EG UK Theory and Practice of Computer Graphics (2013)
Silvester Czanner and Wen Tang (Editors)

Accurate and marker-less head tracking using depth sensors

Martin Breidt[1 †], Heinrich H. Bülthoff[1,2†] and Cristóbal Curio[1†]

[1]Max Planck Institute for Biological Cybernetics
Tübingen, Germany

[2]Deptartment of Brain and Cognitive Engineering
Korea University, Seoul, Korea

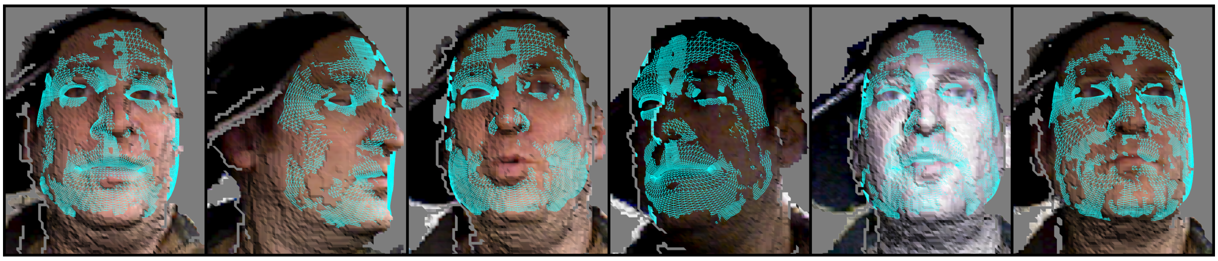


Figure 1: *Examples of accurately tracked head poses (green wireframe) in the presence of speech (1), large head rotation (2), facial expressions (3), illumination changes (4,5), large eye rotation (6).*

Abstract

Parameterized, high-fidelity 3D surface models can not only be used for rendering animations in the context of Computer Graphics (CG), but have become increasingly popular for analyzing data, and thus making these accessible to CG systems in an Analysis-by-Synthesis loop. In this paper, we utilize this concept for accurate head tracking by fitting a statistical 3D model to marker-less face data acquired with a low-cost depth sensor, and demonstrate its robustness in a challenging car driving scenario. We compute 3D head position and orientation with a mesh-based 3D shape matching algorithm that is independent of person identity and sensor type, and at the same time robust to facial expressions, speech, partial occlusion and illumination changes. Different strategies for obtaining the 3D face model are evaluated, trading off computational complexity and accuracy. Ground truth data for head pose are obtained from simultaneous marker-based tracking. Average tracking errors are below 6mm for head position *and below* 2.5° for head orientation, *demonstrating the system's potential to be used as part of a non-intrusive head tracking system for use in Augmented Reality or driver assistance systems.*

Categories and Subject Descriptors (according to ACM CCS): I.3.6 [Computer Graphics]: Methodology and Techniques—Interaction techniques I.4.8 [Image Processing and Computer Vision]: Scene Analysis—Surface Fitting

1. Introduction

Estimating accurate 3D head pose is important in many human-computer interfaces, in particular in the context of Augmented Reality (AR) systems. The recent introduction of head-up displays in the automotive industry has added to the general interest in AR technology. The desire for markerless and uncalibrated operation required in this industry, combined with its high quality standards, presents a particular challenge to the application of AR to the car driving context. While conventional AR systems try to recover 3D position and orientation information (pose) of imaging devices (e.g. head-mounted cameras, or cameras embedded in mobile devices), an AR system used in a car requires estimating the head pose of the driver.

Besides the typical AR use case of perspectively correct projection of 3D content into the visual field of the user, accurate head pose information has additional benefits in a

† {firstname.lastname}@tuebingen.mpg.de

DOI: 10.2312/LocalChapterEvents.TPCG.TPCG13.041-048

car driving scenario, e.g. for assessing user attention direction and, thus, overall situational awareness. In particular, advanced driver assistance systems (ADAS) have to cope with the problem of false or exaggerated responses after detecting potential emergency situations, which could result in reduced customer acceptance. Therefore, one of the most prominent application areas discussed in the automotive industry is judging the driver state and intention while maneuvering a car [TT11], for which the driver's head is a significant source of information.

Production-line sensors embedded in cars place high demands in terms of robustness and accuracy on the data analysis: changing environment information can influence sensor behavior; driver head pose undergoes subtle as well as large changes; user identities vary widely across an entire car fleet. In addition, the driver experience should be affected as little as possible, disqualifying marker-based approaches or those that require explicit user calibration. With the advent of low-cost 3D sensors, such as Microsoft Kinect, some of the shortcomings of purely video-based approaches can be bypassed with little additional cost.

In this work we present a head tracking system based on data acquired with a *Primesense* depth sensor, to which we fit and track a 3D face model. The data was recorded in real-world car driving situations, including strong illumination changes caused by entering and exiting tunnels. We evaluate our person-independent head tracking approach by comparing it to ground truth data acquired with a commercial, marker-based head tracking system synchronized to the depth data. In addition to the car driver scenario, we also show the system's applicability to challenging data of partially occluded faces due to extensive facial hair.

Our paper is structured as follows: In Section 2 we briefly review related work. Our recording setup and data acquisition is described in Section 3. Part of the processing pipeline outlined in Section 3.2 is our pose estimation algorithm described in Section 3.3. Section 4 evaluates our proposed system against ground truth. The discussion in Section 5 concludes this paper.

2. Related work

Whereas recent progress has been made towards pose estimation from 3D depth data [FDG*12], most work either used high quality sensor data only [WBB*09], or sacrificed accuracy for real-time performance [FWGVG11]. Our proposed system achieves accurate head tracking purely on low quality depth data. For a recent survey on head pose estimation see [MCT09]. Recent approaches for head pose estimation have started to exploit depth sensor cues in particular to efficiently estimate yaw, pitch, roll, and 3D position with regression trees on a frame-by-frame basis [FWGVG11, FDG*12]. That work demonstrates the potential to infer head pose from inaccurate depth data. Dantone *et al.* [DGFVG12] reported promising results based on similar algorithms for texture based landmark detection, yet their efficacy and robustness for 3D head pose estimation under car driving conditions needs to be shown. Breidt *et al.* [BBC11] demonstrate robustness in the analysis of 3D facial expression data from noisy Time-of-Flight sensor data. Weise *et al.* [WBLP11] computed 3D head pose as part of their real-time character animation based on Kinect depth data. So far, no work on 3D reconstruction of head pose has been reported for out-of-laboratory applications. In this paper we investigate in particular the potential of a 3D model-based approach to track faces in depth data with a generative face model under various driving situations.

3. Data acquisition

Our aim was to evaluate the precision and reliability of our head tracking system under real-world conditions, i.e. a car driving scenario in normal traffic conditions.

3.1. Recording setup

In order to properly record 3D data in a car we integrated two recording modalities in one setup. Figure 2 shows the major components. With this setup, our automatic head pose estimation from 3D depth data can be evaluated against marker-based head tracking data.

3.1.1. Depth sensor

As depth sensor, we used a *Primesense Carmine 1.09*, based on the same technology as Microsoft Kinect but with the advantage of being USB-powered and of smaller size. In addition, it is optimized for short range sensing, with a specified depth range of 0.35m – 1.4m and a vendor-specified field of view of 57.5° × 45° (H × V). Data was recorded with 30 frames/s at a resolution of 320 × 240 pixels for depth and color. The sensor was mounted in front of the driver on top of the dashboard (bottom left in Figure 2), at a distance of approximately 75cm to the driver, resulting in the driver's face covering approximately 50 × 70 pixels (see Figure 3). Our algorithm is largely independent of the actual sensor characteristics, as long as the depth data is suitable for triangulation.

3.1.2. Head tracking ground truth

For ground truth head pose data, we chose a marker-based *NaturalPoint TrackIR 5* system, consisting of an infrared monocular camera running at 120 frames/s at a resolution of 640 × 480 pixels and a rigid retro-reflective tracking target of known geometry. It has a specified horizontal Field of View of 51.7° and produces a full 6-DOF head pose. To improve the marker detection, a custom tracking target was built with identical dimensions as the original but with larger, spherical markers, rigidly attached to a baseball cap. The driver wore the cap sideways to optimize the orientation

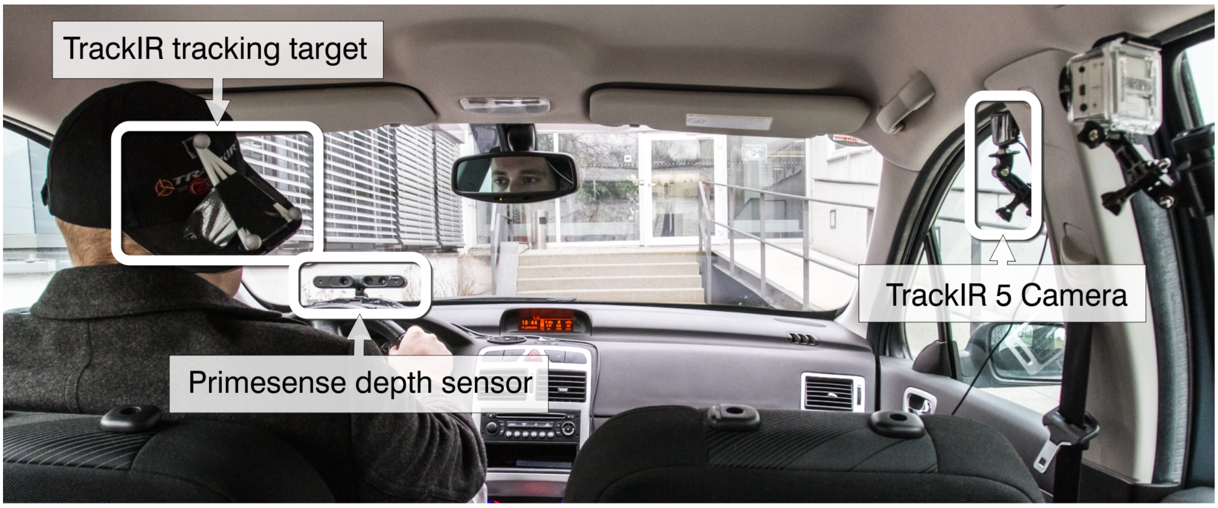


Figure 2: *Recording setup: Depth data of the driver's face was recorded with a* Primesense *sensor on the dashboard. For ground truth data, the driver was wearing a marker-based target, which was tracked by a* TrackIR 5 *camera mounted to the side window. In addition, a reference video was also recorded from behind using a* GoPro HD Hero2 *camera (top right).*

range in which the tracking was stable. The TrackIR camera was mounted to the front-seat passenger window (right in Figure 2) at a distance of approximately 70 centimeters to the tracking target.

3.1.3. Data recording

For our experiment, we recorded the depth and color stream of the 3D sensor along with the 6-DOF head tracking data. All data was time-stamped and logged directly onto a laptop for subsequent analysis. Time synchronization was achieved using an LED trigger device visible to all sensors. Care was taken not to move the cap during the entire course of the recording. After the driving sessions, a 3D scan of the entire head including the cap was taken, in order to measure the spatial relationship between cap, tracking target and face.

In addition to the car driving scenario, we also collected data in an office environment from a heavily bearded participant. With this dataset, we investigated the robustness of our system with respect to occlusions due to facial hair.

It should be pointed out that the color video stream is not mandatory for successful estimation of the head pose as long as a reasonable initialization can be found for the model fitting procedure described in Section 3.3.

3.2. Processing pipeline

All datasets were subjected to the same processing pipeline described in the following section.

3.2.1. Data corpus

For the driver scenario we collected a total of 5837 frames, which include merging into traffic, turning, lane changes, crossroad maneuvers, and a tunnel passage. 1140 additional frames were collected in the office environment while the bearded participant performed large head rotations. Depth data was triangulated using the regular pixel grid of the depth sensor, discarding measurements at a distance larger than 90 cm and triangles with an area larger than 30mm^2.

3.2.2. Merging modalities

First, the relative orientation of the tracking target to the face was determined by aligning the geometry of the tracking markers found in the 3D scan of the head to the 3D marker positions computed by the tracking system. Next, both the 3D scan and the tracking data were transformed such that the 3D scan optimally fitted into the first frame of the triangulated depth data, which provided the global coordinate system for all subsequent analysis.

The higher temporal resolution of the tracking system was resampled to match the frame rate of the depth sensor, with occasional linear interpolation of missing data, caused by short occlusion of the tracking target, using data from neighboring frames. The first frame of the color stream of the sensor was subjected to a face detector [KBFS05]. The resulting bounding box (see Figure 3) was used to initialize the subsequent model fitting analysis described in Section 3.3. This is the only use of the color stream in the entire analysis and could be easily replaced by other initialization schemes, e.g. by finding the tip of the nose [BJH*09]. For the comparison of different approaches for defining a face model (Section 3.3.1), the face detection window was also used to extract geometry from the first frame of the triangulated depth data (see Figure 3 top left). This rectangular shape snapshot was then later registered to the rest of the sequence.

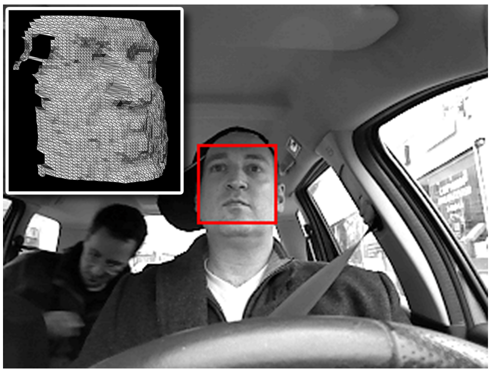


Figure 3: *Driver face detected in the color stream (converted to gray). Corresponding 3D data shown top left.*

3.3. Head tracking using depth data

For estimation of the head pose from recorded depth data, we apply a variant of the model-based fitting described in [BBC11] which builds upon a robust version of the well known Iterative Closest Point (ICP) algorithm [BM92]. We rigidly align a triangulated head model to the measured depth data, using a BFGS quasi-Newton solver for optimizing rotation and position (defining head pose), represented as matrix $\mathbf{R}$ formed from Euler angles, and translation vector $\mathbf{t}$. Our model uses a linear 3D face model that is obtained by applying PCA to 200 3D scans of neutral faces which were put into dense correspondence [BV99]. We retained 95% of the data variance using $m = 43$ principal components $\mathbf{d}_j$ as basis shapes, with $\mathbf{n}$ denoting the average head identity of the corpus. Before applying PCA, the original mesh (75k vertices) was reduced to an optimized mesh of 3980 vertices (Fig. 4). For head pose estimation, we minimized the following energy term

$$E(\mathbf{w},\mathbf{R},\mathbf{t}) = \frac{1}{k}\sum_{i=1}^{k} \min_{\mathbf{x}\in\mathcal{D}} ||\mathbf{R}(\mathbf{n}(i) + \sum_{j=1}^{m} w_j \mathbf{d}_j(i))) + \mathbf{t} - \mathbf{x}||^2, \tag{1}$$

representing the mean distance error between $k = 700$ evenly sampled vertices on the model surface and their respective closest points $\mathbf{x}$ on the data surface $\mathcal{D}$ (triangulated depth data from the sensor, also see inset in Figure 3). Further, $\mathbf{w}$ denotes the identity model weights corresponding to the loadings of the m model components, and index (i) the evenly subsampled 3D vertex vector. The j'th shape vector of the PCA model is defined by $\mathbf{d}_j$ complementing the average head $\mathbf{n}$ of the shape model. During optimization for identity, we enforce an L_2-norm on the PCA weights $\mathbf{w}$ as regularization. For temporal pose tracking, we keep $\mathbf{w}$ fixed and optimize for parameters $\mathbf{R}$ and $\mathbf{t}$ only.

Robust alignment To improve robustness against the significant noise and other errors present in the low-cost depth sensor data, only 50% of the closest point pairs with the smallest residuals were used in computing the fitting error of Equation 1 at each iteration step, thus ignoring the 50% farthest matches, similar to Fractional ICP [PLT07]. This proved to be very effective in cases where the face was rotated sideways, displayed facial expression or speech, presence of additional objects in the depth data, or partial occlusion of the face by significant amounts of facial hair (see Section 4.2).

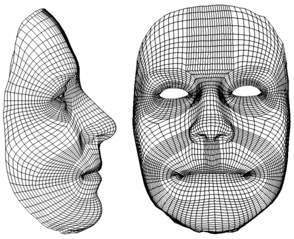


Figure 4: Identity-averaged face ***n*** *with optimized topology.*

3.3.1. Model types

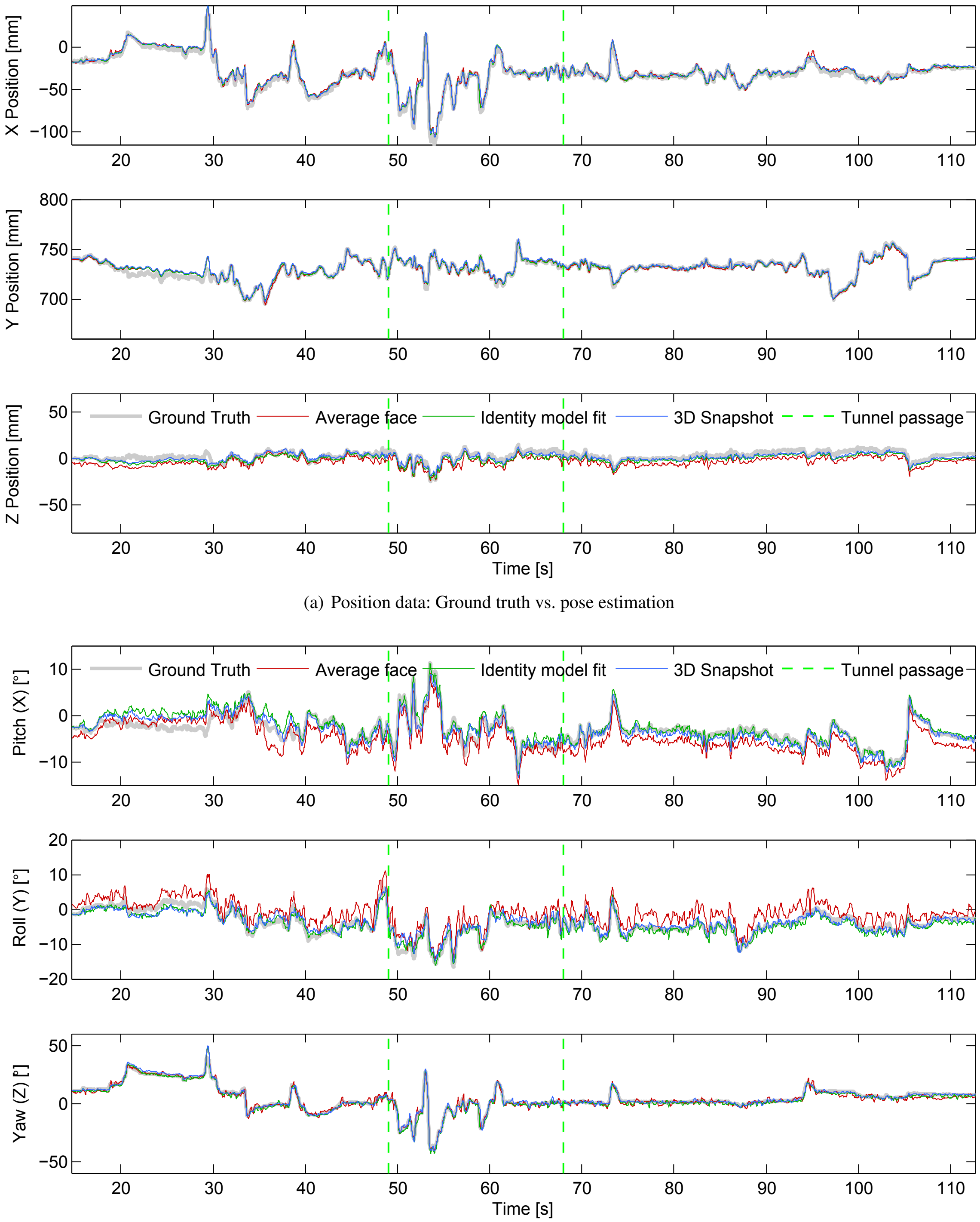
To investigate the contribution of the actual type of face model for tracking, we compared the performance of

- the *Average Face* $\mathbf{n}$ of the full statistical shape model (Eq. 1 with $w_j = 0$),
- a full *Identity Model fit*, matched to the neutral face of the participant once by estimating $\mathbf{w}$ in Eq. 1, and
- a *3D Snapshot*, directly extracted from the first frame of the unmodified depth data (see Figure 3).

Our process for head tracking is twofold: For the first two approaches listed above, an initial head pose is estimated by aligning an identity-averaged 3D face model to the depth data. The *3D Snapshot* only requires a face detection step to extract a surface mesh. The models were then further aligned to the sensor data, i.e. tracked for the rest of the sequence by estimating $\mathbf{R}$ and $\mathbf{t}$ in Eq. 1. For temporal tracking, a linear prediction of the head pose parameters from the current rate of change (i.e. velocity) proved sufficient to cope with large motion. The fitted shape model and the 3D snapshot geometrically match the participant's facial anatomy closely, whereas the identity-averaged face exhibits some major differences. Conversely, no additional preprocessing is required for the averaged face.

4. Evaluation

To evaluate the accuracy of our head tracking system, we compared it to the ground truth head pose data obtained from the tracking system described in Section 3.1.2.



(a) Position data: Ground truth vs. pose estimation

(b) Orientation data: Ground truth vs. pose estimation

Figure 5: *Head pose estimation data of sequence* Drive A *for the different model types. Green dashed lines mark the tunnel passage. Note that the different vertical axes use different scale values.*

4.1. Driving scenario

Figure 5 shows ground truth data for sequence *Drive A* in comparison with pose estimations produced by the three different model types described in Section 3.3.1. As the *3D Snapshot* model does not provide an absolute head pose, but only the pose relative to the initial snapshot, we initialized the snapshot pose with that of the *Identity Model Fit* for easier comparison. *Average Face* and *Identity Model Fit*, on the other hand, provide an implicit, absolute head pose. Analogous we initialized the ground truth coordinate system to the one provided by the *Identity Model Fit*.

Estimation Error Figure 6 plots the absolute error for the three head tracking types with respect to the marker-based ground truth. Positional error is calculated as Euclidean distance between the estimated head origin and the one provided by the head tracking system; orientation error is calculated as the minimum angle needed to rotate the estimated coordinate systems into the ground truth. For this, Euler rotation angles for the two orientations to be compared were converted into quaternion representations $\mathbf{q}_1, \mathbf{q}_2$; the quaternion distance $\mathbf{q}_d = \mathbf{q}_1^{-1}\mathbf{q}_2$ was then converted back into angle $\alpha = 2\arccos(\mathbf{q}_d)$.

Accuracy In order to compare our results with related work, we have calculated overall accuracy for position and orientation for all three model types. Figure 7 plots obtained accuracy over accuracy threshold.

The *Identity Model Fit* show 95% accuracy at an error threshold of 4.13° for orientation and 8.21mm respectively for position. For the *3D Snapshot* model, 95% accuracy is achieved at an error threshold of 3.88° and 8.22mm. Finally, the *Average Face* model reaches 95% accuracy at an error threshold of 6.10° and 11.39mm.
In comparison to Fanelli *et al.* [FDG*12], our head pose estimation exhibits accuracies of 99.80% for orientation and 97.69% for position when applying 10° and 10mm thresholds, in comparison to their reports of 94.7% and 73.0% accuracy for orientation and position. Figure 1 shows the accurate alignment of the *Identity Model Fit* to triangulated depth data for different head poses and facial expressions in sequence *Drive A*. Picture 4 and 5 from the left show the robustness of the system with respect to illumination changes caused by a tunnel passage (indicated by the green vertical dashed lines in Figure 5 and Figure 6) which introduced major illumination changes into the data but did not affect the estimation quality.

4.2. Office scenario: Hairy tracking

We were also interested in testing our tracking approach on partially occluded faces. For this, we recorded data of a participant wearing a beard that covers large portions of the face (Fig. 8). Remarkably, the *Identity Model* still fits accurately to the face underneath the beard, which can be largely attributed to the robust distance measure employed during optimization. In particular, for 10° and 10mm accuracy thresholds, tracking still achieves 100.0% and 89.7% accuracies for head orientation and translation. Similar values are obtained for the *3D Snapshot* and the *Average Face* model, with the latter degrading earlier for lower accuracy thresholds (see Fig. 7(d)–7(f)).

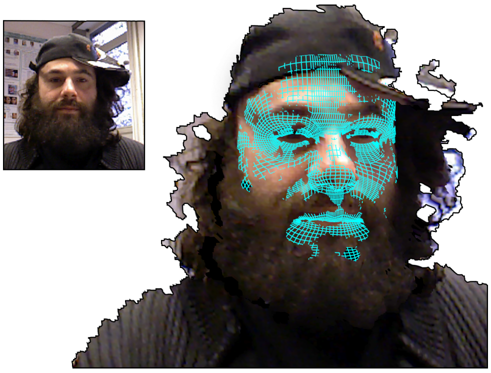


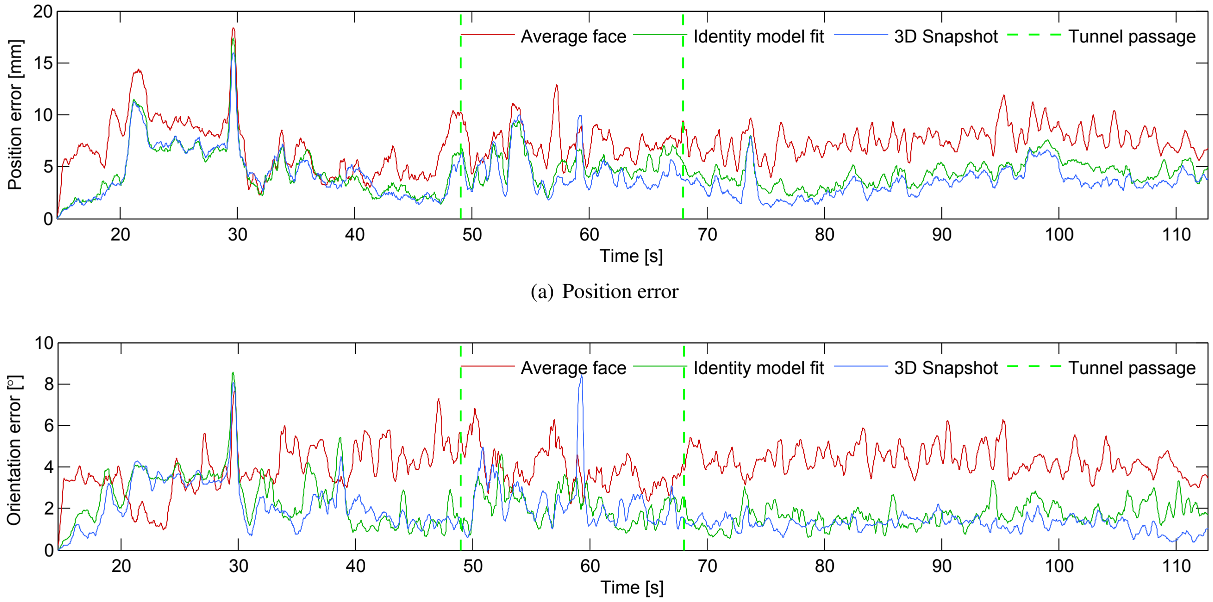
Figure 8: Identity Model *successfully tracked to the noisy sensor data despite occlusion due to facial hair.*

4.3. Evaluation summary

Table 1 compares median position and orientation errors for all sequences and model types. As expected, the *3D Snapshot* model produces the smallest errors, but requires a correct face detection step and cannot provide absolute head pose information. On the other end, the *Average Face* model has the highest median errors but does not require additional preprocessing such as face detection or identity model fitting. The *Identity Model Fit* provides high accuracy and absolute head pose information, even for challenging data.

5. Discussion

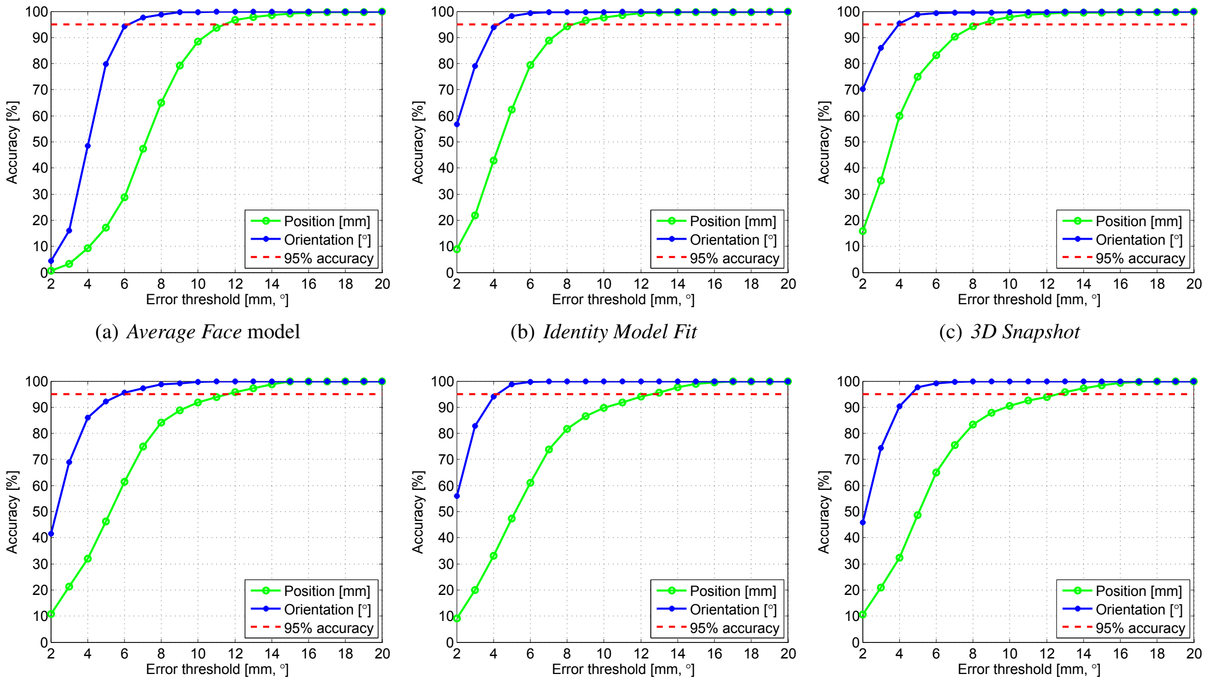
In this paper we have reported results for 3D head tracking from depth sensor data using an adaptive 3D face model in a car driver context. Tracking accuracy of head pose exceeds that reported in previous work on independent head pose detection. Our generative model makes no assumption on the sensor or the driver, is robust to noise and illumination changes, and provides accurate reconstruction results even for challenging data acquired outside the laboratory. Robust identity fitting demonstrates the ability to deal with faces that cannot be entirely captured by our statistical head model, e.g. due to extensive facial hair. As compared to a pure mesh-based approach (*3D Snapshot*) without face domain knowledge, we have shown that the model-based tracking approach provides the coordinate system required in order to accurately estimate gaze. The pure snapshot-based ap-



(a) Position error

(b) Orientation error

Figure 6: *Head pose estimation errors for sequence* Drive A. *Green dashed lines mark the tunnel passage. For better readability, errors were filtered with a moving-window box filter of* 0.5 *seconds width.*



(a) *Average Face* model (b) *Identity Model Fit* (c) *3D Snapshot*

(d) *Average Face* model (e) *Identity Model Fit* (f) *3D Snapshot*

Figure 7: *Accuracy of the head pose estimation for position and orientation. Graphs (a)–(c) show data for the* Drive A *sequence, (d)–(f) data for the* Office *sequence.*

				Average Face error		*Identity Model Fit* error		*3D Snapshot* error	
Sequence	Duration	Yaw Range	Pitch Range	Position	Orientation	Position	Orientation	Position	Orientation
Drive A	2940 frames	$[-41°, 50°]$	$[-13°, 11°]$	7.16mm	4.04°	4.38mm	1.80°	3.60mm	1.49°
Drive B	2563 frames	$[-62°, 61°]$	$[-9°, 15°]$	7.46mm	4.35°	7.58mm	2.43°	7.62mm	1.93°
Office	1140 frames	$[-46°, 45°]$	$[-19°, 24°]$	5.28mm	2.26°	5.21mm	1.83°	5.07mm	2.17°
Mean	6643 (total)	$[-50°, 52°]$	$[-13.7°, 16.7°]$	6.63mm	3.55°	5.72mm	2.02°	5.43mm	1.86°

Table 1: *Median error between head pose estimation using different model types, and ground truth.*

proach is of similar performance but lacks the absolute pose information.

Outlook Overall, we presented an Analysis-by-Synthesis technique to determine accurate 3D head pose, using a parametrized 3D face model, similar to those used in CG animation and rendering. The parameters of this generative model are optimized to explain the observed sensor data. In contrast, efficient but less accurate discriminative models have been previously suggested by means of regression forests [FDG*12]. Only few approaches have attempted to combine both, discriminative and generative, methods, e.g. for human pose tracking [CG05]. Developing the fusion of such approaches is expected to be a fruitful research direction. A real-time implementation could open a new window to applications that require high accuracy and reliability; in automotive mass markets, passenger safety applications such as adaptive airbag control, gaze direction for attentive state monitoring, or perspective correction for head-up displays in an Augmented Reality system could benefit from this work. In addition, the *Identity Model Fit* approach could also be used biometrically to determine facial identity of the driver, similar to [BV03]. The addition of facial deformation information to the shape model [BBC11] would allow to estimate facial expression parameters for a more detailed driver state analysis.

Acknowledgements

This work was in part sponsored by DFG project Perceptual Graphics PAK 38 CU 149/1-1/2 and EU Project TANGO (ICT-2009-C 249858). It was also supported by WCU (World Class University) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (R31-10008). We would like to thank Mario Kleiner for not shaving for a while.

References

[BBC11] Breidt M., Bülthoff H. H., Curio C.: Robust semantic analysis by synthesis of 3d facial motion. In *Automatic Face & Gesture Recognition (FG 2011)* (2011), IEEE, pp. 713–719. 2, 4, 8

[BJH*09] Breitenstein M. D., Jensen J., Hølund C., Moeslund T. B., Gool L. V.: Head pose estimation from passive stereo images. In *Proc. 16th Scandinavian Conference on Image Analysis* (2009), SCIA '09, Springer, pp. 219–228. 3

[BM92] Besl P. J., McKay N. D.: A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence 14*, 2 (Feb. 1992), 239–256. 4

[BV99] Blanz V., Vetter T.: A morphable model for the synthesis of 3d faces. *ACM SIGGRAPH* (1999), 187–194. 4

[BV03] Blanz V., Vetter T.: Face recognition based on fitting a 3d morphable model. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 25*, 9 (2003), 1063–1074. 8

[CG05] Curio C., Giese M. A.: Combining view-based and model-based tracking of articulated human movements. In *Application of Computer Vision (WACV/MOTIONS '05)* (2005), vol. 2, IEEE, pp. 261–268. 8

[DGFVG12] Dantone M., Gall J., Fanelli G., Van Gool L.: Real-time facial feature detection using conditional regression forests. In *Computer Vision and Pattern Recognition (CVPR 2012)* (2012), IEEE, pp. 2578–2585. 2

[FDG*12] Fanelli G., Dantone M., Gall J., Fossati A., Van Gool L.: Random forests for real time 3d face analysis. *International Journal of Computer Vision* (2012), 1–22. 2, 6, 8

[FWGVG11] Fanelli G., Weise T., Gall J., Van Gool L.: Real time head pose estimation from consumer depth cameras. *Pattern Recognition* (2011), 101–110. 2

[KBFS05] Kienzle W., Bakir G., Franz M., Schölkopf B.: Face detection - efficient and rank deficient. *Advances in Neural Information Processing Systems 17* (2005), 673–680. 3

[MCT09] Murphy-Chutorian E., Trivedi M. M.: Head pose estimation in computer vision: A survey. *Transactions on Pattern Analysis and Machine Intelligence 31*, 4 (2009), 607–626. 2

[PLT07] Phillips J. M., Liu R., Tomasi C.: Outlier robust ICP for minimizing fractional RMSD. In *Proceedings of the Sixth International Conference on 3-D Digital Imaging and Modeling* (Washington, DC, USA, 2007), 3DIM '07, IEEE Computer Society, pp. 427–434. 4

[TT11] Tran C., Trivedi M. M.: Vision for driver assistance: Looking at people in a vehicle. *Visual Analysis of Humans* (2011), 597–614. 2

[WBB*09] Walder C., Breidt M., Bülthoff H., Schölkopf B., Curio C.: Markerless 3d face tracking. *Pattern Recognition (DAGM 09)* (2009), 41–50. 2

[WBLP11] Weise T., Bouaziz S., Li H., Pauly M.: Realtime performance-based facial animation. *ACM Transactions on Graphics (Proceedings SIGGRAPH 2011)* (August 2011). 2