Pose-Invariant Face Recognition in Videos for Human-Machine Interaction

Bogdan Raducanu[1] and Fadi Dornaika[2,3]

[1] Computer Vision Center, 08193 Bellaterra, Barcelona, Spain
bogdan@cvc.uab.es
[2] University of the Basque Country UPV/EHU, San Sebastian, Spain
[3] IKERBASQUE, Basque Foundation for Science, Bilbao, Spain
fadi_dornaika@ehu.es

Abstract. Human-machine interaction is a hot topic nowadays in the communities of computer vision and robotics. In this context, face recognition algorithms (used as primary cue for a person's identity assessment) work well under controlled conditions but degrade significantly when tested in real-world environments. This is mostly due to the difficulty of simultaneously handling variations in illumination, pose, and occlusions. In this paper, we propose a novel approach for robust pose-invariant face recognition for human-robot interaction based on the real-time fitting of a 3D deformable model to input images taken from video sequences. More concrete, our approach generates a rectified face image irrespective with the actual head-pose orientation. Experimental results performed on Honda video database, using several manifold learning techniques, show a distinct advantage of the proposed method over the standard 2D appearance-based snapshot approach.

1 Introduction

In the field of human-machine interaction, faces play a major role. For instance, socially oriented robots are specifically designed to support richer forms of interactions with humans. Their primary mission is to detect human presence, engage in an interaction and behave in a personalized manner. State-of-the-art face recognition techniques can achieve very high accuracy rates under controlled conditions. However, most of current face recognition systems lack robustness in uncontrolled environments (e.g., outdoor scenarios, homes, offices, etc.), since they are pretty sensitive to pose, lighting, occlusions and other variations (such as the presence of natural or artificial structures: beards, moustaches, glasses, etc.). Hence, the challenge is two-fold: to discriminate between different persons and at the same time to be able to recognize the same person affected by one or several of the aforementioned transformations. In particular, head pose problem has been one of the bottlenecks for most current face recognition techniques, because it changes significantly a person's appearance. In order to generalize the use of robots in the context of human-machine interaction, it is mandatory to increase the robustness of the face recognition approach to correctly identify a person showing an arbitrary head pose.

In the robotics context, the system must continuously deal with an incoming flow of face images and has to guarantee a temporal coherence of a person's identity during the

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part II, LNCS 7584, pp. 566–575, 2012.

whole duration of the interaction process. In many cases, the difficulty arises from the fact that there is only a small time frame to capture a face with a high probability that the grabbed images do not contain the required frontal face images. On the other hand, videos very often provide non-frontal faces. To this end two categories of approaches were proposed. The first category uses manifold learning paradigms [1,2] in which the face subspace is constructed using many examples depicting subjects in different poses. The second category generates frontal face from the input image and then apply classic face recognition methods on the reconstructed frontal face image. The second category can be split into two main kinds of approaches: i) 3D morphable models [3], and ii) View-based methods [4]. View-based methods train a set of 2D models, each of which is designed to cope with shape or texture variation within a small range of viewpoints.

In [5], the authors propose a Local Linear regression method for pose invariant face recognition. The proposed method can generate the virtual frontal view from a given non-frontal face image. The whole non-frontal face image is partitioned into multiple local patches and then linear regression is applied to each patch for the prediction of its virtual frontal patch. The method requires the pose of the non-frontal pose as input in order to predict the frontal face. Following the approach of Active Appearance Models, [6] develops a face model and a rotation model which can be used to interpret facial features and synthesize realistic frontal face images when given a single novel face image. In [7], the authors address the non-frontal face recognition using morphable models. The morphable model serves as a preprocessing step by estimating the 3D shape of novel faces from the non-frontal input images, and generating frontal views of the reconstructed faces at a standard illumination using 3D computer graphics. The transformed images are then fed into state of-the-art face recognition systems that are optimized for frontal views. In [8], the authors use view-based active appearance models to fit to a novel face image under a random pose. The model parameters (combined appearance parameters) are adjusted to correct for the pose and used to reconstruct the face under a novel pose. This preprocessing makes face recognition more robust with respect to variations in the pose.

As can be seen, 3D based and view based methods have many limitations. For instance, the 3D morphable models require 3D scans and have high computational load. Although the Active Appearance Models are faster than the 3D morphable models, they require very tedious learning procedure and may generate very low quality images due to the fact the face textures are limited to offline learned statistical face texture models.

In this paper, we propose a strategy that generates virtual view of rectified faces taken from video sequences. We show that this way, we obtain a significant increase in face recognition rate when compared with standard 2D appearance-based snapshot approach. We also show that by adopting rectified faces simple linear manifold learning techniques (e.g., Principal Component Analysis) can provide very good results for face recognition. Our novel scheme is based on fitting a 3D mesh face representation to input images, which efficiently generates rectified facial view from an arbitrary head pose. Thus, our proposed approach is pose- and expression invariant. The proposed technique has the following properties:

- It compensates for all six degrees of freedom of the face pose.
- It compensates for the facial expression and/or facial actions since the rectified face image (geometrically normalized image) corresponds to a neutral face.
- The technique runs in real-time meaning that rectified faces can be provided online from a camera, which is a crucial constraint in human-robot interaction. It uses online appearance models in order to fit a generic 3D model to input images. It is much more practical then 3D morphable models.

The remainder of the paper is structured as follows: in Section 2, we provide the proposed pose and expression compensation based on a deformable 3D face model. The experimental results on face recognition are presented in Section 3. Finally, we provide some concluding remarks and guidelines for future work in Section 4.

2 Generating Rectified Faces in Continuous Videos

In this section, we present the main stages used for generating rectified face textures from videos. Firstly, we describe the 3D deformable model used in the fitting process. Secondly, we describe the process of generating the rectified face assuming that the fitting parameters are correctly estimated. Thirdly, we sketch out the fitting process.

A Deformable 3D Wireframe Model. Building a generic 3D model of a face is a challenging task. Indeed, such a model should account for the differences between specific human faces as well as between different facial expressions. This modelling was explored in the computer graphics, computer vision and model-based image coding communities. In our study, we use the 3D face model *Candide* [9]. This 3D deformable wire-frame model was first developed for the purpose of model-based image coding. The 3D shape of this model is directly recorded in coordinate form, i.e., the 3D coordinates of the vertices. The theoretical 3D face model is given by the 3D coordinates of the vertices $\mathbf{P}_i, i = 1, \ldots, n$ where n is the number of vertices. Thus, the shape up to a global scale can be fully described by the $3n$-vector $\mathbf{g}$ – the concatenation of the 3D coordinates of all vertices $\mathbf{P}_i$. The vector $\mathbf{g}$ can be written as:

$$\mathbf{g} = \overline{\mathbf{g}} + \mathbf{S}\,\boldsymbol{\tau}_{\mathbf{s}} + \mathbf{A}\,\boldsymbol{\tau}_{\mathbf{a}} \tag{1}$$

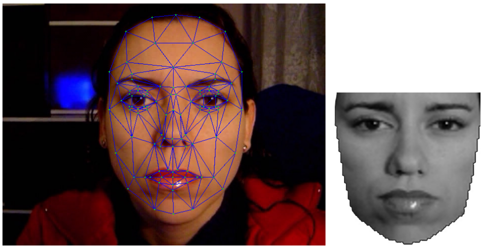
where $\overline{\mathbf{g}}$ is the standard shape of the model, and the columns of $\mathbf{S}$ and $\mathbf{A}$ are the shape and action units, respectively. A shape unit provides a way to deform the 3D wire-frame such as to adapt the eye width, the head width, the eye separation distance, etc. Thus, the term $\mathbf{S}\,\boldsymbol{\tau}_{\mathbf{s}}$ accounts for shape variability (inter-person variability) while the term $\mathbf{A}\,\boldsymbol{\tau}_{\mathbf{a}}$ accounts for the facial action (intra-person variability). The shape and action variabilities can be approximated well enough for practical purposes by this linear relation. Also, we assume that the two kinds of variability are independent. In this study, we use 12 modes for the shape unit matrix and six modes for the action units matrix.

In Equation (1), the 3D coordinates are expressed in a local coordinate system. However, one should relate the 3D coordinates to the image coordinate system (the 2D image coordinates). To this end, we adopt the weak perspective projection model [10]. We neglect the perspective effects since the depth variation of the face can be considered as small compared to its absolute depth[1].

For a given person, $\boldsymbol{\tau}_{\mathbf{s}}$ is constant. Estimating $\boldsymbol{\tau}_{\mathbf{s}}$ can be carried out using either feature-based [11] or featureless approaches [9]. In our recent work, we have shown that some components of the shape control vector can be automatically initialized with a featureless approach [12].

The state of the 3D model is given by the 3D head pose (three rotations and three translations) and the control vector $\boldsymbol{\tau}_{\mathbf{a}}$. This is given by the vector $\mathbf{b}$:

$$\mathbf{b} = [\theta_x, \theta_y, \theta_z, t_x, t_y, t_z, {\boldsymbol{\tau}_{\mathbf{a}}}^T]^T \tag{2}$$



(a) (b)

Fig. 1. Rectified face image based on model fitting. **(a)** an input image with correct fitting. **(b)** the corresponding rectified facial image.

Rectified Facial Texture. Our goal is to generate a rectified facial texture, as shown in figure 1.b, which compensates for the 3D head pose (3 rotational degrees of freedom and the 3 translational degrees of freedom) as well as for the facial expression and/or facial actions. In this Section, we briefly describe how this rectified texture is computed from the input image. In the sequel, this rectified texture will be used for face recognition.

The basic idea is to compute a geometrically normalized face image (rectified face) from the input image by fitting a deformable 3D mesh to the input image.

The 2D mesh associated with the rectified texture is obtained by projecting the standard shape $\overline{\mathbf{g}}$ (wire-frame), using a centered frontal 3D pose, onto an image with a given resolution. The texture of the rectified facial image is obtained by texture mapping from the triangular 2D mesh covering the face in the input image (see figure 1.a) using a piece-wise affine transform, $\mathcal{W}$. Similarly to [13], we have taken advantage of the fact that the barycentric coordinates of the pixels within each triangle are invariant under affine transforms. In other words, since the geometry of the 2D mesh in the rectified image is fixed the barycentric coordinates are fixed and can be computed once for all, which considerably reduces the CPU time associated with the texture mapping process - the warping process.

[1] The perspective projection is the classical pin-hole camera model. The weak perspective projection can be seen as the zero approximation to the perspective projection.

Once an instance of the 3D model (encoded by the vector **b**) is projected onto the input image, the warping process proceeds as follows. The rectified image bounding the fixed 2D mesh is scanned pixel by pixel. For every scanned pixel in this image, we know its triangle as well as its barycentric coordinates within this triangle. Therefore, the 2D location of the corresponding pixel in the input image can be easily inferred using a linear combination of the coordinates of the triangle vertices where the coefficients are given by the barycentric coordinates. The greylevel of the scanned pixel is then set by blending the greylevels associated with the four closest pixels to the non-integer coordinates of the returned location - the bilinear interpolation.

Mathematically, the warping process applied to an input image **y** is denoted by:

$$\mathbf{x}(\mathbf{b}) = \mathcal{W}(\mathbf{y}, \mathbf{b}) \tag{3}$$

where **x** denotes the rectified facial texture and **b** denotes the geometrical parameters. Without loss of generality, we have used two resolution levels for the rectified textures, encoded by 1310 and 5392 facial pixels bounded by 40× 42 and 80×84 rectangular boxes, respectively. Obviously, other levels can be used. Generally speaking as the resolution increases, the fitting accuracy increases. However, by experience we found that the second resolution level is a good trade-off between the accuracy and the computational cost. In the sequel, all results are obtained with this resolution. The fitting and face rectification processes for the 80×84 rectified face image took about 40 ms on a PC equipped with a dual-core Intel processor at 2 Ghz.

Model Fitting: 3D Head Pose and Facial Action Estimation. In the previous section, we have shown that if the 3D deformable model is fitted to the input image, i.e., the geometric parameters **b** are estimated, then generating the rectified face image will be straightforward. In this section, we describe how these parameters are estimated in videos sequences using the temporal tracker developed in [14]. The basic idea is to recover **b** by minimizing a distance between the incoming warped frame and the current appearance of the face. This minimization is carried out using a difference decomposition-like approach [15]. This tracker has two interesting features. First, the statistics of the appearance model are updated online. Second, the empirical gradient matrix is computed for each input frame. This scheme leads to a fast, efficient and robust tracking algorithm. Figure 2 displays the tracking results associated with eight frames of a 750-frame sequence featuring quite large pose variations as well as large facial actions. The sequence is of resolution 720 × 480 pixels.

Pose Invariant Face Recognition. For every input frame, the above tracker provides the geometrical parameters (3D pose and facial actions) of the 3D model. It also provides the rectified face image associated with the input image. This rectified facial image corresponds to a frontal and neutral face. Thus, unlike existing methods, our proposed method simultaneously compensates for 3D pose and facial expression.

Once non-frontal face images are converted to virtual views, face recognition can be easily achieved by using the virtual generated views instead of all the non-frontal face images. The proposed method can be regarded as a preprocessing procedure independent of the following feature extraction and classifier design. Therefore, the proposed method can be combined with any face recognition technologies.

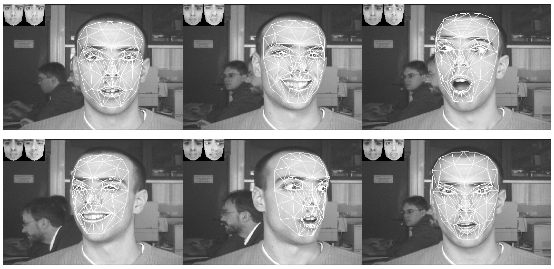


Fig. 2. Face and facial action tracking results using our appearance-based tracker. The snapshots correspond to frames 63, 181, 282, 418, 492, and 683, respectively in a 750-frame video sequence.

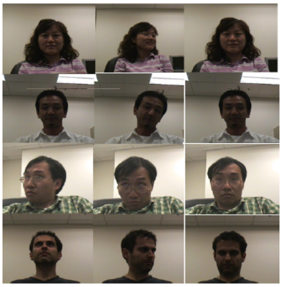


Fig. 3. Some samples from the Honda Video database

3 Experimental Results

Data Preparation. Our approach has been tested on the Honda video database (HVDB) [16,17]. HVDB has been acquired for the purpose of face tracking and recognition. It depicts persons sitting in front of a camera in a totally uncontrolled environment and performing unconstrained in-plane and out-of-plane head motion. Some samples are depicted in figure 3. The resolution of the images is 640x480 pixels and the videos were recorded at 15 frames per second. We selected from this database a subset of 22 video clips belonging to 22 different persons[2]. The rectified face texture was extracted according to the approach presented in the section 2. Some examples of rectified faces, under arbitrary head poses, are presented in figure 4. The iconic images that appear in the upper left corner of the image on the left column have the following meaning: the left represents the temporal average of the facial texture (used by the Online Appearance Model) and the right one, the rectified face (a higher resolution version is presented on the right column). In the end, we collected two datasets: one containing the rectified faces and the other, the cropped ones. Each dataset contains 2317 images organized in 22 classes, with an average of 100 images per class. Both rectified and cropped images were resized to 50×50 pixels.

[2] To guarantee optimal results from the Candide model, we limited the pan/tilt variation in head pose to the interval [-45°,45°].

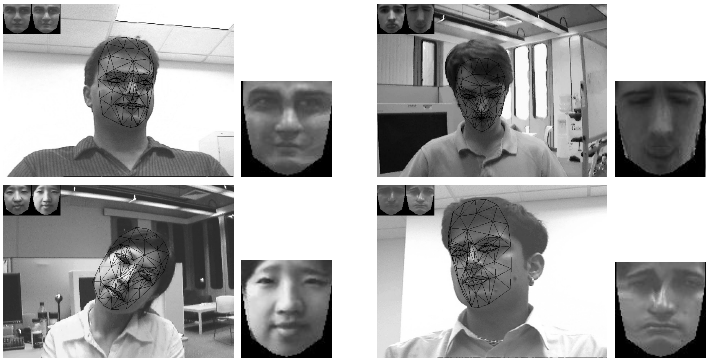


Fig. 4. Some samples of rectified face images from the Honda Video database based on model fitting. **Left column:** the input image with correct fitting. **Right column** the corresponding rectified facial image.

Face Recognition Results. For face recognition evaluation, we applied a manifold representation approach. This implies that the original, high-dimension data, are embedded into a low-dimensional subspace, without any relevant loss of information. The projection function could be either linear or non-linear. More concrete, we used the following manifold learning techniques: Principal Component Analysis (PCA), Locally Linear Embedding (LLE) [18], Locality Preserving Projections (LPP) [19] and Laplacian Eigenmaps (LE) [20]. PCA and LPP belong to the category of linear embedding techniques, while LLE and LE are non-linear. The PCA method estimates orthogonal projection directions that maximize the variance of original data. The LPP method searches a linear projection that preserve the locality of the neighboring data. The LLE algorithm seeks the nonlinear embedding in a neighborhood preserving manner by exploiting the local symmetries of linear reconstructions, and seeking the optimal weights for local reconstruction. The embedding is obtained using the estimated local weights. The LE method seeks the nonlinear embedding by preserving locality. It should be noticed that the LPP method is the linearized version of LE.

For classification purposes, we adopted 10 random splits of data. We split the data in several ratios for training and test: $10\% - 90\%$, $20\% - 80\%$, $30\% - 70\%$, $40\% - 60\%$ and $50\% - 50\%$. The reason we decided to start with such a low percentage of training images (10% is motivated by the fact that we have a pretty high number of instances per class. The classification in the embedded space has been carried out using the Nearest Neighbor (NN) approach. The recognition rates for rectified images and cropped ones are reported in the tables 1 and 2, respectively. From these tables, we can observe that: (i) the use of rectified faces has improved the face recognition rate for all manifold learning techniques used and for all train/test ratios; (ii) the PCA technique has provided the best recognition results for both the cropped and rectified faces; (iii) the LLE method has provided the worst results. This is very consistent with the fact that LLE is not very suited for classification task; and (iv) for the rectified face, the LE method provided better results than the LPP method, whereas for the the cropped faces we got in general the reverse.

Table 1. Best average recognition accuracy using rectified faces

Train	10%	20%	30%	40%	50%
PCA	**99.08%**	**99.89%**	**100.00%**	**100.00%**	**100.00%**
LLE	60.64%	75.02%	79.60%	81.95%	85.76%
LPP	85.47%	94.02%	97.04%	98.33%	98.91%
LE	91.35%	97.87%	99.06%	99.45%	99.65%

Table 2. Best average recognition accuracy using cropped faces

Train	10%	20%	30%	40%	50%
PCA	**74.73%**	**85.86%**	**91.67%**	**94.24%**	**95.85%**
LLE	34.18%	42.71%	47.84%	53.99%	64.53%
LPP	67.54%	80.40%	86.36%	89.62%	91.64%
LE	68.82%	77.48%	82.34%	85.52%	87.62%

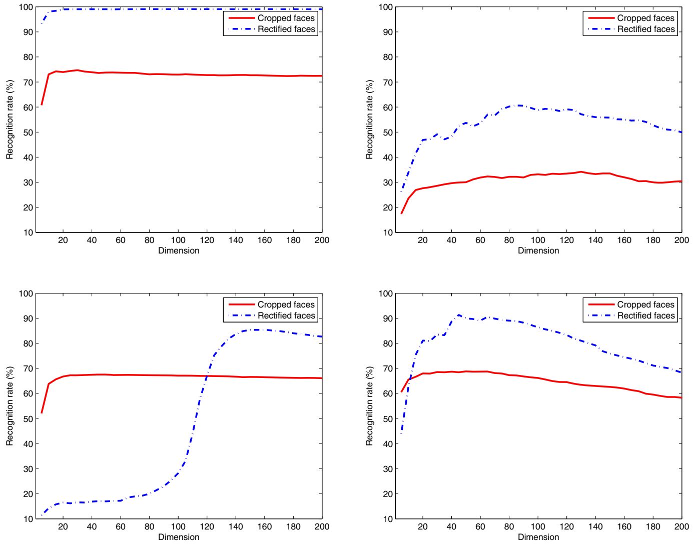


Fig. 5. Comparison of recognition rates between rectified and cropped face images with a 10% ratio of images used for training. The four plots depict the following embedding techniques (left-right and top-bottom order): PCA, LLE, LPP and LE.

In figure 5 we plot a comparison of recognition rates between rectified and cropped face images. The recognition rate is computed for a dimensionality of the embedded space up to 200 (in five-step increments). These plots correspond to the case when 10% of the data is used for training. We can see that the rectified faces have provided better recognition rates than the cropped faces for all dimensions used by PCA, LLE, and LE methods. This holds for LPP method as long as the dimension is above 120. Additionally, we plot in figure 6 a comparison between all the manifold representation

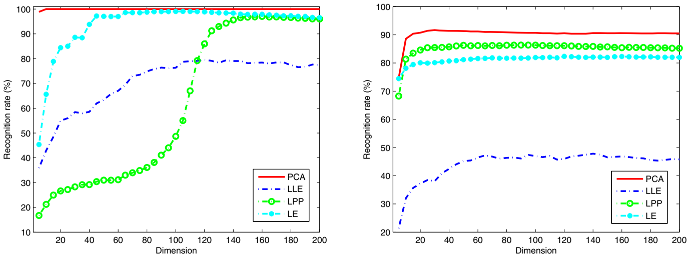


Fig. 6. Comparison of recognition rates between all 4 manifold representation techniques with a 30% ratio of images used for training: rectified faces (top) and cropped faces (bottom)

techniques in the two cases: rectified images and cropped images. This plot corresponds to the case when 30% of the data was used for training. We can observe that PCA-based representation provided the best recognition rates (for both cropped and rectified faces have), even with very few dimensions. This allows us to conclude that simple representation as PCA outperforms more sophisticated manifold representation techniques, which can be seen as an improvement in system's performance.

4 Conclusions and Future Work

In this paper, we proposed a novel approach for robust face recognition in a human-computer interaction scenario. Our method consists in a real time fitting of a 3D deformable model to input images taken from video sequences. More concrete, our approach generates a rectified face image irrespective with the head-pose orientation. This approach is fast and is working in real-time, which is a very important requisite for human-robot interaction. Moreover, experimental results performed on Honda video database, using several manifold learning techniques, show a distinct advantage of the proposed method over the standard 2D appearance-based snapshot approach. Future work will address the problem of video based face recognition using sequences of rectified face images.

Acknowledgements. B. Raducanu is supported by the project TIN2009-14404-C02-00, Ministerio de Educación y Ciencia, Spain.

References

1. Cai, D., He, X., Zhou, K., Han, J., Bao, H.: Locality sensitive discriminant analysis. In: International Joint Conference on Artificial Intelligence, pp. 708–713 (2007)
2. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science 290(5500), 2319–2323 (2000)
3. Blanz, V., Vetter, T.: Face recognition based on fitting a 3D morphable model. IEEE Transactions on PAMI 25(9), 1063–1074 (2003)

4. Cootes, T.F., Wheeler, G.V., Walker, K.N., Taylor, C.J.: View-based active appearance models. Image and Vision Computing 20(9-10), 4165–4176 (2002)
5. Chai, X., Shan, S., Chen, X., Gao, W.: Locally linear regression for pose-invariant face recognition. IEEE Trans. on Image Processing 16(7), 1716–1725 (2007)
6. Shan, T., Lovell, B.C., Chen, S.: Face recognition robust to head pose from one sample image. In: IEEE Intl. Conf. on Face and Gesture Recognition, pp. 515–518 (2006)
7. Blanz, V., Grother, P., Phillips, P.J., Vetter, T.: Face recognition based on frontal views generated from non-frontal images. In: IEEE Intl. Conf. on Computer Vision and Pattern Recognition, pp. 454–461 (2005)
8. Huisman, P., Munster, R., Veldhuis, R., Bazen, A.: Making 2d face recognition more robust using AAMs for pose compensation. In: IEEE International Conference on Face and Gesture Recognition, pp. 113–116 (2006)
9. Ahlberg, J.: An active model for facial feature tracking. EURASIP Journal on Applied Signal Processing 2002(6), 566–571 (2002)
10. Faugeras, O.: Three-Dimensional Computer Vision: a Geometric Viewpoint. The MIT Press (1993)
11. Lu, L., Zhang, Z., Shum, H., Liu, Z., Chen, H.: Model- and exemplar-based robust head pose tracking under occlusion and varying expression. In: Proc. IEEE Workshop on Models versus Exemplars in Computer Vision (CVPR 2001), pp. 1–8 (2001)
12. Dornaika, F., Raducanu, B.: Person-specific face shape estimation under varying head pose from single snapshots. In: IEEE Intl. Conf. on Pattern Recognition, pp. 3496–3499 (2010)
13. Ahlberg, J.: Real-time facial feature tracking using an active model with fast image warping. In: International Workshop on Very Low Bitrate Video (VLBV), Athens, Greece, pp. 39–43 (2001)
14. Dornaika, F., Davoine, F.: On appearance based face and facial action tracking. IEEE Transactions on Circuits and Systems for Video Technology 16(9), 1107–1124 (2006)
15. Gleicher, M.: Projective registration with difference decomposition. In: Proc. of Intl. Conf. on Computer Vision and Pattern Recognition, pp. 331–337 (1997)
16. Lee, K.-C., Ho, J., Yang, M.-H., Kriegman, D.: Visual tracking and recognition using probabilistic appearance manifolds. Computer Vision and Image Understanding 99, 303–331 (2005)
17. Lee, K.-C., Kriegman, D.: Online learning of probabilistic appearance manifolds for video-based recognition and tracking. In: Proc. of Intl. Conf. on Computer Vision and Pattern Recognition, pp. 852–859 (2005)
18. Roweis, S., Saul, L.: Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500), 2323–2326 (2000)
19. He, X., Niyogi, P.: Locality preserving projections. In: Conference on Advances in Neural Information Processing Systems (2003)
20. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation 15(6), 1373–1396 (2003)