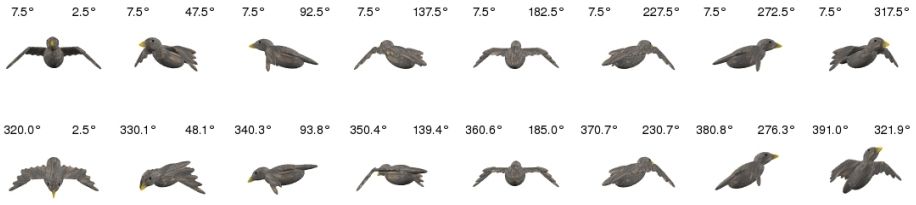


**Fig. 5.** From left to right the views in the two rows show the two variants of trajectories, the view predictors are tested with. The upper is a simple rotation about the vertical axis. The lower starts at 320° horizontal and 2.5° vertical and goes straight to 40° horizontal and 360° vertical and consist of 72 equally distributed views. In order to distinguish between these two trajectories, the first is called "rotate" and the second "whirl". The degrees in the top left corners of the images denote the horizontal rotation and in the top right corners the vertical rotation.

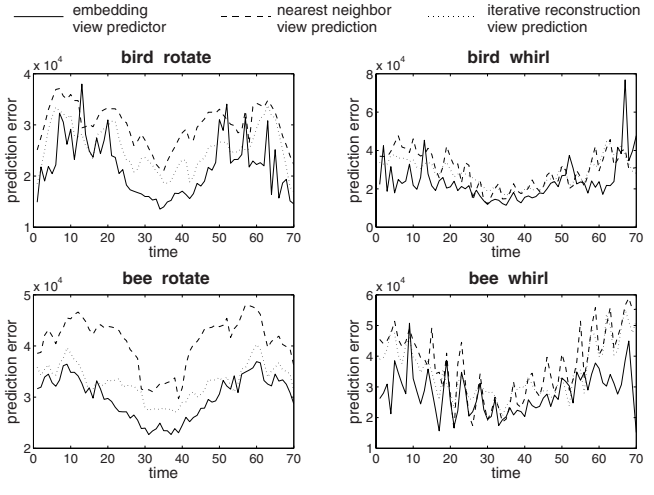


**Fig. 6.** This figure displays the prediction error of the embedding view predictor, nearest neighbor view predictor and iterative reconstruction view predictor for the two trajectories rotate and whirl of the bird and the bee. The error is a sum of absolute difference between the predicted and the actual view. Almost all time the embedding view predictor is superior to the other two.

Furthermore, we analyzed the three predictors concerning their ability to predict further views without being updated with actual views, i.e. we simulated an occlusion of the objects. To this end the three view predictors were again applied to the whirl and rotate trajectories but this time they had to rely solely on their own prediction from the 10th time step on. The results are shown in Fig. 7. It strikes that the embedding view predictor is able to reliably predict up to 10 further views while the other two predictors are only able to predict 2-3 further views. A possible explanation could be the higher ambiguity in the high

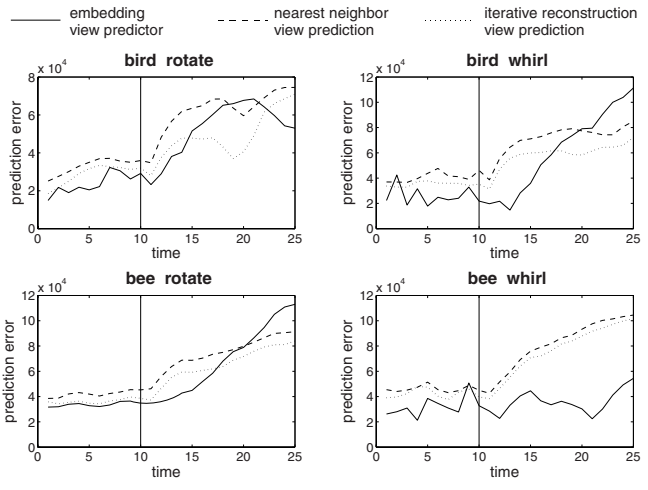


**Fig. 7.** This figure shows the prediction error of the three view predictors applied to the two trajectories rotate and whirl of the bird and the bee. From the 10th time step (view) on the objects are considered to be completely occluded. This means that the predictors have to rely entirely on their own prediction. It can be observed that the embedding predictor can reliably predict up to 10 further views while the other two predictors cannot predict more than two to three views.

dimensional appearance space. This is a hint that predicting on the embedding of the appearance manifold in a low dimensional space is more appropriate for tracking the appearance of objects than predicting directly in the high dimensional appearance space.

## 6 Conclusion

We introduced an approach for predicting views of an object by means of its appearance manifold. By applying Isomap to the various views of an object the appearance manifold of that object can be extracted and embedded into a lower dimensional space. A change of object appearance corresponds to a trajectory on the appearance manifold as well as its embedding. By keeping track of the position of the object on the embedded manifold it is possible to forecast the upcoming appearance. We used an iterative version of the reconstruction idea of LLE in order to map points from the embedding space back into the appearance space and showed that this maps points from the embedding space only to points on the appearance manifold, i.e. only valid views of the object are predicted. Simulations have shown that following the trajectory (and by doing so predicting views) is less error prone using the embedded manifold than its high dimensional equivalent. Furthermore, we have shown that predicting the appearance for several following time steps is also more accurate using the low dimensional embedding. We want to stress that the introduced approach is

no full-fledged real object tracking system but rather a scheme for predicting complex views.

In future work we want to investigate the possibility of using the simplex method for calculating the reconstruction weights as it implicitly constraints the weights to a convex combination. Furthermore, we want to analyze our approach with real objects and integrate it into a tracking architecture based on a view prediction and confirmation model, hopefully boosting the performance of the tracker strongly.

## References

1. Poggio, T., Edelman, S.: A network that learns to recognize three-dimensional objects. Nature 343, 263–266 (1990)
2. Edelman, S., Buelthoff, H.: Orientation dependence in the recognition of familiar and novel views of 3D objects. Vision Research 32, 2385–2400 (1992)
3. Ullman, S.: Aligning pictorial descriptions: An approach to object recognition. Cognition 32(3), 193–254 (1989)
4. Morency, L.P., Rahimi, A., Darrell, T.: Adaptive View-Based Appearance Models. In: Proceedings of CVPR 2003, vol. 1, pp. 803–812 (2003)
5. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by Locally Linear Embedding. Science 290(5500), 2323–2326 (2000)
6. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science 290(5500), 2319–2323 (2000)
7. Zhang, Z., Zha, H.: Principal Manifolds and Nonlinear Dimensionality Reduction via Tangent Space Alignment. SIAM J. Sci. Comput. 26(1), 313–338 (2004)
8. Elgammal, A., Lee, C.S.: Inferring 3D Body Pose from Silhouettes Using Activity Manifold Learning. In: Proceedings of CVPR 2004, vol. 2, pp. 681–688 (2004)
9. Lim, H., Camps, O.I., Sznaier, M., Morariu, V.I.: Dynamic Appearance Modeling for Human Tracking. In: Proceedings of CVPR 2006, pp. 751–757 (2006)
10. Liu, C.B., et al.: Object Tracking Using Globally Coordinated Nonlinear Manifolds. In: Proceedings of ICPR 2006, pp. 844–847 (2006)
11. Saul, L.K., Roweis, S.T.: Think globally, fit locally: unsupervised learning of low dimensional manifolds. Journal of Machine Learning Research 4, 119–155 (2003)