DD-Pose - A large-scale Driver Head Pose Benchmark

Markus Roth[1,2,†] and Dariu M. Gavrila[2,‡]

***Abstract*— We introduce *DD-Pose*, the Daimler TU Delft Driver Head Pose Benchmark, a large-scale and diverse benchmark for image-based head pose estimation and driver analysis. It contains 330k measurements from multiple cameras acquired by an in-car setup during naturalistic drives. Large out-of-plane head rotations and occlusions are induced by complex driving scenarios, such as parking and driver-pedestrian interactions. Precise head pose annotations are obtained by a motion capture sensor and a novel calibration device. A high resolution stereo driver camera is supplemented by a camera capturing the driver cabin. Together with steering wheel and vehicle motion information, *DD-Pose* paves the way for holistic driver analysis.**

Our experiments show that the new dataset offers a broad distribution of head poses, comprising an order of magnitude more samples of rare poses than a comparable dataset. By an analysis of a state-of-the-art head pose estimation method, we demonstrate the challenges offered by the benchmark.

The dataset and evaluation code are made freely available to academic and non-profit institutions for non-commercial benchmarking purposes.

I. Introduction

Visual head pose estimation plays an essential role in human understanding, as it is our natural cue for inferring focus of attention, awareness and intention. For machine vision, the task is to estimate position and orientation of the head from images.

A wide range of uses exists for head pose estimation, either directly or for derived tasks such as gaze estimation, facial identification and expression analysis, when considering natural human-machine interfaces, augmented reality, surveillance and automotive applications. In the automotive domain, there are applications for driver convenience, safety, and conditional automation. For convenience functions, head pose can be used for virtual volumetric head-up displays (HUD), auto-stereoscopic 3D displays and multimodal human-car interfaces. Inferring a driver's pose can benefit in safety applications, as it enables estimation of distraction, intention, sleepiness and awareness. When taking the vehicle's surrounding into consideration, mutual gaze with vulnerable road users (VRU) is of high interest for warning and automatic braking systems [1]. SAE level 3 (conditional automation) involves a possible take-over request to the driver for a transition from autonomous to manual driving mode. Currently, the driver's ability to service such request is maintained by requiring the driver to touch the steering wheel periodically. This could be replaced by a less obnoxious driver awareness recognition system.

Benchmarks (i.e. datasets and evaluation metrics) play a crucial role in developing and evaluating robust head pose estimation methods. A good benchmark not only allows to identify the challenges of a task, but also enables the development of better methods for solving it. An in-car head pose dataset provides difficult illumination conditions, occlusions and extreme head poses. The recent popularity of deep learning methods with their large model complexity stresses the demand for a large dataset [2].

Fig. 1: *DD-Pose* provides precise 6 DOF head pose annotation for 330k stereo image pairs acquired in an in-car environment. The benchmark offers significant out-of-plane rotations and occlusions from naturalistic behavior introduced by complex driving scenarios. Annotations for partial and full occlusions are available for each high resolution driver camera image. An additional camera capturing the interior of the car allows for further multi-sensor driver analysis tasks.

Available head pose datasets have drawbacks in terms of size, annotation accuracy, resolution and diversity (see Table I). To close this gap, we present *DD-Pose*, a large-scale benchmark composed of 330k images from a high resolution stereo driver camera with precise 6 degrees of freedom (DOF) head pose annotations. *DD-Pose* includes a variety of non-frontal poses and occlusions occurring in complex driving scenarios. To extend its use from head pose estimation to more holistic driver analysis, we also supply images from a wide-angle interior camera and vehicle-data, such as velocity, yaw rate and steering wheel angle.

Sample annotations of the benchmark can be found in Figure 1.

II. Related Work

There is an abundance of publicly available image-based head pose datasets dating back nearly two decades [3]–[14] (see Table I).

† markus.r.roth@daimler.com
‡ d.m.gavrila@tudelft.nl
[1] Environment Perception, Daimler AG, Stuttgart, Germany
[2] Intelligent Vehicles, Technical University Delft, The Netherlands

Head pose datasets can be categorized by different aspects, such as *imaging characteristics*, *data diversity*, *acquisition scenario*, *annotation type*, and *annotation technique*. These aspects play an important role on whether and how the dataset identifies challenges of the head pose estimation task.

Imaging characteristics relate to the image resolution, number of cameras, bit depth, frame rate, modality (RGB, grayscale, depth, infrared), geometric setup and field of view.

Data diversity incorporates aspects such as the number of subjects, the distribution of age, gender, ethnicity, facial expressions, occlusions (e.g. glasses, hands, facial hair) and head pose angles. Data diversity is essential to training and evaluating robust estimation models.

Acquisition scenario covers the circumstances under which the acquisition of the head pose takes place. The most important distinction is between in-laboratory [4,7,8,11]–[13] vs. in-the-wild [5,6,9,10,14] acquisition. While the former restricts the data by defining a rather well-defined, static environment, the latter offers more variety through being acquired in unconstrained environments such as outside, thus covering many challenges like differing illumination and variable background. Head movement can be *staged* by following a predefined trajectory or can be *naturalistic* by capturing head movement while the subject performs a different task, such as driving a car.

Annotation type describes what meta-information, such as head pose, comes alongside the image data and how it is represented. *Head pose* is defined by a full 6 degrees of freedom (DOF) transformation from the camera coordinate system to the head coordinate system, covering 3 DOF for translation and 3 DOF in rotation. Head pose datasets differ in how many of those DOFs are provided alongside the images, i.e. whether only a subset of the translation and rotation parameters is given. Ultimately, annotation types differ in their granularity of sampling the DOF space: there are discrete annotation types which classify a finite set of head poses, and there are continuous annotation types which offer head pose annotations on a continuous scale for all DOFs.

There are different *annotation techniques* for obtaining the head pose annotation accompanying each image. The annotation technique has a large impact on data quality. It can be categorized into manual annotations vs. automatic annotations. For manual annotations, human experts annotate the image data according to a label specification [4]. Automatic annotations can be divided into data-based annotations, computed by algorithms on the image data [8,13], and sensor-based annotations, which in turn use an additional hardware sensor for obtaining the head pose for each image [7,11,14].

Manual annotations do not need additional hardware, but are prone to introduce errors and biases. E.g. a human annotator can only annotate in the image plane, thus needing to guess the distance part of the translation of the head [6,10]. There is also inter-annotator variability through different interpretation of the same scene. Additionally, as manual annotations consume human time, its cost scales linearly with the amount of to data to be annotated.

Automatic annotations based on algorithms computing the annotations from the image data are fast to obtain, but induce systematic errors of the underlying algorithm and will not allow to disambiguate between annotation errors and errors induced by the method under test.

Automatic annotations based on sensors make use of additional reference sensors during the data acquisition process. The reference sensor measurements should be calibrated to the head coordinate system and calibrated and synchronized to the camera images. There are different types of reference sensors which differ in their measurement method. Among those are electromagnetic sensors [7,11], inertial sensors, vision-based sensors, 3D scanners [4], optical marker tracking sensors [14], and hybrid combinations of them. An optimal reference sensor for head pose estimation should be accurate, free of drift, robust to disturbance, and measure all 6 DOFs on a continuous scale.

From the aspects mentioned above, we focus on datasets with continuous head pose annotations for all 6 DOF which offer naturalistic scenarios and a large data diversity.

Many recent models for classification and regression tasks are based on deep convolutional neural networks [2]. Their high model complexity demands for a very large number of training examples. Therefore, we also focus on large datasets in terms of number of images.

An overview of currently available datasets is given in Table I. Respective example data can be found in Figure 2.

We subdivide the datasets by their acquisition scenario into two groups, namely generic head pose datasets vs. driving head pose datasets. The latter come with desirable properties such as naturalistic scenarios, a large data diversity and challenging imaging characteristics.

A. Generic Head Pose Datasets

Bosphorus [4] contains 5k high resolution face scans from 105 different subjects. The 3D scans are obtained by a commercial structured-light based 3D digitizer. It offers 13 discrete head pose annotations and with different facial expressions and occlusions.

ICT-3DHP [7] provides 1400 images and depth data from 10 subjects acquired with a Kinect v1 sensor. 6 DOF head pose annotations are measured by a magnetic reference sensor. The authors do not detail on whether calibration and synchronization of the reference sensor measurements to the camera images is performed.

Biwi Kinect [8] consists of 16k VGA images and depth data from 20 subjects depicting the upper body. The data was acquired by a Kinect v1 sensor. 6 DOF head pose annotations are provided by fitting user-specific 3D templates on depth data, which has limitations when occlusions are present. As it is recorded in a laboratory environment, it provides a uniform and static background.

gi4e hpdb [11] contains 36k images from 10 subjects recorded with a webcam in an in-laboratory environment. Head pose annotations are given in 6 DOF using a magnetic reference sensor. All transformations and camera intrinsics are provided. Head pose annotations are given relative to an initial subjective frontal pose of the subject.

SynHead [12] contains 511k synthetic images from 10 head models and 70 motion tracks. The rendered head models are composed with random background images, providing

TABLE I: 2D/3D face datasets with continuous head pose annotations.

Dataset	GT	Year	#Cams x w x h	#Images	#Subjects f/m	Head pose	Reference	Scenarios
Bosphorus [4]	3D	2008	1x1600x1200	5k	45/60	relative	guided	choreographed facial expressions
ICT-3DHP [7]	3D	2012	1x640x480	1k	6/4	relative	magnetic	choreographed large rotations
Biwi Kinect [8]	3D	2013	1x640x480	16k	6/14	relative	guided, ICP	choreographed large rotations (yaw, pitch)
gi4e hpdb [11]	2D	2016	1x1280x720	36k	4/6	relative	magnetic	choreographed large rotations
SynHead [12]	3D	2017	1x400x400	511k	5/5	absolute	synthetic data	70 different motion tracks
UbiPose [13]	3D	2018	1x1920x1080	14k	22[c]	absolute	3DMM	service desk interactions
RS-DMV [5]	2D	2010	1x960x480	13k	6[c]	N/A	N/A	naturalistic driving
Lisa-P [6]	2D	2012	1x640x480	200k	14[c]	relative	POS [15]	naturalistic driving, choreographed large yaw
NDS HPV [9]	2D	2015	1x720x480	2PB[d]	>3100[c]	N/A	N/A	naturalistic driving
VIVA [10]	2D	2016	1x*544	1k	N/A	relative	POS [15]	naturalistic driving
DriveAHead [14]	3D	2018	1x512x424[a]	1M	4/16	absolute	mo-cap	naturalistic driving, parking
DD-Pose (ours)[b]	3D	2019	2x2048x2048	2x330k	6/21	absolute	mo-cap	naturalistic driving, large rotations and translations

[a] only head image crops provided. Mean size 25x50
[b] additional data streams recorded: front facing camera, interior camera facing driver from the rear right
[c] female/male ratio not provided by the authors
[d] number of images not provided. Assumed to be $>10^9$

indoor/office scenery. As this is a generative method for data synthesis, head pose annotations are very accurate. Making use of 10 head models provides little diversity of human facial expressions.

UbiPose [13] features natural role played interactions with 10k frames obtained by a Kinect v2 sensor. 22 subjects are recorded. Head pose was annotated automatically based on the raw footage using initial facial landmark annotations and fitting a 3D morphable model. Annotations not fitting the data were pruned by human annotators. Subjects were captured from a relatively large distance.

B. Driving Head Pose Datasets

RS-DMV [5] contains 13k images from 6 subjects captured in naturalistic outdoor and simulator scenarios. Head pose annotations are not provided.

Lisa-P [6] offers 200k images from 14 subjects with a resolution of 640x480. Head orientation annotations are obtained by using the Pose from Orthography and Scaling (POS) algorithm [15] on manually labeled facial landmarks. By using an orthographic projection, this approach only allows for approximate position and orientation estimates.

NDS-HPV [9] contains 2PB of highly compressed, low resolution images from a naturalistic driving study. It contains images of over 3100 subjects collected over a period of over 2 years. Head pose annotations are not provided, thus restricting its use to qualitative analysis.

The VIVA head pose estimation benchmark [10] is a test set consisting of images with 607 faces, out of which 323 are partially occluded. The naturalistic driving images were selected both from research vehicle recordings and YouTube videos to display harsh lighting conditions and facial occlusions. The head pose annotations of the test dataset are not released, but evaluation is possible by submitting hypotheses through a benchmark website. No training images are provided.

DriveAHead [14] is the nearest neighbor of our proposed benchmark. It features 1M images and depth information acquired by a Kinect v2 sensor during naturalistic driving. 20 different subjects appear in the recordings. Images were collected with a resolution of 512x424 pixels. 6 DOF continuous head pose annotations are obtained by a motion capture system which measures the pose of a marker fixated at the back of the subject's head. The coordinate transformation between the head mounted marker coordinate system and the head coordinate system is calibrated per-subject by measuring the position of 8 facial landmarks of the face of each subject after fixating the head-mounted marker. The transformation between the reference sensor coordinate system and the camera coordinate systems are known, although the calibration process is not described. Alongside, per-image annotations for occlusions and whether the subjects wears glasses or sunglasses is provided.

The large number of image samples enables training of deep convolutional neural networks for head pose estimation. Parking maneuvers and driving on a highway and through a small town results in naturalistic head movements, thus providing distributions of head orientation angles and head positions which are typical for naturalistic drives.

As no intrinsic camera parameters are provided, 3D points in the camera coordinate system cannot be projected into the image space. Consequently, both head position and orientation estimation methods have to implicitly adapt to the specific dataset. DriveAHead provides cut-outs of faces with a mean inter-pupil distance of 35 pixels, thus targeting on methods for low-resolution head pose estimation.

III. DD-Pose - A large-scale Driver Head Pose Benchmark

We introduce *DD-Pose*[1], a large scale head pose benchmark featuring driver camera images acquired during complex naturalistic driving scenarios. The proposed benchmark provides 330k high resolution images from 27 subjects with precise continuous 6 DOF head position and orientation annotations. Occlusions from steering wheel, hands, and accessories such glasses or sunglasses are present and manually annotated as such on a per-frame basis.

High resolution images of the driver's head are acquired by a stereo camera setup mounted behind the steering wheel. Continuous frame-wise head pose is obtained by a optical marker tracker measuring the 6 DOF pose of a marker fixated on the back of each subject's head. We find the per-subject transformation from the head mounted marker to the head coordinate system by a novel calibration device.

[1] Available at https://dd-pose-dataset.tudelft.nl

(a) Bosphorus [4] (b) Biwi Kinect [8]

(c) ICT-3DHP [7] (d) gi4e hpdb [11] (e) SynHead [12]

(f) UbiPose [13] (g) RS-DMV [5] (h) Lisa-P [6]

(i) NDS HPV [9] (j) VIVA [10]

(k) DriveAHead [14]. Content within red box provided.

Fig. 2: Example data of the investigated 2D/3D head pose datasets. The datasets differ in many aspects, such as sensor modalities (RGB, IR, depth), in-lab vs. synthetic vs. naturalistic driving, precision of head pose annotation and resolution.

In addition to the driver stereo camera, the proposed setup uses a wide angle RGB camera depicting the driver from the rear side to allow for upper-body analysis of the driver action. Vehicle parameters such as steering wheel angle, velocity and yaw rate are also part of the benchmark.

All sensors are calibrated intrinsically and extrinsically, such that the coordinate transformations between their coordinate systems are known. Depth information can be extracted from the provided stereo camera images by using a disparity estimation algorithm, e.g. semi-global matching [16]. The optical marker tracker and the stereo driver camera are electrically synchronized, resulting in a head pose measurement free of drift and latency.

DD-Pose offers a broad distribution of poses and challenging lighting conditions like dark nighttime driving, tunnel entrances/exits and low standing sun. 12 driving scenarios were conducted to gain highly variant, yet naturalistic images of the driver. 9 driving scenarios comprise drives through a big German city with lane merges, complex roundabouts, parking, and pedestrian zones with pedestrian interactions. In addition to driving scenarios, we provide 3 standstill scenarios covering a broad range of head poses and a scenario with mobile phone use.

Overall, *DD-Pose* offers a variety of naturalistic driving data which we believe is crucial for development and evaluation of head pose estimation algorithms in unconstrained environments. With 4 megapixels per camera and a mean inter-pupil distance of 274px, *DD-Pose* offers around 60 times more face pixels than DriveAHead to extract features from fine-grained face structures such as eye gaze and evaluate whether high resolution is a benefit to the methods under test.

A. Contributions

Our contributions by supplying *DD-Pose* to the scientific community are manyfold: (a) the driver analysis benchmark from naturalistic driving scenarios features a broad distribution of head orientations and positions with an order of magnitude more samples of rare poses than comparable datasets (see Figures 5 and 6), (b) the high resolution stereo images allow for analysis of resolution, depth, and taking image context around faces into account, (c) the supplemental camera of the driver cabin, combined with steering wheel and vehicle motion information, pave the way for holistic driver analysis, rather than head pose only.

Example data of the proposed benchmark is shown in Figure 1.

B. Scenarios

The definition of driving scenarios has an essential impact on the distribution of the head pose and textural variability of the data. E.g., a drive along the highway would be very biased towards a frontal pose and not be beneficial to train and evaluate head pose estimation methods. We favor non-frontal poses by implicitly forcing the driver have to look out of the car, e.g. by interacting with pedestrians in a pedestrian zone, and instructing the driver to read shop names on the side of the street. Yet, to be representative of naturalistic drives, we included scenarios of standard traffic manoeuvres, such as passing zebra crossings, highway merges, roundabouts, parking and turning the vehicle. To provide more extensive poses, scenarios while standing are included, where the driver is instructed to fixate his or her gaze on predefined locations within the car, forcing large head rotations and translations, and making a phone call.

The scenarios of *DD-Pose* are defined in Table II, alongside with their intended properties on data variability.

For the in-car gaze fixation scenario (Table II, #9) we define the following protocol: the car stands still with the steering wheel in straight position. The subject is asked to turn the head to point at a predefined set of targets in the car. A button is to be pressed by the subject for the period he or she is fixating the object, thus annotating the time stamps of fixation ad-hoc. Among the targets are mirrors, in-car buttons and displays.

In summary, these carefully-chosen scenario definitions result in a large variance in head rotation and head translation, but also facial expressions.

C. Hardware Setup and Coordinate Systems

We equipped a research vehicle with a stereo camera facing the driver (each 2048x2048 px, 16 bit, IR sensitive). It is mounted near the speedometer. An infra-red LED illuminates the driver. A wide angle interior camera (RGB)

#	Description	Rot	Trans	Occl	Stw Occl	Facial ex	Illum var	Ped inter	Remark
0	generic driving	low	low	low	med	high	med	low	talking
1	zebra crossing	low	low	low	low	med	med	high	crossings and bus stops
2	merge	high	med	low	low	med	med	low	mirrors, look over shoulder
3	tunnel	low	low	low	low	med	high	low	entrance, exit
4	roundabout	high	med	low	low	med	med	low	also multi-lane roundabout
5	ped zone	high	med	low	high	med	med	high	incl. two-step turn
6	intentional occl	med	med	high	med	high	med	low	occlusions, facial expressions
7	shop name reading	high	med	med	low	high	med	high	shops left and right
8	parking	high	high	med	high	high	med	med	parking in
9	in-car fixation	high	med	med	no	high	med	low	no driving
10	large translations	med	high	med	no	med	med	low	no driving
11	large rotations	high	med	med	no	med	med	low	no driving
12	hand-held calling	high	med	high	no	med	med	low	no driving

Rot: rotation; Trans: translation; Occl: occlusion; Stw occl: steering wheel occlusions;
Facial ex: facial expressions; Illum var: illumination variance; Ped inter: pedestrian interaction.

TABLE II: Driving scenario definitions and the resulting features of the proposed benchmark. 12 scenarios are defined to implicitly enforce a broad distribution of head poses and texture.

captures the driver's cabin from the rear side. We mounted an optical marker tracker on the rear right behind the driver. The optical marker tracker can measure the 6 DOF pose of a marker consisting of multiple IR retroreflective spheres. The subject wears such a marker on the back of his or her head, which is fixated using a rubber band.

The driver stereo camera, LED illumination and optical marker tracker are electrically triggered at 15 Hz. The other sensors are synchronized.

We designed a head calibration device which defines the head coordinate system when attached to the driver's head while being simultaneously being measured by the optical marker tracker.

Each camera, the optical marker tracker, the head mounted marker, the driver's head and the car's chassis define a coordinate system. We define a transformation between two coordinate systems A and B as a homogeneous matrix $T^{A\to B}$ which transforms a homogeneous point p^B into p^A by $p^A = T^{A\to B} \cdot p^B$.

See Figure 3 for a visual overview of the sensors, their coordinate systems and the transformations in between them.

D. Optical Marker Tracker to Driver Camera Calibration

$T^{\text{cam\_driver\_left}\to\text{marker\_tracker}}$ and the camera intrinsic parameters are obtained simultaneously by a calibration routine which makes use of 3D checkerboard corner positions. We obtain the 3D checkerboard corner positions inside the marker tracker coordinate system by attaching retro-reflective spheres to the checkerboard, thus making it a marker measurable by the optical marker tracker. With the 3D checkerboard corner positions and their corresponding 2D projections in the image, a bundle adjustment method is used to optimize intrinsic and extrinsic camera parameters, such as focal lengths, principal points, distortion parameters and rectification parameters [17]. $T^{\text{cam\_driver\_left}\to\text{marker\_tracker}}$ is obtained as a by-product of the optimization.

E. Marker to Head Calibration

We define the head coordinate system as follows. The origin is located in the nasion of the head. The x-axis points in frontal direction. The y-axis points towards the left ear. The z-axis points upwards; it touches the chin centrally. The xz-plane mirrors the head.

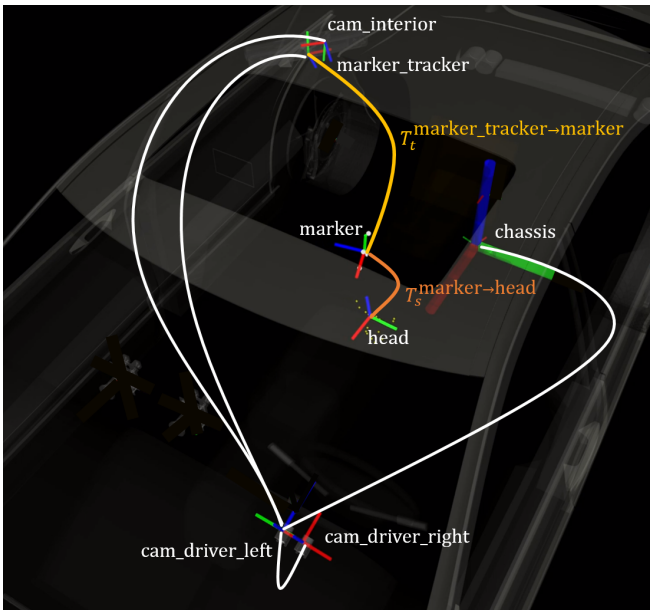

Fig. 3: In-car hardware setup, coordinate systems and transformations. White arcs denote static transformations acquired once during the setup calibration process. The yellow arc denotes the transformation $T_t^{\text{marker\_tracker}\to\text{marker}}$ being measured by the optical marker tracker for each frame at time t. The orange arc denotes the transformation $T_s^{\text{marker}\to\text{head}}$ being calibrated once per subject s. All transformations are provided with *DD-Pose*.

We designed a calibrator to attach to the driver's head during the per-subject calibration process. It provides a notch to touch the nasion. A chin slider is adjusted such that it touches the chin centrally. Two cheek sliders are slid against the head such that they touch the cheeks with equal force, thus defining symmetry about the xz-plane. It is also equipped with retroreflective spheres such that its pose can be measured by the optical marker tracker. Its coordinate system is defined such that it coincides with the head coordinate system above. When it is attached properly, the per-subject transformation between marker and head is then $T_s^{\text{marker}\to\text{head}} := T_t^{\text{marker}\to\text{calibrator}}$. This process has to be performed once per subject and is valid as long as

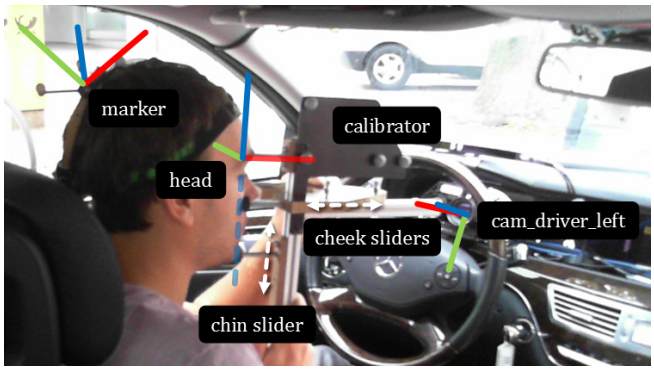

Fig. 4: The per-subject head calibration process. A calibrator whose pose can be measured by the optical marker tracker is attached to the head by touching the nasion and both chin and cheek sliders in proper position.

the marker is fixated at the subject's head. The calibration process is illustrated in Figure 4.

F. Data Preprocessing

Depending on the driver's head pose, the retroreflective spheres of the head-worn marker are visible in the camera image. To avoid models to overfit to these, we remove them. We extend the approach of [14], where the projected locations of the spheres are filled with interpolations of the values of their surroundings. As markers will mostly be hidden behind the subjects' head, we employ a heuristic to only blur the spheres which are likely visible. The heuristic is based on an empirically found range of head poses and conservatively set, i.e. rather fill hair or face border than leave spheres visible.

G. Occlusion Annotations

We manually annotated each driver camera images for its occlusions based on the visibility of facial landmarks, as defined in [18]. **none**: all 68 landmarks visible; **partial**: at least one landmark occluded; **full**: all landmarks occluded.

H. Dataset splits

To allow for a fine-grained evaluation, we split the data into the disjoint subsets *easy*, *moderate*, and *hard* depending on the angular distance of the measured head pose from a frontal pose (looking directly into the driver camera) α_f and the presence of occlusion. **easy**: $\alpha_f \in [0, 35)^\circ \wedge \text{occl} \in \{\text{none}\}$; **moderate**: $(\alpha_f \in [0, 35)^\circ \wedge \text{occl} \in \{\text{partial}\}) \vee (\alpha_f \in [35, 60)^\circ \wedge \text{occl} \in \{\text{none}, \text{partial}\})$; **hard**: $\alpha_f \in [60, \infty)^\circ \vee \text{occl} \in \{\text{full}\}$;

IV. Dataset Analysis

DD-Pose comprises recordings of 27 subjects, of which 21 are male and 6 are female. The average age is 36 years. The youngest and oldest driver are 20 and 64 years old.

There are 330k measurements of the driver stereo image camera along with interior camera images. Head pose measurements are available for 93% of the images. The proportion of the dataset splits is (*easy*, *moderate*, *hard*) = (55%, 33%, 12%).

For the left driver camera images, 5% are fully occluded, 19% are partially occluded (not counting glasses or sun glasses) and 76% have no occlusion. In 41% of the images, the driver wears glasses, in 1% sunglasses.

There are 13 scenarios, out of which 9 are driving scenarios (#0 - #8) and 4 are non-driving scenarios (#9 - #12); see Table II. The shortest scenario (#3, tunnel entrance/exit) is on average 24s long. The longest scenario (#5, pedestrian zone) is on average 211s long.

The mean inter-pupil distance is 274px (cf. DriveAHead: 35px [14]).

The distribution of head orientation angles of *DD-Pose* and DriveAHead [14] is depicted in Figure 5. The angles vary in the following ranges, ignoring outliers with less than 10 measurements in a 3° neighborhood: roll $\in [-63..60]^\circ$; pitch $\in [-69..57]^\circ$; yaw $\in [-138..126]^\circ$. The mean pitch angle is -20°, caused by the driver camera mounted at the speedometer.

The distribution of head position occurrences of *DD-Pose* and DriveAHead [14] is depicted in Figure 6. *DD-Pose* covers a broad volume of head locations.

Overall, *DD-Pose* offers an order of magnitude more data for off-centered head poses than comparable datasets [14].

V. Evaluation

To show that the proposed benchmark contains challenging imagery, we evaluate the performance of two head pose estimation methods on it.

A. Head Pose Estimation Methods

One method is the head pose prior, which always assumes the head to be present in the mean pose obtained from the dataset. The second method performs head pose estimation by localizing facial landmarks and solving the Perspective-n-Point (PnP) Problem.

Prior: on a dataset with a large amount of frontal poses, this method is expected to perform very well, despite performing bad on rare poses. The mean head position of *DD-Pose* wrt. the camera is $\bar{t} = (0.011\text{m}, 0.006\text{m}, 0.608\text{m})$. The mean rotation is $yaw = -6.6^\circ$, $pitch = -20.1^\circ$, $roll = 0.7^\circ$.

OpenFace 2.0: the second method we evaluate is OpenFace 2.0 [19], a state-of-the-art face analysis toolkit. Head pose estimation is performed by localization of facial landmarks via Convolutional Experts Constrained Local Model (CE-CLM). The facial landmarks are assigned to a 3D landmark model in head coordinates. The pose is found via solving the Perspective-n-Point (PnP) problem, i.e. finding the pose of the head coordinate system with respect to the camera coordinate system which minimizes the projection error. We use the pretrained models from the authors [19], but transform the pose such that it fits the head coordinate system defined above. The model uses multi-view initialization to account for extreme poses.

B. Evaluation Metrics

Evaluation metrics play an important role on evaluating the performance of the methods for the specific task. The task of head pose estimation is evaluated for position and orientation separately.

Recall: recall defines on which percentage of the images a head hypothesis from head pose estimation method exists.

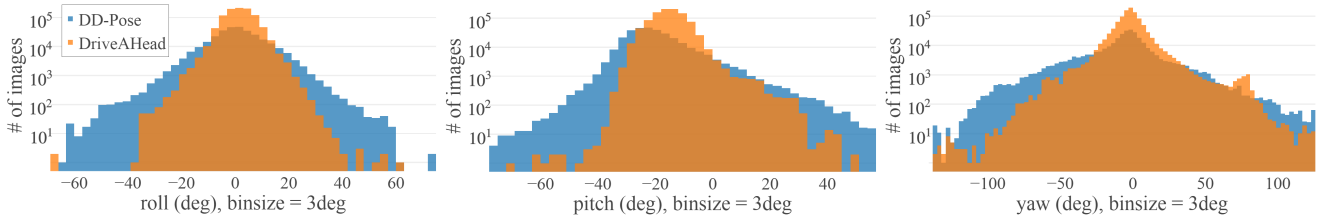

Fig. 5: Distribution of head orientation angles of the proposed benchmark *DD-Pose* and DriveAHead [14] with respect to a frontal pose into the camera. While both datasets cover a broad range of orientations, *DD-Pose* supplies an order of magnitude more data for non-frontal head orientations.

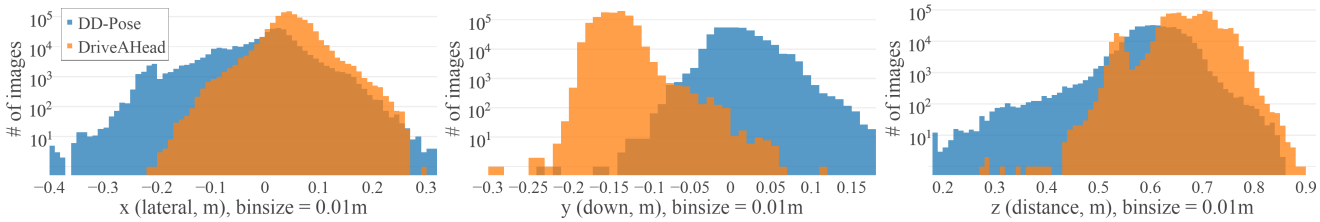

Fig. 6: Distribution of head positions of *DD-Pose* and DriveAHead [14] in the camera coordinate system. Although the action volume of the driver is limited in the driver's seat, the datasets differ in their position distribution. *DD-Pose* covers a larger lateral space, is unbiased in y-direction and also depicts very nearby heads.

Images without a hypothesis are left out when evaluating position and orientation.

Position: we evaluate the mean Euclidean distance for each axis and the Euclidean distance between ground truth head origin and hypothesis head origin.

Orientation: the commonly used metric *mean angular error (MAE)* can be performed on each of the three rotation angles separately or by computing a single rotation angle between ground truth and hypotheses. In both cases, outliers will have a small weight on biased datasets, e.g. with many frontal poses and a few extreme poses. For an unbiased evaluation of head rotation, we use *balanced mean angular error (BMAE)* introduced in [14]. It splits the dataset in bins based on the angular difference from the frontal pose and averages the MAE of each of the bins:

$$\mathrm{BMAE}_{d,k} := \frac{d}{k} \sum_i \phi_{i,i+d}, i \in d\mathbb{N} \cap [0,k]$$

where $\phi_{i,i+d}$ is the MAE of all hypotheses, where the angular difference between ground truth and frontal pose is between i and $i+d$. During evaluation, we use bin size $d := 5°$ and maximum angle $k := 75°$.

C. Recall

The prior method, by construction, has a recall of 1.0. The recall of OpenFace 2.0 on the whole dataset is 0.76 and for the subsets (*easy*, *moderate*, *hard*) = (0.95, 0.65, 0.16). A more fine grained analysis on the recall value depending on the angular distance from the frontal pose is found in Figure 7. One can see the influence of the definition of the subsets. While the *easy* subset offers a large recall as it covers unoccluded heads with angles up to 35°, the *moderate* subset covers the partial occlusions in this range with a lower recall. The overall recall drops with increasing angle.

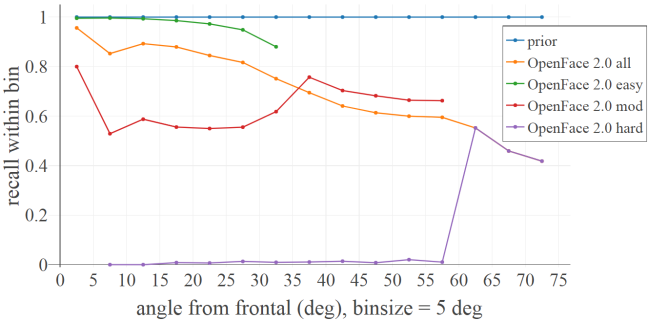

Fig. 7: Recall depending on angular difference from frontal pose. The recall of OpenFace 2.0 on the whole dataset drops with increasing rotation from the frontal pose.

Subset	Prior				OpenFace 2.0			
	x	y	z	L_2	x	y	z	L_2
all	40	21	36	66	8	8	41	44
easy	23	19	32	49	5	6	31	33
moderate	54	21	38	78	12	10	58	63
hard	83	27	46	107	44	30	134	148

TABLE III: Position errors (mm). Errors along all axes and Euclidean Distance L_2 for the subsets.

D. Position

The errors in head position estimation are listed in Table III. The errors on the prior method implicitly denote statistics of the distribution of the subsets. The L_2 error increases from 5cm to 11cm from the easy to the hard subset, caused by a larger position variance around the mean position in the measurements. OpenFace 2.0 localizes the head position in x and y direction for the *easy* and *moderate* subsets within 1cm, increasing up to 4cm for the *hard* subset. OpenFace 2.0 has approximately 4-5 times larger errors in z direction than for the other two dimensions.

Subset	Prior		OpenFace 2.0				
	MAE	BMAE	MAE	BMAE	roll	pitch	yaw
all	20	32	9	16	5	4	4
easy	11	14	5	5	3	3	2
moderate	27	26	14	13	8	6	8
hard	45	34	33	31	13	9	27

TABLE IV: Overall mean angular errors (MAE) and balanced mean angular errors ($BMAE_{5,75}$) in degrees of the head pose estimation methods for the subsets; MAE for roll, pitch, yaw of OpenFace 2.0 (deg).

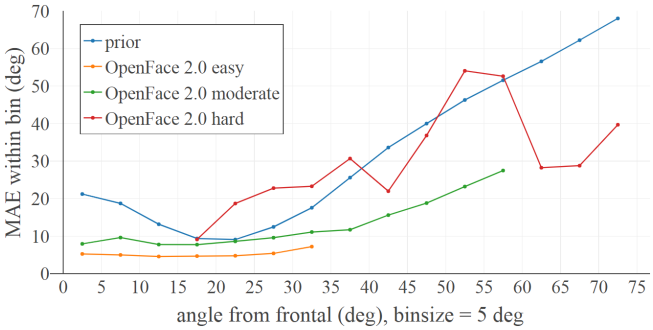

Fig. 8: Mean angular errors (MAE). All methods increase in terms of MAE for more extreme poses.

E. Orientation

An overview of the mean angular errors (MAE) and balanced mean angular errors (BMAE) of the methods on *DD-Pose* is given in Table IV. Figure 8 depicts the MAE depending on the angular difference from a frontal pose.

The prior method implicitly denotes statistics on the orientation measurement distribution around the mean orientation. The MAE increases from 11° to 45° from the easy subset to the hard subset, showing the increasing variance for the more difficult subset.

The MAE of OpenFace 2.0 ranges from 5° on the *easy* subset to 33° on the *hard* subset, i.e. the error increases by more than a factor of 6 when facing more challenging poses and occlusions. For comparison: the reported MAE of OpenFace 2.0 is 2.6° on the BU dataset [3] and 3.2° on the ICT-3DHP dataset [19].

VI. Conclusions

In this paper, we introduced *DD-Pose*, a large-scale driver head pose benchmark featuring multi-camera images of 27 drivers captured during 12 naturalistic driving scenarios. The benchmark contains 330k frames with high resolution stereo images from a driver camera, accompanied by an interior camera and driving meta data such as velocity and yaw rate. It provides per-frame head pose measurements and occlusion annotations. Precise head pose is measured by a novel calibration device. All sensors are fully-calibrated and synchronized.

The experiments showed, that *DD-Pose* provides challenges for a current state-of-the-art method due to its richness in extreme non-frontal head poses.

We therefore recommend *DD-Pose* for training and benchmarking of head pose estimation methods which have to perform robustly in challenging conditions.

References

[1] M. Roth, F. Flohr, and D. M. Gavrila, "Driver and pedestrian awareness-based collision risk analysis," in *2016 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 6 2016, pp. 454–459.

[2] M. Braun, S. Krebs, F. Flohr, and D. Gavrila, "EuroCity Persons: A Novel Benchmark for Person Detection in Traffic Scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 5 2019.

[3] M. La Cascia, S. Sclaroff, and V. Athitsos, "Fast, reliable head tracking under varying illumination: an approach based on registration of texture-mapped 3D models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 4, pp. 322–336, 4 2000.

[4] A. Savran, N. Alyüz, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Bosphorus Database for 3D Face Analysis," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2008, pp. 47–56.

[5] J. Nuevo, L. M. Bergasa, and P. Jiménez, "RSMAT: Robust simultaneous modeling and tracking," *Pattern Recognition Letters*, vol. 31, no. 16, pp. 2455–2463, 12 2010.

[6] S. Martin, A. Tawari, E. Murphy-Chutorian, S. Y. Cheng, and M. Trivedi, "On the design and evaluation of robust head pose for visual user interfaces," in *Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications - AutomotiveUI '12*. New York, New York, USA: ACM Press, 2012, p. 149.

[7] T. Baltrusaitis, P. Robinson, and L. P. Morency, "3D Constrained Local Model for rigid and non-rigid facial tracking," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2012.

[8] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool, "Random Forests for Real Time 3D Face Analysis," *International Journal of Computer Vision*, 2013.

[9] J. Paone, D. Bolme, R. Ferrell, D. Aykac, and T. Karnowski, "Baseline face detection, head pose estimation, and coarse direction detection for facial data in the SHRP2 naturalistic driving study," in *2015 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 6 2015, pp. 174–179.

[10] S. Martin, K. Yuen, and M. M. Trivedi, "Vision for Intelligent Vehicles & Applications (VIVA): Face detection and head pose challenge," in *2016 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 6 2016, pp. 1010–1014.

[11] M. Ariz, J. 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, 7 2016.

[12] J. Gu, X. Yang, S. De Mello, and J. Kautz, "Dynamic facial analysis: From Bayesian filtering to recurrent neural network," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017.

[13] Y. Yu, K. Funes Mora, and J.-M. Odobez, "HeadFusion: 360° Head Pose tracking combining 3D Morphable Model and 3D Reconstruction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2018.

[14] A. Schwarz, M. Haurilet, M. Martinez, and R. Stiefelhagen, "DriveAHead A Large-Scale Driver Head Pose Dataset," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, vol. 2017-July. IEEE, 7 2017, pp. 1165–1174.

[15] D. F. DeMenthon and L. S. Davis, "Model-based object pose in 25 lines of code," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 1992, pp. 335–343.

[16] H. Hirschmüller, "Stereo Processing by Semi-Global Matching and Mutual Information," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 328–341, 2008.

[17] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000.

[18] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark Localization Challenge," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013.

[19] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "OpenFace 2.0: Facial Behavior Analysis Toolkit," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 5 2018, pp. 59–66.