FI-CAP: ROBUST FRAMEWORK TO BENCHMARK HEAD POSE ESTIMATION IN CHALLENGING ENVIRONMENTS

Sumit Jha and Carlos Busso

Multimodal Signal Processing (MSP) Lab, Department of Electrical and Computer Engineering
University of Texas at Dallas, Richardson TX 75080, USA

Sumit.Jha@utdallas.edu, busso@utdallas.edu

ABSTRACT

Head pose estimation is challenging in a naturalistic environment. To effectively train machine-learning algorithms, we need datasets with reliable ground truth labels from diverse environments. We present Fi-Cap, a helmet with fiducial markers designed for head pose estimation. The relative position and orientation of the tags from a reference camera can be automatically obtained from a subset of the tags. Placed at the back of the head, it provides a reference system without interfering with sensors that record frontal face. We quantify the performance of the Fi-Cap by (1) rendering the 3D model of the design, evaluating its accuracy under various rotation, image resolution and illumination conditions, and (2) comparing the predicted head pose with the location of the projected beam of a laser mounted on glasses worn by the subjects in controlled experiments conducted in our laboratory. Fi-Cap provides ideal benchmark information to evaluate automatic algorithms and alternative sensors for head pose estimation in a variety of challenging environments, including our target application for *advanced driver assistance systems* (ADAS).

Index Terms— Head pose estimation, benchmark for head pose.

1. INTRODUCTION

Tracking rigid head motion has useful applications in various areas. The position and orientation of the head can give crucial information about the visual attention of a subject [1], emotional state [2] and engagement in a conversation [3]. In a driving environment, head pose provides useful information about the driver's visual attention [4,5]. Head motion can provide non-verbal cues during human interaction [6]. Head pose estimation is also a crucial pre-step in facial expression analysis and face landmark location [7]. The wide range of applications for head pose estimation has established the need for a stand-alone tool that automatically estimates head pose in a variety of challenging environments. A challenge in designing such a system is to obtain continuous annotated data for head pose of a subject in diverse situations, providing ground truth labels to train robust classifiers.

Several studies have explored benchmark approaches to annotate the position and orientation of the head. Previous studies have used magnetometers [8], *Inertial Measurement Units* (IMUs) [9], motion capture systems [10] and fiducial markers [4]. These approaches have limitations. For example, magnetometer provides unreliable information when metal is present in the environment. Data obtained from IMU is very noisy, especially in environments with additional vibrations such as vehicles, which is our target application. While motion capture systems can provide highly accurate information, they require a specialized and controlled setup. Using fiducial markers can occlude the face if the setting is not properly designed. There is a need for a system that can be used with minimal effort to reliably provide ground truth labels for head pose estimation in unconstrained scenarios.

This study presents Fi-Cap, a cap-like design with fiducial markers that addresses most of the challenges to benchmark continuous head poses in challenging scenarios. The study is inspired by the work of Jha and Busso [11], where a head band with AprilTags was used to track the head motion of a driver. Fi-Cap improves this design in three significant aspects: (1) the size of each square is bigger and, therefore, the fiducial markers can be detected more accurately, (2) the design considers markers in the horizontal and vertical directions, increasing the precision for pitch, yaw and roll rotations, and (3) the cap is worn on the back of the head without interfering with any sensor used to record the subject's face. The design of the Fi-Cap system allows multiple fiducial markers to be seen regardless of the head rotation of the target subject. As long as few of the tags are visible, the system is able to provide reliable frame-by-frame estimations which can be used as ground truth for training and evaluating head pose estimation algorithms.

We conduct evaluations to validate the accuracy of the Fi-Cap system. In a virtual environment, we simulate the proposed design by rendering a virtual character wearing the Fi-Cap system. We use different rotations, illumination conditions, and image resolutions. The system is very robust against these variations. We also conduct experiments in our laboratory, where we ask subjects to wear the Fi-Cap system. The subjects also wear a laser mounted on glasses, projecting a trackable mark on a white screen signaling their head pose. The evaluation suggests that with a few seconds of calibration, the system provides reliable head pose estimation.

The Fi-Cap design opens opportunities to advance head

This work was supported by Semiconductor Research Corporation (SRC) / Texas Analog Center of Excellence (TxACE), under task 2810.014.

978-1-5386-1737-3/18/$31.00

pose estimation algorithms that work on real, challenging conditions. Using Fi-Cap, we plan to collect data in naturalistic driving environment where current head pose estimation algorithms seem to struggle [11]. To Illustrate the potential of Fi-Cap, we conduct a preliminary recording during naturalistic conditions of a driver wearing the proposed Fi-Cap. The results show that at least one tag is observed in 99.8% of the frames. Furthermore, we are able to detect four or more tags in 99.2% of the frames, increasing the reliability of head pose estimation. These large naturalistic databases are needed for designing robust algorithms for *advanced driver assistance systems* (ADAS).

2. RELATED WORK

Head pose estimation is an important research problem with implications in many fields. Murphy-Chutorian and Trivedi [12] and Czupryński and Strupczewski [13] provided surveys with advances in head pose estimation systems. Some of the off-the-shelf facial processing tools such as IntraFace [14], OpenFace [15] and zFace [16] include state-of-the-art head pose estimation algorithms from regular cameras. Other studies have explored the use of depth data to estimate head pose using sensors such as the Kinect [17]. Since recent studies have started to use more advanced sensors such as Radar [18] or Ultrasound [19] to track human movements, it is expected that these technologies will also be employed to predict head pose. A prerequisite for development in this area is data with annotated head pose labels. Ideally, the data should be collected across diverse environmental settings. While humans can easily distinguish between coarse head poses, it is difficult to reliably quantify the exact position and orientation. Therefore, reference systems are required to provide reliable ground truth to train machine-learning algorithms for head pose estimation.

Head pose estimation databases have relied on different reference systems for head pose. A common approach in early recordings was to ask the subject to look at predefined locations. In the Pointing'04 database [20], subjects were asked to sit in the center of the room and look at different markers on the walls. They placed markers on the walls creating a discrete grid with 13 horizontal locations and nine vertical locations. A similar approach was used for the Bosphorus database [21], which was collected by asking the subjects to perform seven different yaw rotations, four pitch rotations and two combinations of yaw and pitch rotations. They provided qualitative instructions for pitch rotation (e.g., downwards, slight upwards), so the head pose are not necessarily consistent across subjects. These approaches require the subjects to direct their head toward the target mark, while avoiding any eye movement. Therefore, they are prone to errors. A slightly more accurate method was employed by Rae and Ritter [22]. They mounted a laser on top of the subject's head to verify the required head orientation. While these studies have provided important milestones to advance the research in head pose estimation, more advanced systems require accurate ground truth for head pose. Moreover, this approach can only provide information for pitch and yaw movements, ignoring roll rotation.

Magnetic sensors such as the flock-of-bird provide a helpful method for annotating head pose in all six degrees of freedom. The setup consists of a transmitter that can track the position and orientation of magnetic sensors. The subjects are asked to place the sensor on their head, which is tracked by the transmitter. Ba and Odobez [8] collected a database with two hours of video in office and meeting environments with head pose annotations provided by magnetic sensors. Ariz et al. [23] also collected a database with 10 subjects performing guided head poses, as well as free head motions. They designed a head band with magnetic flock-of-birds to track head motions. Magnetic sensors provide an accurate method for continuous annotation of head pose. However, they are limited by the environment, since the presence of metal can cause high fluctuation in the data [12].

Inertial measurement units (IMUs) are another option to track head movements. Most widely used and cheaper IMUs such as InertiaCube^2 provide rotation angles, but not positions. Morency et al. [9] used IMU sensors as a reference to design an adaptive view-based appearance model for head pose. Tawari et al. [24] used a pair of calibrated IMUs in a car to counter the effect of vibrations from the car. A reference IMU was placed inside the car and the second IMU was worn by the driver to obtain the head pose. While IMUs provide an effective method to track motions, they are highly susceptible to noise in the form of micromotions. These artifacts should be eliminated by using noise reduction methods such as Kalman filters. Also, there is a drift observed in the recording, making the data suitable only for short recordings.

Motion Capture (MoCap) systems are also useful tools to track motions. The systems rely on active markers with synchronized LEDs, or passive markers with reflective surfaces. These markers are placed on the object of interest, which are tracked by the system. The IEMOCAP database [10] was collected with a MoCap system, recording the head movement of the subjects (facial expressions and hand movements were also collected). A similar approach was used for the MSP-AVATAR database [25], which includes dyadic interactions. MoCap systems can also be useful in diverse environments. Murphy-Chutorian et al. [26] designed an algorithm to estimate head pose from monocular camera in a car. They used a setup with multiple MoCap reflectors and cameras placed at different angles to provide reference head poses. Schwarz et al. [17] also collected a driver head pose dataset using a Kinect camera, recording 2D and 3D data. They used a MoCap system with markers placed at the back of the head to capture ground truth for the driver's head motion. While MoCap systems provide accurate ground truths for head poses, the setup is often expensive and require specialized cameras for the recordings.

Our approach is similar to MoCap systems, without the

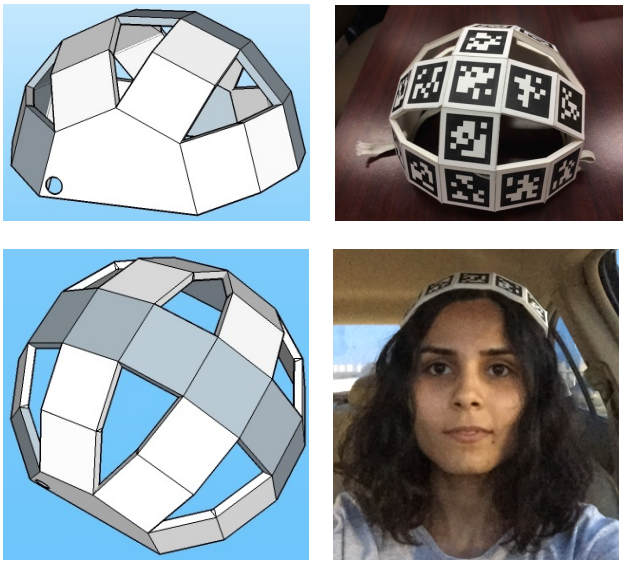

Fig. 1. Design of Fi-Cap, which is worn on the back avoiding occlusions with primary sensors.

need of expensive sensors. We use barcode markers that can be easily detected using a simple RGB camera. Tag based markers such as ARToolkit [27], ARTags [28] and AprilTags [29] have found multiple applications in robotics, augmented reality and computer vision. AprilTags [29] are 2D fiducial markers with black and white patterns. The patterns can be detected in an image. Since the design and the size are known, it is easy to estimate the six degrees of freedom of the surface containing the AprilTag. This paper builds on the framework proposed by Jha and Busso [11]. Their study presented some of the challenges in head pose estimation in a driving setting, where the head pose labels were obtained with an AprilTag-based headband. A main limitation of the headband used in their experiment is the occlusion of the upper part of the face, which interferes with current head pose estimation algorithms. This study solves this limitation with a cap structure that a subject wears in the back of the head, avoiding interference with most sensors (depth, radar, RGB cameras).

3. FI-CAP

Figure 1 shows the design of the proposed Fi-Cap, which rely on 2D fiducial tags. The design aims to provide reliable frame-by-frame head pose information without occluding the face of the subjects. Since Fi-Cap is worn on the back of the head, the design does not interfere with primary sensors used for head pose estimation. Fi-Cap has 23 different AprilTags that can be individually detected, providing a robust framework to detect the orientation of the cap. A reliable estimation can be obtained as long as few markers are visible. The system requires an extra camera behind the subject to record the position and orientation of the Fi-Cap. Since the detection of the tags is purely from images, the setup is simple and robust. This section explains the design of the Fi-Cap system.

3.1. Structure of Fi-Cap

The structure of the Fi-Cap system is designed to increase angular resolution for pitch and yaw rotations. We achieve this goal by creating a 3D structure with multiple fiducial markers along the vertical and horizontal directions. This structure also facilitates reliable estimation for roll rotation. The design provides enough diversity such that multiple squares are always visible for any reasonable head rotation.

The Fi-Cap is a 3D cap with 23 square faces (4cm $\times$ 4cm). The angle between adjacent faces is 24.5 degrees. A unique AprilTag of size 3.2cm x 3.2cm is placed at the center of each of the 23 faces. The size of these fiducial markers are twice as big as the AprilTags used on the headband introduced by Jha and Busso [4]. The size is important as it facilitates the location of the tags using automatic algorithms. An elastic band is attached to fit the cap on the head.

3.2. Estimating Orientation of Fi-Cap from AprilTags

We obtain the orientation of the Fi-Cap from the AprilTags. The first task is to detect the AprilTags. We use the algorithms proposed by Olson [29]. The camera parameters and the true size of the tag are used to estimate the position and orientation of each AprilTag with respect to the camera coordinate system. Then, the shape of each tag can be rebuilt using the estimate of the tag pose. We estimate the corner points of each of the visible tags as shown in Figure 4 (see blue dots on the corners of the tags), which we use to compute the pose of the Fi-Cap from the RGB images. The corner points of the tags are combined to obtain a single mesh of all the visible AprilTags, creating a partial shape of the Fi-Cap structure from the image. This partial mesh is compared with a reference model of the Fi-Cap that has registered all the 92 existing corners of the AprilTags. This reference is obtained with various frames from multiple angles under controlled conditions (illumination, rotation, resolution). We rely on the Kabsch algorithm [30] to obtain an initial estimate of the transformation between the pose of the Fi-Cap and the reference using all the visible corners. Subsequently, unreliable points that do not fit the original mesh are removed to obtain a better estimate. This approach is similar to the *iterative closest point* (ICP) algorithm. We repeat this process until the estimation error of the visible corners is below a given threshold (notice that this metric can be used as a reliability of the head pose estimation). This approach provides the estimate of the position and orientation of the cap with respect to the reference model.

3.3. Calibration

Once the position and orientation of the cap is estimated, we estimate the transformation from the Fi-Cap to the actual head pose (rotation matrix). The key goal in this calibration is to determine the head pose that is considered frontal with respect to our reference system (i.e. no rotation in any axis). This process is conducted at the beginning of the recordings, since

(a) neutral illumination (b) high illumination (c) low illumination

Fig. 2. Rendered images with different illuminations.

the actual placement of the Fi-Cap varies across subjects, and even across sessions. The experimental evaluation shows that the system can be calibrated with only a few seconds.

The calibration process can be conducted with different approaches. For example, the subject can be asked to look at a few points, directing his/her face toward target points. Alternatively, we can rely on automatic head pose estimation under good environment conditions. An advantage of the second approach is that the estimation of the algorithm and Fi-Cap are under the same reference system and can be directly compared.

4. VALIDATION USING SIMULATIONS

Before evaluating the design in a real world setting, we render an animation of a virtual character wearing a design of the Fi-Cap in Blender. Since we can program the orientation of the head, we can estimate the robustness of the system as we change the illumination and head orientation. We generate videos at different image resolutions, changing the illumination. We estimate the head pose orientation of a driver, collected in a previous study conducted by our laboratory. We use the head pose sequence to render the animation (1,800 frames – 60s at 30 fps). The approach aims to explore the performance of the system in realistic head orientations observed during a naturalistic driving scenario.

Three different illumination conditions are used: neutral, high and low illuminations. Figure 2 shows an example for each condition. For neutral illumination, we place three lamps around the head that provide uniform illumination (Fig. 2(a)). For high illumination, we add three spotlights on the Fi-Cap, generating highly saturated images (Fig. 2(b)). For low illumination, we use a single light source in the front right corner that causes shadow artifacts in the image (Fig. 2(c)). All the videos are created in two resolutions: 960×540 pixels (540p) and 1920×1080 pixels (1080p). We create six videos in total.

The rendered videos are the input of our Fi-Cap system, which detects the orientation and position of the head of the virtual character for each frame. The estimations are compared with the actual orientation used to render the animations. The error between the measurements is obtained in terms of the arccosine of the dot product of the two rotation quaternions (ϕ_3 in Huynh [31]).

$$\Delta(q_1, q_2) = \arccos(|q_1.q_2|) \quad (1)$$

Table 1 reports the mean, median and 95 percentile error between the estimated and the provided angles. While the mean and median shows the overall statistics of the error, the 95 percentile mark shows the worst case scenario, ignoring outliers. The table shows that the best results are obtained using neutral illumination at 1080p resolution. The median error is only 0.36° and the 95 percentile error is 0.66°. We can reliably estimate the head orientation for this condition. The performance degrades for the neutral setting at lower resolution. We observe a median error of 1.18° and the 95 percentile error is 1.95°. Shadows and illumination add extra challenges in the estimation as some of the tags may not be visible in the image. However, the results are still good. Even for the most challenging cases, the 95 percentile error mark is less than 3°. While we observe additional artifacts when looking at a real world situations, this evaluation suggests that the framework can reliably estimate the orientation of the head.

Table 1. Estimation error of head pose in our simulations.

Data	Mean error [°]	Median error[°]	95 percentile error[°]
540p neutral	**1.17**	**1.18**	**1.95**
540p high	**1.45**	**1.61**	**2.10**
540p low	**1.17**	**0.95**	**2.87**
1080p neutral	**0.39**	**0.36**	**0.66**
1080p high	**1.12**	**1.15**	**1.46**
1080p low	**0.93**	**1.01**	**1.30**

5. VALIDATION IN LABORATORY SETTING

We design a second experiment to validate the use of Fi-Caps in a controlled laboratory setting. For this purpose, we use a glass frame with a laser attached at the center. The head movement can be tracked by locating the position at which the laser beam is projected on a screen (similar to Rae and Ritter [22]) (Fig. 4). We asked subjects to wear both the laser frame and the Fi-Cap. We record the data with a camera placed behind the subjects head, such that both the screen and the Fi-Cap are visible in the camera view. We asked the subjects to freely look at arbitrary locations on the screen. The true location of the laser on the screen is estimated using template matching. We only consider frames where the beam location is accurately estimated. We collect data from 10 different subjects using a GoPro Hero6 camera. Each subject is recorded for about 90 seconds. Among all these frames, the laser beam was accurately detected in 21,683 frames.

For each subject, we need to calibrate the Fi-Cap system, aligning the cap with the direction of the laser beam. We use a portion of the data for this purpose and use the rest of the data to test the accuracy of the system. We explore different number of samples per session to find the ideal calibration of the system, as explained in Section 3.3. We evaluate the first 10, 100 or 500 frames for the calibration, where the remaining frames are used to evaluate the system. We also evaluate selecting 100 random frames from all over the video, using

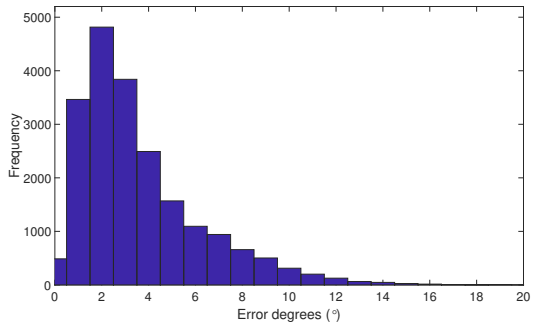

Fig. 3. Histogram of head pose error when we calibrate the system with the first 100 samples per session.

Table 2. Estimation error in head pose in laboratory settings.

Data Calibration	Mean error [°]	Median error [°]	95 percentile error [°]
First 10	**4.93**	**3.8**	**11.64**
First 100	**3.64**	**2.86**	**8.64**
First 500	**3.34**	**2.86**	**7.03**
Random 100	**2.91**	**2.30**	**6.92**

the remaining frames to calculate the error.

Table 2 shows the error for different calibration conditions. We observe a slightly higher error compared to the virtual environment. When using 10 samples, we observe that the median error is 3.8° while the 95 percentile error is 11.64°. Increasing the number of samples for calibration decreases the error. The median error is less than 3° if we use 100 samples for calibration (less than four seconds). With 500 samples, there is no significant change in the median error, although the 95 percentile error is reduced. This result implies that a few seconds of recording at the beginning is enough for calibrating the position and orientation of the Fi-Cap to the head pose. While calibrating with random samples seem to give us a better results with only 2.3° median error, this approach would require to have reference values available throughout the video which is unfeasible in real world situations. Figure 3 shows the histogram of the head pose errors when we calibrate the system with the first 100 samples of each session. The error in most of the frames is below 5°.

6. NATURALISTIC DRIVING RECORDINGS

To evaluate Fi-Cap in a naturalistic driving environment, we collect one hour and eight minutes of data of a driver wearing the device. We collect the data with two GoPro cameras (60fps), one facing the driver, and one behind the driver facing the tags. The data was collected during daytime in different roads including residential areas and highways. Out of 246,643 frames, we are not able to detect any tag in only 445 frames, providing a head pose estimation in 99.8% of the frames. We estimate the head pose with at least four tags in 244,581 frames (99.2%), suggesting that it is feasible to reliably estimate the head position and orientation of the driver.

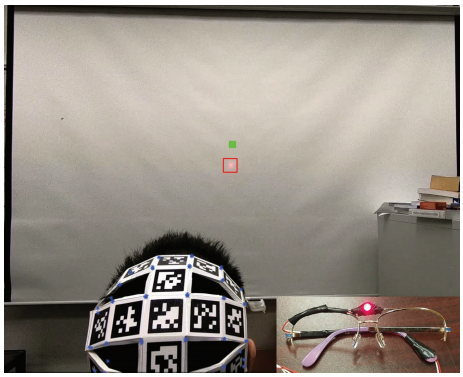

Fig. 4. Setting for the recordings. The laser is mounted on glasses projecting a beam on the screen, which is automatically tracked. The green point is the projection of the head pose estimation using Fi-Cap.

This is a significant improvement over the headband proposed by Jha and Busso [11].

7. CONCLUSIONS

This paper presented Fi-Cap, a framework that uses 2D fiducial points to reliably estimate the position and orientation of the head. The cap is worn in the back of the head, avoiding occlusion with primary sensors. The 3D structure of the Fi-Cap provides good resolution in the vertical and horizontal directions. The cap has big tags that can be easily detected using automatic algorithms. For reasonable head poses, multiple tags are always visible providing enough information to reconstruct the orientation of the design. Fi-Cap provides the infrastructure to collect large databases with continuous, frame-by-frame annotated head poses in diverse environments. In addition to our primary driving application, Fi-Cap can also be helpful in collecting head pose data in various settings in human-human interactions, human-computer interactions or human-robot interactions.

We plan to collect a database in naturalistic driving settings, which can be helpful in designing robust head pose estimation algorithms. Providing the annotated data will be crucial to solve the challenges observed by current head pose estimation tools in naturalistic driving scenarios. Maybe the answer is to use other non-invasive sensors such as infrared, radar or ultrasound. In each of these cases, the benchmark head pose labels can be obtained with our Fi-Cap. Advances in this area will lead to better driver behavior modeling for ADAS, and better designs for smart systems for safety and infotainment.

8. REFERENCES

[1] S. Jha and C. Busso, "Probabilistic estimation of the driver's gaze from head orientation and position," in *IEEE International Conference on Intelligent Transportation (ITSC)*, Yokohama, Japan, October 2017, pp. 1630–1635.

[2] C. Busso, Z. Deng, M. Grimm, U. Neumann, and S. Narayanan, "Rigid head motion in expressive speech animation: Analysis and synthesis," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 3, pp. 1075–1086, March 2007.

[3] B. Xiao, P. Georgiou, B. Baucom, and S.S. Narayanan, "Head motion modeling for human behavior analysis in dyadic interaction," *IEEE transactions on multimedia*, vol. 17, no. 7, pp. 1107–1119, July 2015.

[4] S. Jha and C. Busso, "Analyzing the relationship between head pose and gaze to model driver visual attention," in *IEEE International Conference on Intelligent Transportation Systems (ITSC 2016)*, Rio de Janeiro, Brazil, November 2016, pp. 2157–2162.

[5] A. Doshi and M.M. Trivedi, "On the roles of eye gaze and head dynamics in predicting driver's intent to change lanes," *IEEE Transactions on Intelligent Transportation Systems*, vol. 3, no. 10, pp. 453–462, September 2009.

[6] C. Breazeal, "Toward sociable robots," *Robotics and autonomous systems*, vol. 42, no. 3-4, pp. 167–175, March 2003.

[7] X. Yu, J. Huang, S. Zhang, W. Yan, and D. N. Metaxas, "Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model," in *IEEE International Conference on Computer Vision (ICCV 2013)*, Sydney, NSW, Australia, December 2013, pp. 1944–1951.

[8] S. O. Ba and J. M. Odobez, "Evaluation of multiple cue head pose estimation algorithms in natural environements," in *IEEE International Conference on Multimedia and Expo (ICME 2005)*, Amsterdam, The Netherlands, July 2005, pp. 1330–1333.

[9] L.P. Morency, A. Rahimi, and T. Darrell, "Adaptive view-based appearance models," in *IEEE Computer Vision and Pattern Recognition (CVPR 2003)*, Madison, WI, USA, June 2003, vol. 1, pp. 803–810.

[10] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J.N. Chang, S. Lee, and S.S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, December 2008.

[11] S. Jha and C. Busso, "Challenges in head pose estimation of drivers in naturalistic recordings using existing tools," in *IEEE International Conference on Intelligent Transportation (ITSC)*, Yokohama, Japan, October 2017, pp. 1624–1629.

[12] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 607–626, April 2009.

[13] B. Czupryński and A. Strupczewski, "High accuracy head pose tracking survey," in *International Conference on Active Media Technology (AMT 2014)*, D. Ślzak, G. Schaefer, S.T. Vuong, and Y.S. Kim, Eds., vol. 8610 of *Lecture Notes in Computer Science*, pp. 407–420. Springer Berlin Heidelberg, Warsaw, Poland, August 2014.

[14] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2013)*, Portland, OR, USA, June 2013, pp. 532–539.

[15] T. Baltrušaitis, P. Robinson, and L. P. Morency, "Constrained local neural fields for robust facial landmark detection in the wild," in *IEEE International Conference on Computer Vision Workshops (ICCVW 2013)*, Sydney, Australia, December 2013, pp. 354–361.

[16] L.A. Jeni, J. F. Cohn, and T. Kanade, "Dense 3d face alignment from 2d videos in real-time," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2015)*, Ljubljana, Slovenia, May 2015, pp. 1–8.

[17] A. Schwarz, M. Haurilet, M. Martinez, and R. Stiefelhagen, "DriveAHead - a large-scale driver head pose dataset," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2017)*, Honolulu, HI, USA, July 2017, pp. 1165–1174.

[18] Q. Wan, Y. Li, C. Li, and R. Pal, "Gesture recognition for smart home applications using portable radar sensors," in *IEEE Engineering in Medicine and Biology Society (EMBC 2014)*, Chicago, IL, USA, August 2014, pp. 6414–6417.

[19] A. Das, I. Tashev, and S. Mohammed, "Ultrasound based gesture recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, New Orleans, LA, USA, March 2017, pp. 406–410.

[20] N. Gourier, D. Hall, and J.L. Crowley, "Estimating face orientation from robust detection of salient facial structures," in *ICPR International Workshop on Visual Observation of Deictic Gestures*, Cambridge, UK, August 2004, pp. 1–9.

[21] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Bosphorus database for 3D face analysis," in *Biometrics and Identity Management (BioID 2008)*, B. Schouten, N.C. Juul, A. Drygajlo, and M. Tistarelli, Eds., vol. 5372 of *Lecture Notes in Computer Science*, pp. 47–56. Springer Berlin Heidelberg, Roskilde, Denmark, May 2008.

[22] R. Rae and H.J. Ritter, "Recognition of human head orientation based on artificial neural networks," *IEEE Transactions on Neural Networks*, vol. 9, no. 2, pp. 257–265, March 1998.

[23] M. Ariz, J. Bengoechea, A. Villanueva, and R. Cabeza, "A novel 2D/3D database with automatic face annotation for head tracking and pose estimation," *Computer Vision and Image Understanding*, vol. 148, pp. 201–210, July 2016.

[24] A. Tawari, S. Martin, and M. M. Trivedi, "Continuous head movement estimator for driver assistance: Issues, algorithms, and on-road evaluations," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 2, pp. 818–830, April 2014.

[25] N. Sadoughi, Y. Liu, and C. Busso, "MSP-AVATAR corpus: Motion capture recordings to study the role of discourse functions in the design of intelligent virtual agents," in *1st International Workshop on Understanding Human Activities through 3D Sensors (UHA3DS 2015)*, Ljubljana, Slovenia, May 2015, pp. 1–6.

[26] E. Murphy-Chutorian, A. Doshi, and M.M. Trivedi, "Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation," in *IEEE Intelligent Transportation Systems Conference (ITSC 2007)*, Seattle, WA, USA, September-October 2007, pp. 709–714.

[27] H. Kato and M. Billinghurst, "Marker tracking and HMD calibration for a video-based augmented reality conferencing system," in *IEEE/ACM International Workshop on Augmented Reality (IWAR 1999)*, San Francisco, CA, USA, August 1999, pp. 85–94.

[28] M. Fiala, "Artag, a fiducial marker system using digital techniques," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005)*, San Diego, CA, USA, June 2005, vol. 2, pp. 590–596.

[29] E. Olson, "AprilTag: A robust and flexible visual fiducial system," in *IEEE International Conference on Robotics and Automation (ICRA 2011)*, Shanghai, China, May 2011, pp. 3400–3407.

[30] W. Kabsch, "A discussion of the solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A.*, vol. A34, no. Part 5, pp. 827–828, September 1978.

[31] D.Q. Huynh, "Metrics for 3D rotations: Comparison and analysis," *Journal of Mathematical Imaging and Vision*, vol. 35, no. 2, pp. 155–164, June 2009.