AutoPOSE: Large-scale Automotive Driver Head Pose and Gaze Dataset with Deep Head Orientation Baseline

Mohamed Selim[1], Ahmet Firintepe[2], Alain Pagani[1] and Didier Stricker[1]
[1]*German Research Center for Artificial Intelligence (DFKI), Trippstadter Str. 122, Kaiserslautern, Germany*
[2]*BMW Group, Munich, Germany*

https://autopose.dfki.de

Keywords: Driving, Head Pose Estimation, Deep Learning, Infrared Camera, Kinect V2, Eye Gaze.

Abstract: In computer vision research, public datasets are crucial to objectively assess new algorithms. By the wide use of deep learning methods to solve computer vision problems, large-scale datasets are indispensable for proper network training. Various driver-centered analysis depend on accurate head pose and gaze estimation. In this paper, we present a new large-scale dataset, AutoPOSE. The dataset provides $\sim$ 1.1M IR images taken from the dashboard view, and $\sim$ 315K from Kinect v2 (RGB, IR, Depth) taken from center mirror view. AutoPOSE's ground truth -head orientation and position- was acquired with a sub-millimeter accurate motion capturing system. Moreover, we present a head orientation estimation baseline with a state-of-the-art method on our AutoPOSE dataset. We provide the dataset as a downloadable package from a public website.

1 INTRODUCTION

Public datasets have tremendously pushed forward computer vision research in the recent years. Objective comparisons of new algorithms on exact same data is essential for assessing contributions. In addition, since the rise of deep learning methods, large-scale datasets have become crucial to realize research and development.

There is a large interest in car interior human-centered applications, such as driver attention monitoring, driver intention prediction, and driver-car interaction. All these technologies requires as basis the head pose and gaze of the driver. The head pose describes the head position and orientation in the car, whereas the gaze is the direction of the driver's view.

Recent datasets provide either head pose or gaze or have an automotive context. However, none of them contains the combination of all of them. Thus, we propose AutoPOSE, which is the first dataset providing combined driver head pose and gaze for in-car analysis.

In more detail, our contributions are:

- We provide a large-scale, accurate, driver head pose and eye gaze dataset.
- The dataset contains images acquired from two different camera positions in our car simulator and provides different image types: dashboard (IR, $\sim$ 1.1M) and center mirror (RGB, Depth, IR, $\sim$ 315K each).
- All frames are annotated of the dataset with information about driver's activity, accessories (glasses) and face occlusion.
- We provide baseline results for head orientation estimation task where we evaluate POSEidon network on our dataset.

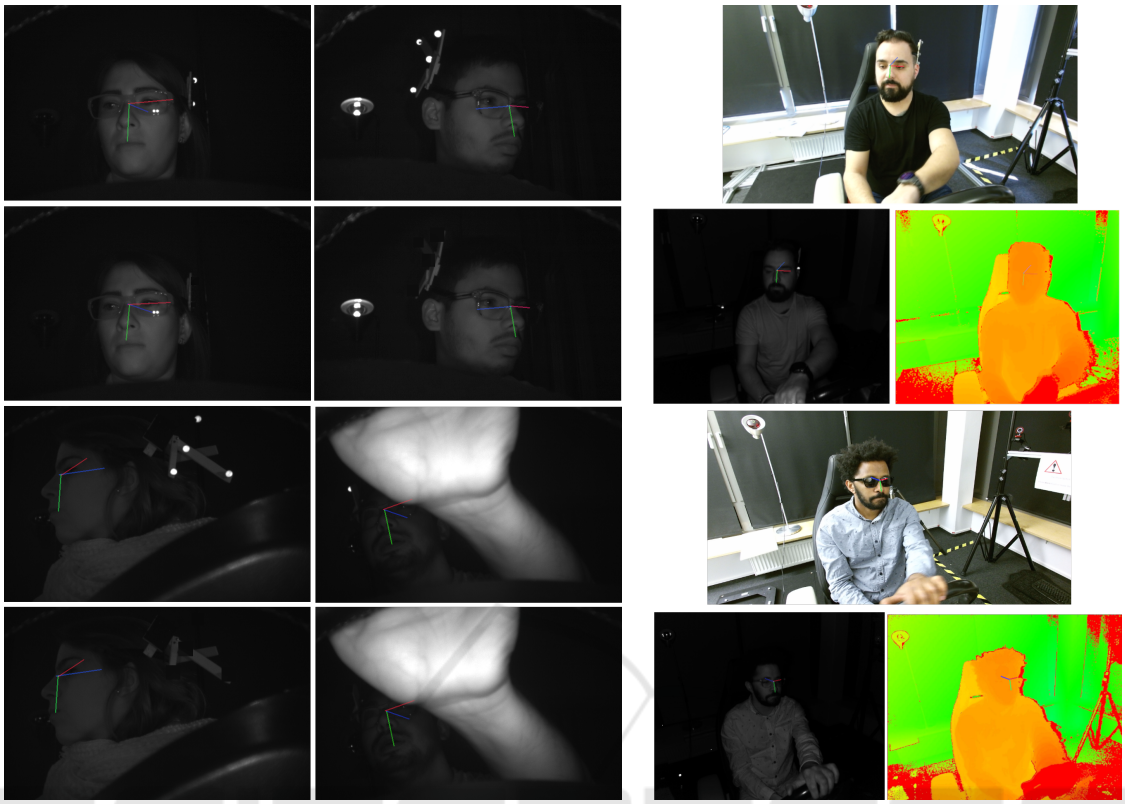
In the remainder of the paper, we discuss the related work on head pose estimation and on latest head pose datasets in section 2. We present our new dataset in section 3, and explain in detail how we acquired it in section 4. We discuss and evaluate the head orientation estimation algorithm on our dataset in section 5. In section 6, we conclude and summarize our work.

2 RELATED WORK

2.1 Related Datasets

In 2017, two new head pose datasets were introduced, the DriveAHead (Schwarz et al., 2017) and Pandora (Borghi et al., 2017). The DriveAHead proposed a novel head reference system (or head coordinate system), defining where the head center is, and how the

Selim, M., Firintepe, A., Pagani, A. and Stricker, D.
AutoPOSE: Large-scale Automotive Driver Head Pose and Gaze Dataset with Deep Head Orientation Baseline.
DOI: 10.5220/0009330105990606
In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020) - Volume 4: VISAPP*, pages 599-606
ISBN: 978-989-758-402-2; ISSN: 2184-4321
Copyright © 2022 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved



(a) IR Camera - Dashboard (b) Kinect v2 - Center rear mirror

Figure 1: (a) Row 1: RAW images with head target reflective markers visible, Row 2: post-processing - markers covered. Second column shows gaze annotation lamp(b) Kinect color, IR, and depth(color mapped) images. Note: Intensity was improved for visibility and printing purposes.

x, y, and z-axis of the head are defined related to specific facial landmarks. We also use the same head reference system in our AutoPOSE dataset. The DriveA-Head provided IR and depth frames from a Kinect camera in a real driving scenario. The authors did not provide accuracy measures of the motion capturing system while driving, although the motion of the car will affect the tracking system calibration accuracy. The dataset is suitable for deep learning frameworks.

Pandora (Borghi et al., 2017) is a large scale dataset that is also suitable for deep learning frameworks. However, the authors did not specify a head reference system (head center and rotation axis). In addition, the subjects were acting to be driving on a normal chair in front of a wall. In our AutoPOSE, we provide data captured in a real car cockpit with cameras placed at the dashboard and the center mirror location. Moreover, we use a well-defined head reference coordinate system.

In 2015, the MPIIGaze (Zhang et al., 2015) dataset was introduced containing RGB images only. The subjects were gazing at known points at a computer screen. As RGB cameras are highly affected by sunlight, they are not suitable for driving scenarios (Schwarz et al., 2017). In our AutoPOSE, we provide IR images from two perspectives (dashboard, center mirror) with 3D gaze target ground truth in a driving environment.

In summary, all existing datasets have specific drawbacks. AutoPOSE provides ground truth in a controlled environment, that ensures ground truth correctness and quality. Moreover, we provide frame annotations were the subjects performed the required task while having no glasses on, with clear glasses on and with sunglasses on. All frames were manually annotated. The dataset provides two camera views (dashboard, and center mirror) in a car cockpit with gaze target ground truth and occlusion annotations.

2.2 Head Pose Estimation

Approaches for head pose estimation are performed either on 2D information like RGB (Baltrusaitis et al., 2018; Ranjan et al., 2019), or IR images (Schwarz et al., 2017), or on 3D information like depth maps (Borghi et al., 2017; Borghi et al., 2018). The selec-

tion of the suiting input type depends also on the category of an approach. Three main categories have been defined to classify approaches: feature-based, 3D model registration and appearance-based approaches (Fanelli et al., 2011; Meyer et al., 2015; Borghi et al., 2017). Feature-based approaches need defined facial features like eye corners or mouth corners, which are then localized in frames to perform pose estimation. These approaches can work on 2D as well as 3D information. In (Barros et al., 2018), two different feature-based approaches have been combined to regress head pose, the approaches being defined facial landmarks on the face and keypoints computed by motion. The approach requires 2D images only.

3D model registration derives a head model from the data and regresses a head pose depending on the derived 3D information. This can be done based on 2D and 3D or both. (Papazov et al., 2015) uses facial point clouds and matches them with possible poses.

Appearance-based approaches take the whole information provided into consideration and try to regress a pose. They are generally learning-based methods. This can be either a raw 2D image or a depth map, as in the DriveAHead-approach (Schwarz et al., 2017). The DriveAHead-approach uses both, 2D-IR-images and depth information to regress a pose. The POSEeidon-framework (Borghi et al., 2017; Borghi et al., 2018) uses 3D-information only to derive other types of information like motion and grayscale image to regress the 3D orientation.

The baseline method we use in this paper is based on deep neural networks, which has proven to have high potential for the head pose estimation task as shown by (Borghi et al., 2017; Borghi et al., 2018; Ahn et al., 2015; Ahn et al., 2018), however, requiring large amounts of data.

3 AutoPOSE DATASET

We introduce a new headpose and eye gaze dataset. We captured our images using two cameras placed at two different positions in the car simulator in our lab. One camera, is an IR-camera placed at the dashboard of the car, and targeted at the driver. The second camera is a Kinect v2 placed at the location of the center miror of the car providing 3 image types, [IR, depth (512x424 pixels)], and RGB (1920x1080) images. The dataset consists of 21 sequences. Our 21 subjects were 10 females and 11 males. The dashboard IR camera was running at 60 fps, giving in total **1,018,885** IR images. The Kinect was running at 30 fps, giving in total **316,497** synchronized RGB, depth, and IR images.

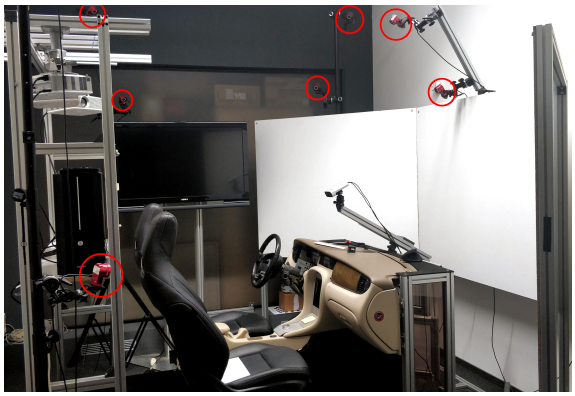


Figure 2: Driving simulator at our lab. The red circles highlights some of the motion capture system cameras.

It was not possible to capture the data using both cameras at the same time, because the strong IR light emitted by the Kinect was affecting the image captured by the camera located at the dashboard. Consequently, we decided to capture the data first with the dashboard IR camera, then capture with the Kinect. In each capturing sequence, the subject was asked to perform the tasks listed in Table 1. First, the subject was instructed about all the tasks required. The subject performed pure rotations as much as possible, followed by free natural motion, with and without face occlusions using his/her hand. Later, the gaze tasks which are described later in detail in subsection 4.5.

All tasks were first performed without any glasses on the face of the subject. Later on, all tasks were performed again with clear glasses on, then with sunglasses on. After acquiring the data with the dashboard camera, the whole experiment was repeated again using the Kinect camera while turning the dashboard IR camera off. It is noted that the subjects were faster in performing the tasks again for the Kinect sequence, thus leading to less frames for the Kinect sequence. Also, 4 Kinect sequence were discarded due to technical issues that lead to invalidating them. All tasks for all of our 21 subjects were manually annotated stating the start/end frame, along with the task performed, and glasses existence with its type.

3.1 Head Coordinate System

As introduced in subsection 2, existing datasets have different head coordinate system definition. In other words, when treating the head as a rigid body, it is required to define the x, y, and z axis of the head, and the head center. In our dataset we decided to follow the head coordinate system definition proposed in (Schwarz et al., 2017), which adds more consistent data to the community. The definition, requires

Table 1: Number of frames per annotation of the IR-Dashboard camera.

	No glasses	Clear glasses	Sunglasses	Neck scarf	Total
Pure yaw rotation	12k	12.5k	13.5k	11.7k	**50k**
Pure pitch rotation	12k	11.7k	12.6k	11k	**47.5k**
Pure roll rotation	13k	12k	12.5k	11k	**48.5k**
Free natural motion	374k	153k	158k	22k	**705k**
Free natural motion - Hand near face actions	40k	-	-	-	**40k**
Gaze point 1 - Left mirror	7.2k	6.5k	6.7k	-	**20.5k**
Gaze point 2 - Right mirror	7.6k	7k	7k	-	**21.6k**
Gaze point 3 - Dash board	7.5k	7.2k	7.5k	-	**22.2k**
Gaze point 4 - looking forward at the road	7.3k	6.6k	7.1k	-	**21.1k**
Gaze point 5 - Back mirror	6k	6.8k	7.1k	-	**19.9k**
Gaze point 6 - Media center	7.4k	6.6k	6.6k	-	**20.6k**
	495k	**230k**	**238k**	**55k**	**1M**

8 landmarks on the face, which are four eye corners, two nose corners, and two mouth corners. The head center is the 3D mean point of the four eye corners. The The x-axis is defined to be the 3D vector that starts at the head center, and passes between the left eye corners. The y-axis is computed as follows. The 3D mean point of the two nose corners and two mouth corners is projected on the plane whose norm is the x-axis. The projected point and the head center define the y-axis of the head. Finally, the z-axis is the cross product of the x and y axis.

4 DATA ACQUISITION

In order to acquire reliable and accurate ground truth for AutoPOSE, we used a sub-millimeter accurate motion capturing system, the OptiTrack, which consists of 12 Flex13 IR cameras, running at 120 fps. We calibrated the system using the Motive software of OptiTrack. At the beginning of each recording, the system was calibrated by waving a calibration wand that has three markers at known distances. The calibration sequence is used to estimate the intrinsics and extrinsics of each camera in the setup. The system tracks the reflective markers and provides the 3D position and orientation of the defined rigid bodies.

In our dataset, the subject put on a rigid tool containing 8 reflective markers at the back of the head, we refer to it as the head target. The calibration software computed a mean 3D error for the markers tracking of 0.32 mm. We also attached 8 markers to the IR camera at the dashboard, and also 8 markers to the Kinect v2 camera. In the tracking software, we set the rigid bodies to be tracked only if all markers are visible and tracked. This ensures the most accurate tracking possible of the subject head and the cameras.

By applying image processing and projecting known 3D head target marker locations on the 2D images, we erase the markers in order not to provide more realistic images for learning. Figure 1 (a) shows pictures with and without the markers.

4.1 System Synchronization and Calibration

In this section, we describe in details our system synchronization and calibration, which allows us to have the subject's head pose (orientation and translation) with respect to the camera coordinate system for each frame.

In order to synchronize the images of the cameras and the tracking information, the cameras and the tracking system were running on the same computer. The captured images were saved along with the timestamp of the computer. Also, the tracking information from OptiTrack were saved on the same machine along with the timestamp. This enabled us to synchronize the images with the 3D information. Since, the tracking system is running at 120 Hz and the cameras are running at 60 Hz (dashboard IR) and 30 Hz (Kinect), we were able to select the best matching 3D information for each frame in the dataset. The average difference between the time stamps is max 5 ms.

4.2 Camera - Handeye Calibration

We placed spherical reflective markers on the camera body, thus in each frame we get the position and orientation of the camera body in our reference coordinate system. Our aim is to find the head pose, described as orientation and translation in the camera coordinate system. Consequently, the rigid transformation between the camera body (defined by our re-

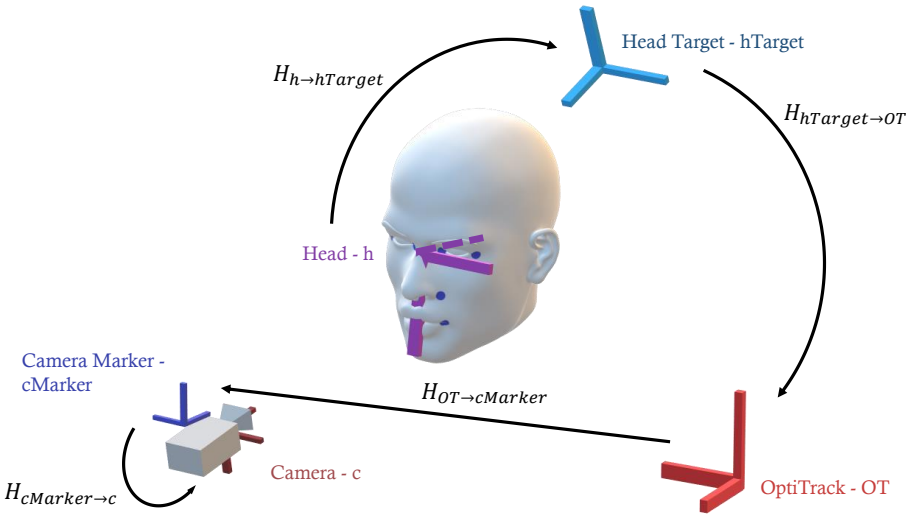


Figure 3: (a) 3D demonstration of the setup coordinate systems. (b) Sample image showing the subject wearing the head target, Kinect at the center mirror position, and gaze annotation IR lamp at the back.

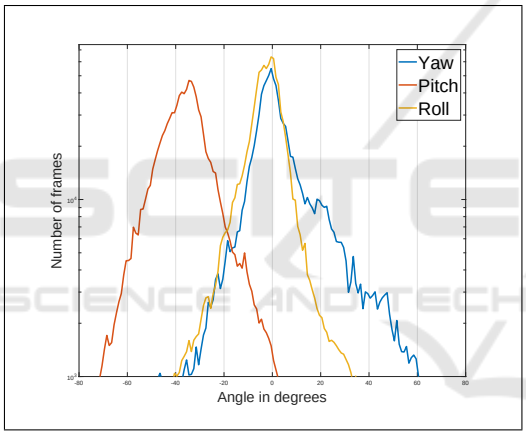


Figure 4: AutoPOSE yaw, pitch, and roll angles histogram.

flective markers as a rigid body in OptiTrack), and the camera coordinate system must be computed.

We used the handeye calibration algorithm (Tsai and Lenz, 1989) to find the required rigid transformation. We attached spherical reflective markers to both cameras (IR, and Kinect). We calibrated our two cameras with 50 images. The re-porjection error is 2.19 pixels for the IR camera. The error for the Kinect v2 RGB camera is 3 pixels, and 2.3 pixels for the Kinect's IR camera.

4.3 Head Calibration

In order to compute subject-specific head reference system, he/she puts on the head target, so that it rests at the back of the head, and does not occlude any part of the face. The experiment coordinator puts on the subject's 8 special facial markers from OptiTrack at the designated positions mentioned before. We record with OptiTrack a calibration sequence of 1 min where the subject rotates his/her head in yaw, pitch, and roll directions.

We compute for each frame the head coordinate system, and find the rigid transformation from the head coordinate system to the head target. Finally, we compute the average transformation among all frames. This defines our subject-specific head calibration. The facial markers are removed, and the subject is now ready for recording the dataset sequence.

In order to compute the calibration error, we use the calibration sequence. We consider the computed head reference system as ground truth. We apply the computed transformation on the head target, this gives us the recovered head pose. The calibration error in is as small as **1.02 mm** for translation and **1.6 degrees** for the orientation.

4.4 Head Pose

Finally, in order to find the head pose (translation and orientation) in the camera coordinate system, we track the head target and the camera rigid bodies in each frame. We apply transformations in this order

$$H_{h\to c} = H_{cMarker\to c} \cdot H_{OT\to cMarker} \cdot H_{hTarget\to OT} \cdot H_{h\to hTraget} \quad (1)$$

Figure 4 shows the histogram of the yaw, pitch and roll angles of the 1M frames from the dashboard IR camera. The rotations were limited to -90 degrees to +90 degrees. As shown, the pitch angle histogram

is shifted in the negative values of the rotation angles. This is due to the placement of the camera in the dashboard, where it is looking at the face from the bottom, check Figure 1 (a).

4.5 Eye Gaze

In our dataset, we provided annotation for gaze frames. We asked our subjects to gaze at six spherical reflective markers placed at driving-related locations, the dashboard, in front of the driver (representing looking at the road), center mirror, 2 side mirrors, and center of the car (representing media center, climate control). The car markers are tracked by OptiTrack throughout the entire sequence. We asked the subject to gaze at each marker for 5 to 10 seconds.

We think that the best person to tell if he/she is gazing at a point or not is the person him/herself. We placed a button close to the subject. When the subject gazes at a marker, he/she press the button, which turns an IR lamp at the back, visible in the frame, and does not interfere with the camera's IR light. Later, we manually annotated the start and end frame for each gaze target. The ground truth gaze targets can be used to in gaze estimation algorithms assessment in automotive or other fields. To the best of our knowledge, this is the first time to provide gaze target ground truth in automotive field using IR camera from dashboard and from center mirror views.

5 HEAD ORIENTATION ESTIMATION BASELINE

We used the POSEidon-CNN on the IR data to perform head pose estimation. Before training, we conducted preprocessing on the raw images for cleaning and obtaining cropped images of the frames. As a first dataset preparation step, we cleaned the 752x480 pixel images of the IR camera. We kept the frames with yaw rotations higher than 120 degrees for training to increase robustness, but did not consider them in the validation and test set. We additionally equalize and normalize the images.

The authors of (Borghi et al., 2017) and (Borghi et al., 2018) rely on the output of a neural network to regress 2D head position, which they further use for cropping. This outputs the head center in image coordinates (x_H, y_H). We obtained the head center from the ground truth data instead of a neural network. This prevents having additional error in the pose estimation part introduced through another position estimation method. A dynamic size algorithm provided the head bounding box with the acquired head center, the

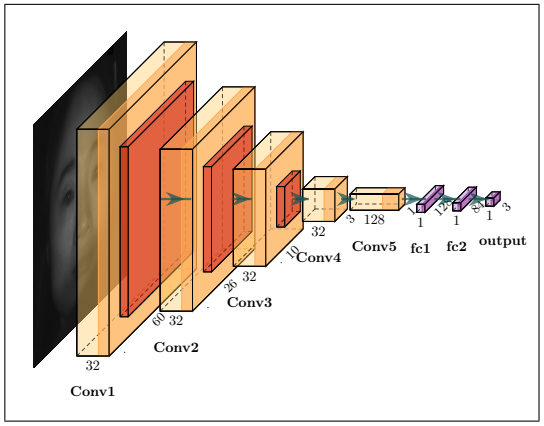


Figure 5: The POSEidon-CNN (Borghi et al., 2018) for 64 pixel images.

width w_H and the height h_H, which are used to crop the frames. We acquired them as described in (Borghi et al., 2017). With the horizontal and vertical focal lengths of the acquisition device, distance D between the head center and the acquisition device and R_x and R_y, which are the average width and height of a face. The head width R_x and height R_y in 3D were defined uniformly as 32 cm, so the head is equal in size inside the cropped images. Moreover, if more than a third of the head were not visible in the frame, we discarded the cropped image.

5.1 Network Architecture

We considered part of a recent head pose estimation framework: the POSEidon-framework (Borghi et al., 2018). The framework relied on depth data and did not perform Head Pose Estimation on IR images. The head pose estimation part in the framework is based on three different branches, which considers depth maps, grayscale images generated from depth maps and motion images. All branches were trained with the same CNN architecture. The output of the three branches is fused in the end. We obtained the CNN, which each branch in the framework used separately (Figure 5).

The model exploited Dropout as regularization (σ = 0.5) at the two fully connected layers.

We trained and tested the described model on the IR data of the dashboard IR camera, providing baseline results for the dataset.

5.2 Network Training

We trained the Deep Neural Network on the cropped images of the dataset. We selecedt training and test setup including loss function and training, validation

and test set definition accordingly. To evaluate the model, we chose metrics for benchmarking on the dataset.

For training, we defined our loss function as presented in (Borghi et al., 2017; Borghi et al., 2018) to put more focus on the yaw, which is predominant in the automotive context. Our labels range from -180 degree to 180 degree. We used a weighted L_2 loss between label and prediction, where we weighed the difference between them on the yaw with 0.45, pitch with 0.35 and roll with 0.2. Furthermore, we took 19 of the 21 sequences of the subjects for training. We use one sequence for the validation set and one for testing. The training was done in batches with a size of 128, where the batches were chosen randomly.

5.3 Evaluation Metrics

To provide a good benchmarking foundation, meaningful metrics for the head pose estimation task are required. Thus, we chose 4 metrics as a basis for further benchmarking.

The first metric is the angle estimation error, that we refer to as Mean Absolute Error (MAE).

$$MAE := \frac{1}{n}\sum_{i=1}^{n}|y-\widetilde{y}| \tag{2}$$

It provides an easily comprehensible metric. Computing it on one axis or all axis result in the total estimation error on the respective input. Another metric is the Standard Deviation (STD), providing further insight to the error distribution around the ground truth.

The third metric is the Root Mean Squared Error (RMSE) to weigh larger errors higher.

$$RMSE := \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y-\widetilde{y})^2} \tag{3}$$

It takes the squared difference of the predicted value and the ground truth value, weighing larger errors higher. Thus, high variation in predictions of an algorithm result in a higher overall error compared to the mean without squaring the values. Computing the mean over one or all axis and subsequently calculating the square root of the outcome produces the same unit as the predictions and ground truth, thus making it more understandable.

The last metric is the Balanced Mean Angular Error (BMAE) as defined in (Schwarz et al., 2017), which provides further insight as it takes different ranges into consideration. The authors base their metric on the unbalanced amount of different head orientations due to driving and its bias towards frontal orientation. The BMAE addresses this:

$$BMAE := \frac{d}{k}\sum_{i}\phi_{i,i+d}, i \in d\mathbb{N}\cap[0,k], \tag{4}$$

$\phi_{i,i+d}$ is the average angular error. In contrast to (Schwarz et al., 2017) which computes the difference based on quaternions, we compute it as $|y-\widetilde{y}|$ for all labels y and predictions $\widetilde{y}$, where the absolute distance of ground truth angle y to zero lies between i and i + d. During our evaluation, we set the section size d to 5 degrees and maximum degree k to 120.

We tested the previously presented POSEidon-model on the metrics to provide a baseline for future head pose estimation benchmarking.

5.4 Results

Our evaluations for head orientation estimation on all metrics are shown in table 2.

Table 2: Results on the 64x64 pixel cropped images of Poseidon trained and tested on our dataset.

Metric	Pitch	Roll	Yaw	Avg
MAE	2.96	3.16	3.99	3.37
STD	4.63	3.93	7.82	5.46
RMSE	4.73	4.55	7.98	5.97
BMAE	7.10	9.42	19.05	11.86

The results showed the performance of the POSEidon-CNN on our 64 pixel images. We observed that the CNN had a lower error than 3.5 degree on the MAE. The BMAE shows that the networks performed worse if more extreme poses with less examples are weighted equally as more common poses. In general, we noted that the yaw is more challenging as the network performed worse on the yaw on all metrics compared to the pitch and roll.

6 CONCLUSION

In this paper, we introduced a new large-scale driver head pose and eye gaze dataset. We discussed in detail, the head and camera calibration pipeline that enabled us to have the head pose described in the camera frame. We captured data from two positions in our car simulator for 21 subjects (10 females and 11 males). We collected 1.1M images from the dashboard IR camera and collected 315K images for each type from Kinect v2 (RGB, Depth, IR). We acquired the ground truth head pose of all frames of the dataset head pose using a sub-millimeter accurate motion capturing system. Moreover, we annotated the frames of the dataset with information about driver's activity, face accessories (clear glasses, and sunglasses) and face occlusion.

Based on our dataset, we selected a state-of-the-art method to generate a baseline result on the IR data

for head orientation estimation task.

ACKNOWLEDGEMENT

This work was partially funded by the company IEE S.A. in Luxembourg. The authors would like to thank Bruno Mirbach, Frederic Grandidier and Frederic Garcia for their support. This work was partially funded by the German BMBF project VIDETE under grant agreement number (01|W18002).

REFERENCES

Ahn, B., Choi, D.-G., Park, J., and Kweon, I. S. (2018). Real-time head pose estimation using multi-task deep neural network. *Robotics and Autonomous Systems*, 103:1 – 12.

Ahn, B., Park, J., and Kweon, I. S. (2015). Real-time head orientation from a monocular camera using deep neural network. In Cremers, D., Reid, I., Saito, H., and Yang, M.-H., editors, *Computer Vision – ACCV 2014*, pages 82–96, Cham. Springer International Publishing.

Baltrusaitis, T., Zadeh, A., Lim, Y. C., and Morency, L. (2018). Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 59–66.

Barros, J. M. D., Mirbach, B., Garcia, F., Varanasi, K., and Stricker, D. (2018). Fusion of keypoint tracking and facial landmark detection for real-time head pose estimation. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2028–2037.

Borghi, G., Fabbri, M., Vezzani, R., Cucchiara, R., et al. (2018). Face-from-depth for head pose estimation on depth images. *IEEE transactions on pattern analysis and machine intelligence*.

Borghi, G., Venturelli, M., Vezzani, R., and Cucchiara, R. (2017). Poseidon: Face-from-depth for driver pose estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Fanelli, G., Gall, J., and Van Gool, L. (2011). Real time head pose estimation with random regression forests. In *CVPR 2011*, pages 617–624. IEEE.

Meyer, G. P., Gupta, S., Frosio, I., Reddy, D., and Kautz, J. (2015). Robust model-based 3d head pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3649–3657.

Papazov, C., Marks, T. K., and Jones, M. (2015). Real-time 3d head pose and facial landmark estimation from depth images using triangular surface patch features. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ranjan, R., Patel, V. M., and Chellappa, R. (2019). Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):121–135.

Schwarz, A., Haurilet, M., Martinez, M., and Stiefelhagen, R. (2017). Driveahead-a large-scale driver head pose dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–10.

Tsai, R. Y. and Lenz, R. K. (1989). A new technique for fully autonomous and efficient 3d robotics hand/eye calibration. *IEEE Transactions on Robotics and Automation*, 5(3):345–358.

Zhang, X., Sugano, Y., Fritz, M., and Bulling, A. (2015). Appearance-based gaze estimation in the wild. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4511–4520.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS