Single Snapshot-Based 3D Head Pose Initialization for Tracking in a HRI Scenario

F. Dornaika [1,2]
[1] University of the Basque Country
20018 San Sebastian, Spain
[2] IKERBASQUE, Basque Foundation for Science
48011 Bilbao, Spain
fadi_dornaika@ehu.es

B. Raducanu
Computer Vision Center
08193 Bellaterra, Barcelona, Spain
bogdan@cvc.uab.es

Abstract

This paper presents an automatic 3D head pose initialization scheme for a real-time face tracker with application to human-robot interaction. It has two main contributions. First, we propose an automatic 3D head pose and person specific face shape estimation, based on a 3D deformable model. The proposed approach serves to initialize our real-time 3D face tracker. What makes this contribution very attractive is that the initialization step can cope with faces under arbitrary pose, so it is not limited only to near-frontal views. Second, the previous framework is used to develop an application in which the orientation of an AIBO's camera can be controlled through the imitation of user's head pose. In our scenario, this application is used to build panoramic images from overlapping snapshots. Experiments on real videos confirm the robustness and usefulness of the proposed methods.

1. Introduction

The mainstream for 3D head pose and person-specific face shape parameters estimation in video sequences relies on extracting and matching some salient facial features such as the locations and local statistics of the eyes, nose, and mouth in one or more views. A taxonomy of head pose estimation approaches can be found in [14]. Feature-based approaches suffer from self-occlusions and drifting [21, 20]. A solution to overcome the drawbacks of feature-based approaches is given by holistic approaches (appearance-based approaches), which try to analyze the whole facial appearance [4, 11]. For example, Active Appearance Models (AAMs) were mainly used for 2D model fitting and tracking.

Model-based applications exploiting monocular vision systems (the face model is given by a 3D mesh or a range model) need to personalize the face model of the person utilizing the system in order to achieve an accurate estimation. This holds true even with simple 3D models such as cylinders and ellipsoids. Many works have addressed the 3D face tracking in video sequences [22]. However, most of them rely on a manual initialization of the parameters, in the sense that the tracking parameters are provided by the user. Obviously, real world human computer interaction require a full automatic 3D face tracker. Recently many authors used special sensors such as a travelling camera or a 3D scanner in order to build personalized facial shape [2]. These shape models are then used for art production or for 3D face detection and recognition using 3D sensors. Such systems suffer from several shortcomings. Some of the shortcomings can be alleviated by using stereo vision sensors [8]. In [10], the authors propose to infer side-view shape parameters from one single frontal image using learned statistical correlation between the frontal-view parameters and the side-view parameters. The facial points (MPEG-4 points) and the frontal view parameters (relative distances) are extracted from the frontal image using some heuristics and prior knowledge.

In this paper, we present a novel approach for the automatic 3D head pose estimation and tracking from monocular videos. More concretely, our work presents two contributions. First, we estimate the 3D head pose and person specific face shape parameters from single images. For this purpose, we use a statistical facial texture model and a standard deformable 3D model. This technique is used to initialize a real-time tracker [6]. The proposed approach presents several characteristics that make it very attractive: (i) the initialization step copes with faces under arbitrary pose, so it is not limited only to near-frontal views, (ii) it is person-independent, (iii) it is featureless (no facial features are needed), and (iv) its learning stage is simple. Second, based on this framework, we developed an application that shows how to control an AIBO head movement by imitat-

ing the automatic estimated user's head pose. In our scenario, this application is used to build panoramic images from overlapping snapshots. The whole process described in this paragraph is graphically depicted in figure 1. This work builds on our previous works [6, 7]. More precisely, [6] describes our developed real-time 3D face tracker where the tracking is initialized manually. [7] evaluates the accuracy of 3D head pose parameters obtained by this tracker using dense 3D data. Our current work provides two main contributions: i) automatic initialization of the parameters from one single image, and ii) an application in which the robot camera is controlled using the tracked face orientation.

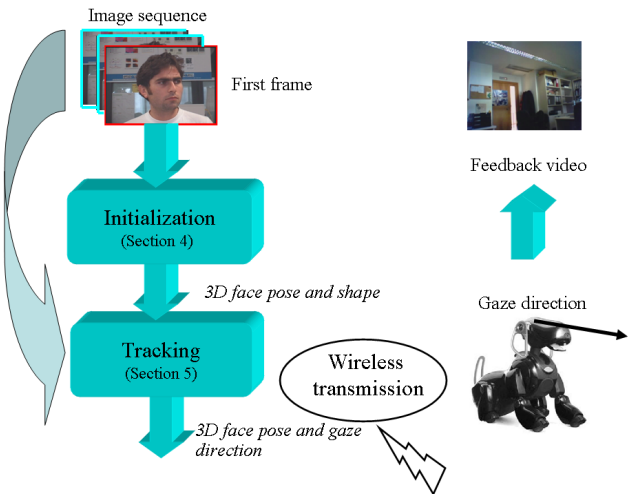


Figure 1. Process flow for a user's gaze imitation using a monocular camera. The user's automatically estimated head pose is continuously controlling the AIBO's head orientation and thus registering the scene.

The proposed holistic approach estimates the 3D pose parameters, as well as the person specific face parameters, by registering the input texture (warped region of the image) to a statistical face texture. Compared to AAMs methods our proposal has two advantages. First, there is no need to compute a Jacobian matrix neither off-line nor online. Second, while AAMs merge both the inter and intra-person shape variabilities, our method separates these variabilities, and therefore the proposed method can be easily and efficiently used for initializing a real time 3D face tracker and facial expression recognizer in videos (both the 3D deformable model and its 3D pose are computed for the first frame in the video sequence). However, it is not clear how these tasks can be performed with AAMs. We stress the fact that our approach does not use neither 2D AAM nor 3D AAM. The only similarity with AAMs is the use of a statistical facial texture model based on Principal Component Analysis (PCA).

The remainder of the paper is organized as follows. Section 2 reviews some related work on head movement imitation with application to robotics. Section 3 introduces some aspects regarding face modelling. Section 4 presents the proposed holistic approach for the initialization of the 3D head pose and face shape parameters. Section 5 summarizes the 3D face tracker. Section 6 presents some qualitative and quantitative evaluations of performance. In section 7 we present an application for a human-robot interaction scenario, based on the framework previously introduced. Section 8 concludes the paper.

2. Related work on head movement imitation

Imitating gestures, in general, and head pose, in particular, represents an attractive research topic in robotics. The reason is twofold: (i) imitative robots require less programming effort compared to 'classical' ones (which need to be programmed explicitly); (ii) at the same time, imitative robots could be used by cognitive researchers to test computational theories, and to validate psychological experiments.

Head movement imitation represents the basis for joint attention, defined as the behavior to look where someone else is looking. From a psychologic point of view, joint attention is of primary interest to study several aspects related to cognitive development. Early work on head pose imitation was reported in [5]. They built a model of head movement imitation based on the computation of optical flow. Their implementation was inspired by a theoretical framework of the active intermodal mapping proposed by Meltzoff et al. [12]. The same theoretical model was used later on by Shon et al. [18] in order to build a robot endowed with the capability to manifest joint attention. Scassellati [17] built a robot able to imitate head nodding. The head movement was detected by the cumulative displacement of user's face in the robot's field of view. Another work which addresses the problem of joint attention development in a infant-like robot based on head movement imitation is reported in [15].

The application presented in our paper show how the orientation of an AIBO's camera can be controlled through the imitation of user's head pose, in real-time. This application is used to build panoramic images from overlapping snapshots. Due to the low complexity and facility to be reproduced, our approach can be easily adopted by any system whose aim is to study efficiently high-level, cognitive aspects, of the human-robot interaction (like joint attention, for example).

3. Modelling faces

A deformable 3D mesh In our work, we used *Candide* 3D face model [1]. This common 3D deformable wireframe model accounts for person specific shape variation as well as for facial animation. The 3D shape of this wireframe

model (triangular mesh) is directly recorded in coordinate form. It is given by the coordinates of its n 3D 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. The vector $\mathbf{g}$ is written as:

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

where $\overline{\mathbf{g}}$ is the standard shape of the model, $\boldsymbol{\tau}_{\mathbf{s}}$ and $\boldsymbol{\tau}_{\mathbf{a}}$ are shape and animation control vectors, respectively, and the columns of $\mathbf{S}$ and $\mathbf{A}$ are the Shape and Animation Units. A Shape Unit provides a means of deforming the 3D wireframe so as to be able to adapt eye width, head width, eye separation distance, etc (see Figure 2). 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 animation (intra-person variability). With this model, the ideal neutral face configuration is represented by $\boldsymbol{\tau}_{\mathbf{a}} = \mathbf{0}$. In this study, we assume that the images are depicting quasi-neutral faces. Thus, the expression for the deformable mesh becomes:

$$\mathbf{g} = \overline{\mathbf{g}} + \mathbf{S}\,\boldsymbol{\tau}_{\mathbf{s}} \quad (2)$$

The shape modes were created manually to accommodate the subjectively most important changes in facial shape. In the model package, the number of modes associated with facial Shape Units matrix $\mathbf{S}$ (inter-person variability) is twelve. However, for the purpose of our study which deals with the automatic image-based extraction of the control vector $\boldsymbol{\tau}_{\mathbf{s}}$ only six components are considered as the most significant indicators of the perceived person-dependent facial shape in a given near frontal facial image. These components are: Head height, vertical position of the eye brows, vertical position of the eye, eyes separation distance, vertical position of the nose, vertical position of the mouth. The remaining components are set to nominal values.

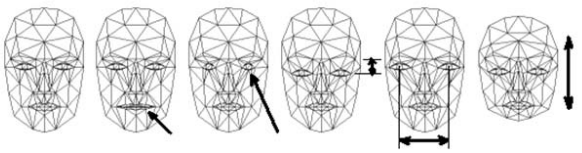


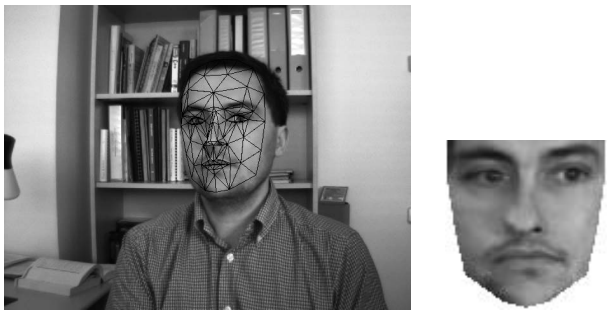
Figure 2. Effects of some facial shape control parameters on the deformable 3D model (standard shape, mouth width, eyes width, eyes vertical position, eye separation distance, head height).

In equation (1), the 3D shape is expressed in a local coordinate system. However, one should relate the 3D coordinates to the image coordinate system. To this end, we should add the six degrees of freedom associated with the 3D face pose. The mapping between the 3D face model and the image adopts the weak perspective projection model.

Thus, the state of the 3D wireframe model is given by the 3D face pose parameters (three rotations and three translations) and the shape control vector $\boldsymbol{\tau}_{\mathbf{s}}$. This is given by the 12-dimensional vector $\mathbf{b}$:

$$\mathbf{b} \;=\; [\theta_x,\;\; \theta_y,\;\; \theta_z,\;\; t_x,\;\; t_y,\;\; t_z,\;\; \boldsymbol{\tau}_{\mathbf{s}}^T\;]^T \quad (3)$$

Shape-free facial patches A facial patch is represented as a shape-free image (geometrically normalized rawbrightness image). The geometry of this image is obtained by projecting the standard shape $\overline{\mathbf{g}}$ using a centered frontal 3D pose onto an image with a given resolution. The texture of this geometrically normalized image is obtained by texture mapping from the triangular 2D mesh in the input image (see figure 3) using a piece-wise affine transform.



(a) **(b)**

Figure 3. **(a)** an input image with correct fitting. **(b)** the corresponding shape-free facial patch.

4. Initializing the 3D head pose and face shape parameters

4.1. Face subspace

The statistical facial texture model describes the appearance variation of the shape-free facial patches $\mathbf{x}$ (see figure 3.**(b)**). These patches are obtained from the training images (individual snapshots or video sequences) by fitting the 3D deformable model to the face. This fitting can be manual or automatic [1]. Using these training patches one can easily build the face subspace. For this purpose we use the Principal Component Analysis (PCA)—a well-known technique used for modelling face subspaces. We assume that we have K shape-free patches. Applying a PCA on the training patches we can compute the mean and the principal modes of variation. The use of linear subspace (PCA) can be justified by 1) the facial images are geometrically normalized, and 2) the proposed method will be carried out in a relatively constrained environment. However, currently we are investigating the use of non-linear manifold learning techniques.

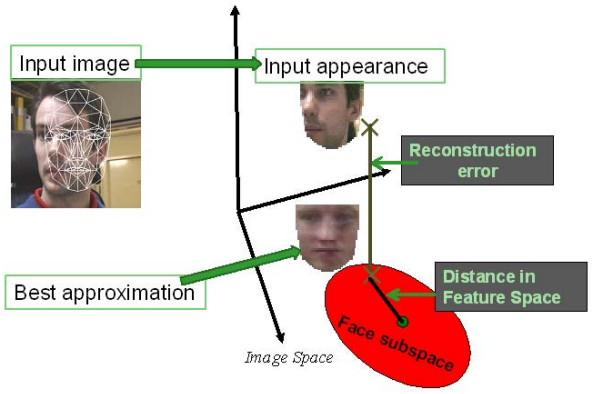


Figure 4. The unknown parameters are estimated by maximizing a likelihood measure taking into account the reconstruction error and the distance in feature space. The face subspace is linear.

4.2. Optimization

The basic idea is to estimate the 3D face pose and shape parameters, i.e. the vector **b**, such that the associated shape-free patch will be as close as possible to the facial sub-space. This can be carried out by maximizing a certain likelihood measure. For this purpose, we use the likelihood measure proposed in [13]:

$$p(\mathbf{x}|\mathbf{b}) \propto \exp\left(-\frac{1}{2}\sum_{i=1}^{M}\frac{\xi_i^2}{\lambda_i}\right)\ \exp\left(-\frac{e}{2\rho^\star}\right) \qquad (4)$$

where e is the reconstruction error, λ_is are the M largest eigenvalues given by the PCA, ξ_is represent the texture projection onto the corresponding M eigenvectors, and $\rho^\star$ is the arithmetic average of the remaining eigenvalues (in the complementary subspace). The reconstruction error is the distance between the original shape-free texture **x** and its projection onto the PCA subspace. The above likelihood measure takes into account two distances (i) the distance-from-feature-space, and (ii) the distance-in-feature-space. These two distances are illustrated in Figure 4. Maximizing this likelihood is equivalent to minimizing the *Mahalanobis* distance over the original textures. The unknown 3D face pose and shape parameters (the vector **b**) can be estimated by seeking the maximum of the likelihood (4):

$$\mathbf{b} = arg\max_{\mathbf{b}} p(\mathbf{x}|\mathbf{b}) \qquad (5)$$

To this end, we use the Differential Evolution (DE) algorithm [16] in order to maximize (4) with respect to the 3D face pose and shape parameters. The DE algorithm is a practical approach to global numerical optimization that is easy to implement, reliable and fast. The crucial idea behind DE is a scheme for generating trial parameter vectors. Basically, DE adds the weighted difference between two population vectors to a third vector. In our case, the initial population is randomly selected between the lower and upper bounds defined for each variable using uniform distributions. The distributions associated with the translational part of the 3D face pose are centered on the output of the 2D face detector [19].

5. 3D face tracking in video sequences

In order to track the face in an arbitrary video sequence we have to estimate the 3D pose face parameters as well as the facial action parameters for every frame in that video. Recall that the 3D face pose and shape parameters associated with the first video frame were estimated by the initialization approach described in the above section. Since the shape parameters $\boldsymbol{\tau}_\mathbf{s}$ are constant for a given video sequence (the static shape of a given subject is not changing over time), these parameters do not to be tracked in the remaining frames. Instead, we need to track the 3D face pose and possibly some facial actions describing local facial motions. For this purpose, we use Equation (1) with $\boldsymbol{\tau}_\mathbf{s}$ being constant. Thus, the 3D deformable model can be described by

$$\mathbf{g} = \mathbf{g}_s + \mathbf{A}\,\boldsymbol{\tau}_\mathbf{a} \qquad (6)$$

where $\mathbf{g}_s$ denotes the static shape of the model. It is estimated by the initialization process described in Section 3 (see Equation (2)). Therefore, the sate vector for every subsequent frame in the video sequence is given by the vector $\mathbf{b}'$

$$\mathbf{b}' = [\theta_x,\ \theta_y,\ \theta_z,\ t_x,\ t_y,\ t_z,\ \boldsymbol{\tau}_\mathbf{a}^T\ ]^T \qquad (7)$$

For fitting every subsequent frame in the video sequence (i.e., estimating the state vector $\mathbf{b}'$), we use the 3D face tracker based on Online Appearance Models [6]. This estimates a full 3D orientation and position of a head, while incorporating additional DOF including movement of the eyebrows and the lips. This appearance-based tracker aims at computing the 3D head pose by minimizing a distance between the incoming warped frame and the current *shape-free* appearance of the face. This optimization is carried out using a gradient descent method. The statistics of the *shape-free* appearance as well as the gradient matrix are updated every frame. This scheme leads to a fast and robust tracking algorithm.

6. Evaluating the initialization

Experiments were conducted to evaluate the performance of the proposed initialization technique in image

snapshots extracted from several video sequences recorded under realistic conditions. More concretely, we recorded 20 videos, belonging to 10 persons, in which people were asked to move their head in an unconstrained manner, covering as many head pose angles as possible. The videos have an average length of 40 seconds and are recorded at 25 frames per second. The distance of the faces from the recording camera ranged from 50 cm to one meter in order to guarantee enough facial image resolution, as required by *Candide* model. The yaw and pitch angles belong to the interval $[-40°, +40°]$.

In this section, we report qualitative and quantitative evaluation of the proposed algorithm.

6.1. Qualitative evaluation

In this section, we present some qualitative results of the initialization scheme proposed in Section 4. Figure 5 illustrates the application of the proposed scheme on three different persons, under varying head poses. The purpose of this figure is to show that the features of the 3D model (deformed 3D mesh) projects correctly onto their corresponding 2D features in the image. Recall that this projection relies on the estimated shape and 3D head pose parameters. One can notice that there is a slight misalignment for the lip corners. This is not due to the algorithm but to the fact that the mouth is not perfectly at its neutral configuration.

Figure 5. Single snapshot-based initialization of the 3D head pose and face shape parameters associated with three different persons.

6.2. Quantitative evaluation

In the previous section, the evaluation of the fitting algorithm was carried out by visual inspection. In this section, we present a quantitative evaluation of the proposed initialization scheme. The aim of our experiments is twofold: to evaluate the accuracy of the estimated 3D head pose over a large number of images, and to evaluate the accuracy of the estimated person-specific shape parameters. Recall that these parameters are constant for a given person. The ground truth for both kinds of parameters (the person-specific shape parameters and the 3D head pose parameters) was established by manually fitting the 3D mesh of the *Candide* model using an interactive graphical interface. In order to show the robustness of our approach, we tested its performance also in the presence of occlusions. We have found that PCA models with 20 principal components are usually enough for representing the face space. More precisely, we found that the retained variance is above 95% of the total variance.

6.2.1 3D head pose accuracy evaluation

We point out that the input to our proposed fitting algorithm is a single snapshot of a face.

In order to test the accuracy of the 3D head pose estimation, we selected about 900 images from our videos. The average errors over the whole data set are summarized in table 1.

$t_x(cm)$	$t_y(cm)$	$t_z(cm)$	θ_x	θ_y	θ_z
0.24	0.26	1.33	4.37°	4.82°	0.73°

Table 1. Deviation in the estimation of the 3D head pose parameters.

The first three values correspond to the 3D spatial location: translation on the horizontal, vertical axis and depth, respectively and the last three values correspond to the three rotation angles: pitch, yaw and roll, respectively. In figure 6, the plots depict the error in estimation of the pitch and yaw for a sequence of 300 snapshots. The error in roll hasn't been represented due to its very small values.

Evaluation in the presence of occlusions Figure 7 shows the fitting results on a partially occluded face. The occlusions affected about half of the shape-free facial image. Despite the presence of these occlusions, the estimated pose and shape parameters do not deviate significantly from their estimated values with no occlusion. Table 2 summarizes the deviations associated with the 3D face pose parameters (Figure 7.**(a)**). These depicted deviations are simply the absolute difference between the estimated parameters in the presence of occlusion and those ones estimated with no occlusion.

$t_x(cm)$	$t_y(cm)$	$t_z(cm)$	θ_x	θ_y	θ_z
0.11	0.2	1.6	1.35°	2.1°	1.64°

Table 2. Deviation in the estimation of the 3D head pose parameters when the face is partially occluded (Figure 7.**(a)**).

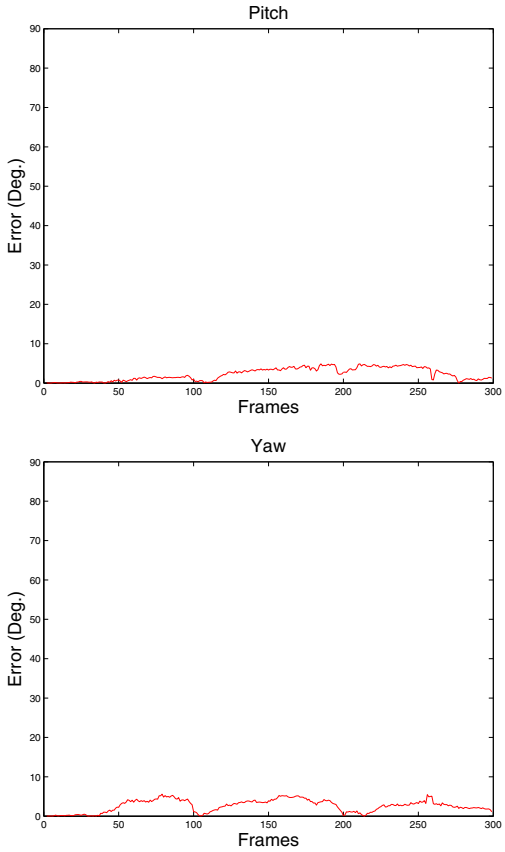
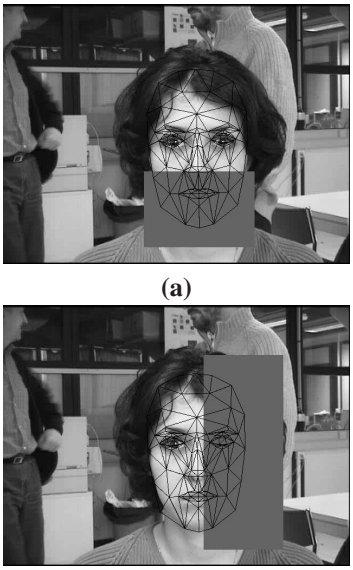


Figure 6. 3D head pose accuracy estimation. The average error for pitch and yaw, respectively



(a)

(b)

Figure 7. 3D face pose and shape estimation when the face is partially occluded.

6.2.2 Person specific face shape parameters evaluation

The error in the estimation of the person specific face shape parameters is reported in terms of the deviation between the estimated shape parameters and the ground-truth values. Recall that the shape parameters control the following features: the vertical position of the eyebrows, the vertical position of the eyes, the eyes separation distance, the vertical position of the nose and the vertical position of the mouth.

Table 3 depicts the average deviation between the ground-truth parameters and the automatically estimated ones over ten different individuals. Recall that the face shape parameters are normalized, i.e., each parameter belongs to the interval $[-1, 1]$. Thus, one can conclude that the largest deviation is associated with the vertical position of the nose 4.57%. This can be explained by the fact that the nose and its surrounding areas are somewhat featureless. From the results in Table 3 it can be seen that the shape parameters are were accurately estimated under different 3D face poses.

	eyebrow	eye	eyes separa.	nose	mouth
Ave. dev.	2.65%	3.27%	1.7%	4.57%	1.22%

Table 3. Average deviation (in $\%$) over ten different individuals.

7. A human-robot interaction scenario

The proposed method for estimating user's head pose and person specific face shape (described in section 4) as well as the real-time tracker (described in section 5) are applied to a human-robot interaction scenario, for mapping an indoor environment (see Figure 1). By mimicking user's face movement, a robot's camera can take successive snapshots of the current perceived region. At the end of the process, the panoramic image of the region of interest is built, from the extracted snapshots by applying an image mosaicking technique [3]. We stress the fact that our proposed framework can be used by any Human-Robot Interaction application that is based on online tracking of subjects' head pose [9].

The experimental setup is depicted in Figure 8. The input to the system consists of a video stream capturing user's face from a fixed camera. The corresponding pitch and yaw angles of the user's face are encoded and sent to the robot through a wireless connection. Without any loss of generality, we used in our experiments Sony's AIBO robot, which has the advantage of being especially designed for interaction with persons. Thus, our application can be considered as a natural extension of AIBO's built-in behaviors. The orientation of robot's head (the robot's camera) is updated online according to the desired direction imposed by the user's face pose. Due to the motors response (as well the com-

munication through the wireless network), there is a very small delay between the desired orientation and robot's response. The inertia associated with the motors is most visible when a change in direction has to be performed. However, if user's movement is reasonably slow, but continuous, a real-time response from the robot can be expected.

Figure 9 illustrates the results of face movement imitation associated with a 691-frame sequence. In this video, the person looks around without any restriction. Only eight frames are shown in the figure. The left column displays the user's face pose and the right column shows the corresponding snapshot of the scene as seen by the AIBO's camera. Figure 10 illustrates a panoramic image computed from the captured individual snapshots. For this purpose, we used the AutoStitch^{TM} application [3], developed by M. Brown and D. Lowe from UBC, Canada. In a broader context, the user and the robot can be very distant from each other. This case corresponds to a telepresence application where the user is exploring a remote (dangerous or inaccessible) spot by only changing his/her head pose. This is the case, for instance, for rescue robots, which can be sent to areas affected by earthquakes or fires in order to perform an exhaustive search of the environment to find survivors. The use of a special display device that visualizes the scene as it is viewed by the remote camera will boost the telepresence feeling. If we assume that the orientation of the fixed camera is aligned with that of the robot camera (in its reference position) then its gaze direction will be equivalent to that of the user.

Figure 8. The experimental setup. The camera (in this image placed between the monitor and the keyboard) perceives user's 3D head pose and transmits it to the robot.

8. Conclusion

This paper described two main contributions. First, we proposed an automatic 3D face pose and person-specific shape initialization scheme for a real-time appearance-based tracker, based on a statistical facial texture model and a standard deformable 3D model. Second, we used the proposed methods—the initialization and 3D tracking—for controlling the movements of an AIBO's camera, by imitating a user's head movements. Applications such as telepresence, virtual reality can directly use the proposed tech-

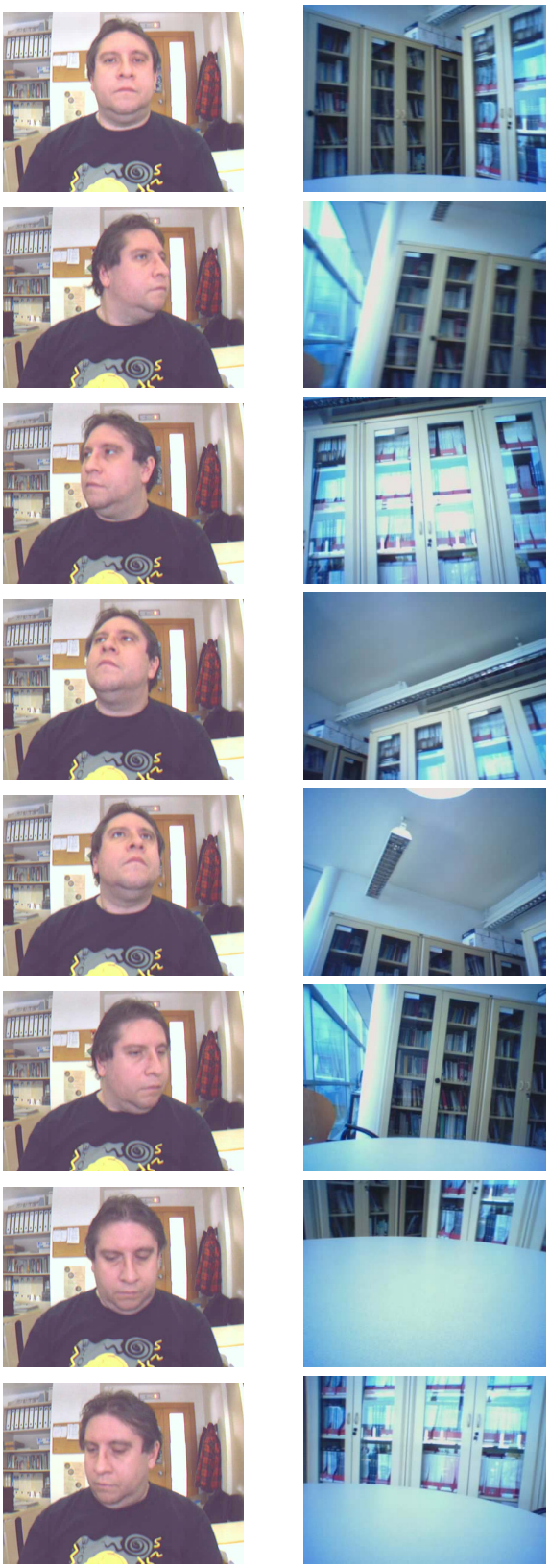


Figure 9. The library sequence. **Left column:** Some input images from the original video. **Right column:** The corresponding snapshot acquired by the controlled robot's camera.

Figure 10. A panoramic view obtained from the individual snapshots shown in Figure 9.

niques. The proposed method has several advantages that make it attractive: (i) the initialization step copes with faces under arbitrary pose, so it is not limited only to near-frontal views, (ii) it is person-independent, (iii) it is featureless (no facial features are needed), and (iv) its learning stage is simple. Future work will be devoted to learn more complex statistical facial texture models, like for instance the ones based on Laplacian Eigenmaps, Gaussian mixture models.

Acknowledgements

B. Raducanu is supported by the projects TIN2009-14404-C02-00 and CONSOLIDER-INGENIO 2010 (CSD2007-00018), Ministerio de Educación y Ciencia, Spain.

References

[1] J. Ahlberg. An active model for facial feature tracking. *EURASIP Journal on Applied Signal Processing*, 2002(6):566–571, June 2002.

[2] M. D. Breitenstein, D. Kuettel, T. Weise, L. Gool, and H. Pfister. Real-time face pose estimation from single range images. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1–8, Anchorage (AK), USA, June 2008.

[3] M. Brown and D. Lowe. Automatic panoramic image stitching using invariant features. *International Journal of Computer Vision*, 74(1):59–73, 2007.

[4] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–684, 2001.

[5] J. Demiris, S. Rougeaux, G. Hayes, L. Berthouze, and Y. Kuniyoshi. Deferred imitation of human head movements by an active stereo vision head. In *Proceedings of the 6th IEEE International Workshop on Robot and Human Communication*, pages 88–93, Sendai, Japan, 1997.

[6] F. Dornaika and F. Davoine. On appearance based face and facial action tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 16(9):1107–1124, September 2006.

[7] F. Dornaika and A. Sappa. Evaluation of an appearance-based 3D face tracker using dense 3D data. *Machine Vision and Applications*, 19(5-6):427–441, October 2008.

[8] S. Gurbuz and N. Inoue. Real-time head pose estimation using reconstructed 3D face data from stereo image pair. In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 785–788, Honolulu (HI), USA, April 2007.

[9] M. Hasanuzzaman and H. Ueno. *Face Recognition*, chapter Face and gesture recognition for Human-Robot interaction, pages 149–182. I−TECH, 2007.

[10] C. J. Kuo, R. Huang, and T. Lin. 3D facial model estimation from single front-view facial image. *IEEE Trans. on Circuits and Systems for Video Technology*, 12(3):183–192, 2002.

[11] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, 2004.

[12] A. Meltzoff and M. Moore. Explaining facial imitation: A theoretical model. *Early Development and Parenting*, 6:179–192, 1997.

[13] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):696–710, 1997.

[14] E. Murphy-Chutorian and M. Trivedi. Head pose estimation in computer vision: a survey. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 31(4):607–626, 2009.

[15] Y. Nagai. Joint attention development in infant-like robot based on head movement imitation. In *Third International Symposium on Imitation in Animals and Artifacts*, pages 87–96, 2005.

[16] K. V. Price, J. A. Lampinen, and R. M. Storn. *Differential Evolution: A Practical Approach To Global Optimization*. Springer, 2005.

[17] B. Scassellati. Imitation and mechanisms of joint attention: A developmental structure for building social skills on a humanoid robot. In C. Nehaniv, editor, *Computation for Metaphors, Analogy and Agents*, volume LNCS1562, pages 176–195. Springer-Verlag, Berlin Heidelberg, 1999.

[18] A. Shon, D. Grimes, C. Baker, and R. Rao. A probabilistic framework for model-based imitation learning. In *Proc. of the 26th Annual Meeting of the Cognitive Science Society*, 2004.

[19] P. Viola and M. Jones. Robust real-time object detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[20] J. Wang, E. Sung, and R. Venkateswarlu. Em enhancement of 3d head pose estimated by perspective invariance. In N. Sebe, M. Lew, and T. Huang, editors, *Computer vision in human-computer interaction: ECCV 2004 Workshop on HCI*, volume 3058 of *LNCS*, pages 187–199. Springer-Verlag, Berlin-Heidelberg, 2004.

[21] U. Weidenbacher, G. Layher, P. Bayerl, and H. Neumann. Detection of head pose and gaze direction for human-computer interaction. In E. A. et al., editor, *Perception and Interactive Technologies*, volume 4021 of *LNAI*, pages 9–19. Springer-Verlag, Berlin-Heidelberg, 2006.

[22] Z. Wen and T. Huang. *3D Face Processing: Modeling, Analysis and Synthesis*. Kluwer Academic Publishers, 2004.