The More, the Merrier? A Study on In-Car IR-based Head Pose Estimation

Ahmet Firintepe[1,3], Mohamed Selim[2], Alain Pagani[2] and Didier Stricker[2,3]

***Abstract*— Deep learning methods have proven useful for head pose estimation, but the effect of their depth, type and input resolution based on infrared (IR) images still need to be explored. In this paper, we present a study on in-car head pose estimation on the IR images of the AutoPOSE dataset, where we extract 64×64 and 128×128 pixel cropped head images. We propose the novel networks Head Orientation Network (HON) and ResNetHG and compare them with state-of-the-art methods like the HPN model from DriveAHead on different input resolutions. In addition, we evaluate multiple depths within our HON and ResNetHG networks and their effect on the accuracy.**

Our experiments show that higher resolution images lead to lower estimation errors. Furthermore, we show that deep learning methods with fewer layers perform better on head orientation regression based on IR images. Our HON and ResNetHG18 architectures outperform the state-of-the-art on IR images on four different metrics, where we achieve a reduction of the residual error of up to 74%.

I. Introduction

In-car driver or passenger observation aiming for driver state prediction and action recognition has become very popular in the last years within the computer vision community. Research on fundamental computer vision tasks like detection of facial features, human pose or head pose estimation pushed by safety-related use cases like driver drowsiness detection or ADAS-related assistive functions became even more relevant for the automotive industry.

In general, computer vision algorithms achieve better results if the images can be pre-processed or normalized. As driving can be done during a sunny day or in the dark at night, classical RGB cameras can capture scenes with extremely varying illumination conditions. Thus, infrared images come in handy as the images are less dependent on the global illumination. This makes them interesting for tasks like head pose estimation in cars. IR image-based head pose estimation may have different requirements for deep learning, as they only have one channel and thus contains less information, but more consistent looking scenes than RGB images. This paper aims for a study on IR images with different deep networks and resolution sizes to get more insights on whether larger networks and higher resolution sizes are relevant for this task. For this study, we evaluate state-of-the-art deep learning methods on the AutoPOSE dataset.

In more detail, our contributions are:

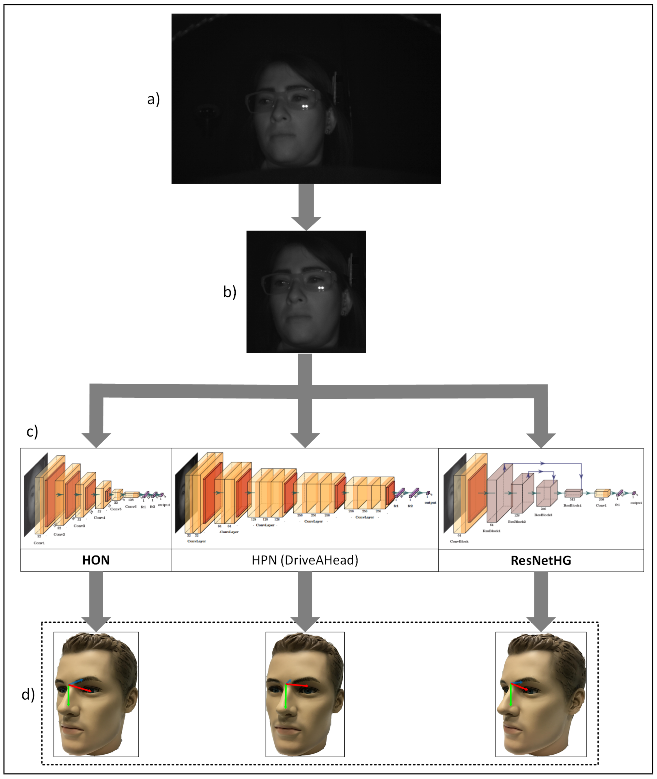


Fig. 1. Pipeline of the evaluation. a) On top is an example frame from the AUTOPose dataset. b) In the second step, the images are being cropped to head size and a resolution of 128×128 pixel images. Additionally, a 64×64 pixel version is being generated. c) In a next step, the different networks are used to regress a head orientation. The architectures used on the 128×128 pixel images are shown here. d) The outcome is further compared and evaluated on different metrics.

- We provide algorithms for head pose estimation where we outperform the state-of-the-art on IR images with our HON and ResNetHG networks.
- We analyze and compare the effect of deeper networks and more resolution on in-car head pose estimation performance based on IR images.
- We prove performance gains with fewer layers in deep neural networks on IR images.
- We show that IR images with higher resolution result in a lower pose error.

In the remainder of the paper, we discuss the related work on head pose estimation and on latest IR head pose datasets in section II. We discuss head pose estimation methods on the AutoPOSE dataset in section III and present our evaluation in section IV.

[1] are with BMW Group Research, New Technology, Innovations, Garching (Munich), Germany ahmet.firintepe@bmwgroup.com

[2] are with German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany {mohamed.selim, alain.pagani, didier.stricker}@dfki.de

[3] are with TU Kaiserslautern, Germany

II. Related work

A. Head Pose Estimation

Computer vision methods for head pose estimation either utilize 2D information like RGB [1], [2], IR images [3], or 3D information like depth information [4], [5]. The specific head pose estimation approach and its category determine a suitable input type selection. Three main categories of head pose estimation approaches exist: 3D model registration, feature-based and appearance-based approaches [6], [7], [4]. 3D model registration derives 3D information in form of a head model from the data. The derived information is used to regress a head pose. In this category, either 2D or 3D information or both can be made use of. As a 3D-only approach of this category, Papazov et al. use facial point clouds to match them with possible pose candidates [8]. Ghiass et al. fit a 3D morphable model using the depth data and including the RGB information of the image, thus utilizing 2D and 3D information [9]. The model is used to predict the pose.

Definition of facial features like eye or mouth corners are needed in feature-based approaches. These are then being localized in 2D or 3D to perform a pose estimation. Barros et al. combine two different feature-based approaches to derive head pose [10]. One approach is the usage of defined facial landmarks on the face. The second one uses keypoints computed by motion. The approach requires 2D images only. Yang et al. combine 2D and 3D information [11]. HOG features [12] are being extracted from RGB and depth images to perform a head pose estimation.

Appearance-based approaches make use of the complete information available to regress a pose and are in most cases learning-based methods. The information can be either a raw 2D image or a depth map, as in the HPN approach in DriveAHead [3]. HPN uses both, 2D and 3D information in form of IR images and depth information to regress a head pose. In the POSEeidon-framework [4], [5], only 3D information is used. Other types of information like motion and grayscale image are being derived to regress the 3D orientation.

Recent works have shown that deep neural network have a high potential for head pose estimation [4], [5], [13], [14]. Therefore, we exclusively use deep neural networks in our study, which require large amounts of data.

B. IR Head Pose Datasets

IR images are advantageous for in-car scenarios, as it dramatically lowers dependency to changing light source direction when driving. To utilize this vital advantage of IR images, we considered three IR based head pose datasets. One of them is the DriveAHead [3] dataset, introduced in 2017. Two more recent head pose datasets based on IR images are DD-Pose [15] and AutoPOSE [16].

The DriveAHead dataset consists of about 1M frames which were recorded with the resolution of 512×424 pixels. The dataset provides only cropped images, the mean size being 25x50 [15]. Thus, it is not suitable for our study which focuses on higher resolution levels.

The dataset DD-Pose consists of 330K 2048×2048 pixel binocluar stereo IR images [15]. It is recorded while driving, thus containing natural movements. At the same time, as the recording was done in an uncontrolled environment, the motion of the car while driving affects the tracking system accuracy.

The dataset AutoPOSE provides around 1.1M 752×480 pixel IR images and was recorded in a car simulator [16]. Subsequently, it contains less natural movement but more correct ground truth and higher accuracy.

As the goal of our study is to analyze the impact of different deep neural networks and image sizes on the pose estimation, we need a large dataset with a small ground truth error. Thus, we utilize the AutoPOSE dataset to train our deep neural networks as it contains more data and little ground truth error.

III. Deep Neural Network Architectures

We use different networks on the IR data to perform and evaluate head pose estimation. We conduct pre-processing on the raw images, where we clean the images first based on head visibility. Afterwards, we generate head cropped images. Figure 1 gives an overview of the evaluation pipeline.

A. Dataset preparation and cropped image generation

In a first preparation step, we sort out frames, where we keep the frames with rotations higher than 120 degrees for training to increase robustness, but eliminate them from the validation and test set. In addition, we equalize and normalize the images.

Borghi et al. use the output of a different neural network to compute the 2D head position, which they then use for cropping the image [4], [5]. Similarly, any open source library such as [17] can be used to find a head bounding box. As the orientation is more volatile and more crucial in a driving scenario, we do not want to add imprecision through position estimation in this orientation evaluation. Thus we do not perform head position estimation. Instead, we obtain the head center from the ground truth data. This prevents having additional error in the pose estimation part introduced through another position estimation method. Subsequently, we determine the head center in image coordinates (x_H, y_H). The head bounding box is deduced from the acquired head center, which is defined by the width w_H and the height h_H, used to crop the frames. 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 help deducing a dynamic size bounding box. The head width R_x and height R_y in 3D are defined uniformly as 32 cm, so the head is equal in size inside the cropped images. Additionally, we discard the cropped image, if more than a third of the head is not visible in the frame. We generate two options to evaluate on different resolution levels, one being 64×64, the other 128×128 pixels. We train and evaluate a variety of networks on the generated images for 3D head pose regression.

B. HPN model (DriveAHead)

At first, we consider one of the most recent, learning-based head pose estimation algorithms on IR data: The HPN model [3], [18], originally created as a baseline estimator for the DriveAHead dataset. We reimplement and train the model from scratch on the AutoPOSE dataset with the same initial learning rate α = 0.001 with the Adam optimizer [19]. We only change the output layer to regress euler angles to match all other models in this paper which regress euler angles. We perform no further changes on the model.

C. Head Orientation Network - HON

Secondly, we design our own, efficient network named ”Head Orientation Network” (HON) inspired by VGG [20] and the model of the POSEidon-framework [5] (Figure 2).

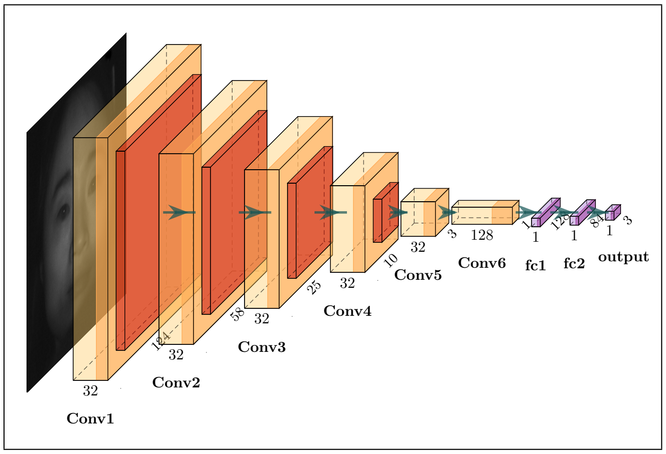


Fig. 2. Our Head Orientation Network for 128 pixel images.

We use this model only on 128×128 pixel images, as there is a baseline with a network in the AutoPOSE paper for 64×64 pixel cropped images. We train HON with an Adam optimizer with the initial learning rate α = 0.0001. We exploit Dropout as regularization (σ = 0.5) at the two fully connected layers. Finally, we develop one more, novel approach for the head pose estimation task.

D. ResNet-based model

As an alternative to the aforementioned models, we developed a model based on ResNet [21] and Hourglass [22] architectures. The model maintains information and feature within building blocks of ResNet with skip connection as described in [21]. In addition, lower level features from the head of earlier layers are being connected with later layers working on higher level head features with hourglass like skip connections [22]. Thus, coarse and fine-grained features of the head are being utilized for a head pose regression. We choose ResNet-18, which we further elaborate in subsection IV-B. Figure 3 shows an overview of the architecture, which we refer to as ”ResNetHG” in general and ”ResNetHG18” for the ResNet-18 variant in the following.

We realize the additional two skip connections by first applying a Convolution with a stride of 8 or 2, respectively. Then, the output of the source block is being added to the output of the destination block, additionally applying the ReLu-Activation.

We trained and tested the described models on cropped images from the AutoPOSE dataset on two different scaling levels.

IV. Evaluation

We train the Deep Neural Networks on the 64×64 and 128×128 pixel cropped images of the dataset. An exception is HON, which was specifically designed for the 128×128 pixel images.

For training, we define our loss function as done by Borghi et al. [4], [5] and already used in the baseline of AutoPOSE [16]. It puts more focus on the yaw, which is predominant in the automotive context. It is the weighted L_2 loss between label and prediction, where the difference between them is weighed differently: the yaw with 0.45, pitch with 0.35 and roll with 0.2. Furthermore, we also take 19 of the 21 sequences of the subjects for training and use one sequence for the validation set and one for the test set. Thus, the test and validation sets consist of around 50k images each, whereas the rest is used for training. We train the networks in randomly chosen batches with a size of 128. We train the models on 4 Nvidia Geforce GTX 1080 Ti until convergence. To evaluate the models, we use the benchmarking metrics defined in the AutoPOSE paper. These metrics are briefly described in the following section.

A. Evaluation metrics

For evaluation, we use the same 4 metrics defined for benchmarking purposes in the AutoPOSE paper [16]. y and $\widetilde{y}$ define the labels and the predictions, respectively.

The first metric is the angle estimation error or Mean Absolute Error (MAE).

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

We compute it on all axis seperately and on all axis at once for the total estimation error. The second metric is the Standard Deviation (STD), for further insight to the error distribution around the ground truth.

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

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

RMSE penalizes high variation in predictions of an algorithm, which result in a higher error. 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). It enables further insight as it takes the unbalanced amount of different head orientations due to driving and its

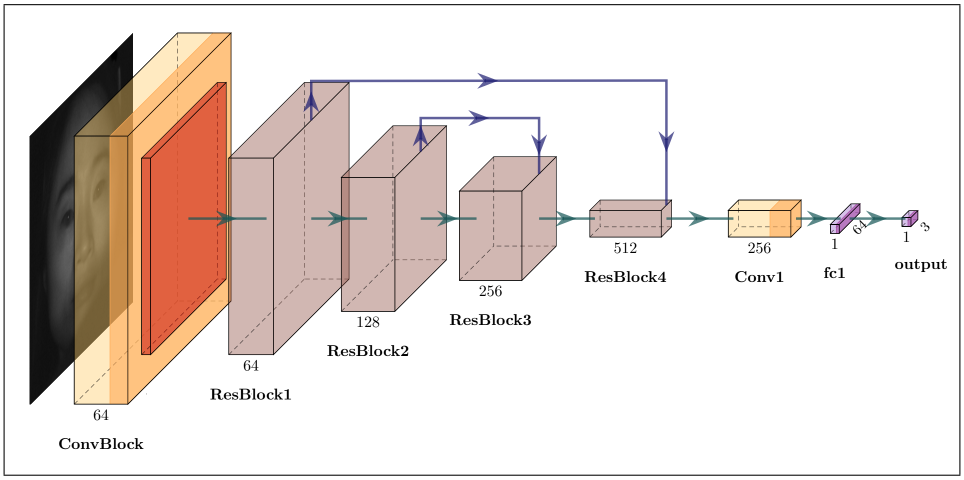


Fig. 3. The ResNetHG18 architecture. The ”ConvBlock” and the ”ResBlock”s are as described in [21]. We added connections from ”ResBlock1” to ”ResBlock4” and ”ResBlock2” to ”ResBlock3”. The model is suitable for both cropped image sizes.

bias towards frontal orientation by defining different ranges into consideration:

$$BMAE := \frac{d}{k}\sum_{i=1} \phi_{i,i+d}, i \in d\mathbb{N} \cap [0, k], \tag{3}$$

$\phi_{i,i+d}$ is defined as the average angular error. For our evaluation, we use the same section size d to 5 degrees and maximum degree k to 120 as selected by Selim et al. [16]. We tested the previously presented, trained models on the metrics to analyze the effect of different networks with different sizes and varying input resolutions.

B. Results

The training of the models gave us insight on what type of architectures are more suitable for IR images on different resolution levels as we trained on 64×64 and 128×128 pixel images.

We also evaluate different network depths on ResNetHG and the HON model to deduce what amount of layers for IR image-based head pose regression is more fitting. After comparing ResNet-50, ResNet-34 and ResNet-18 as a basis including the added skip connections, ResNet-18 showed the most promising results for the task and data at hand. In general we found out that the more layer we have in the ResNetHG and HON, the worse the estimation becomes. This is illustrated in figure 4, where we plot the error in degree on three metrics of of the aforementioned variants of the HON and the ResNetHG architecture. Both results may be caused due to the IR images being grayscale and having one channel. In comparison to RGB images which usually have 24 bit information, grayscale images only have 8 bit information, leading to overparametrization in deeper networks. Therefore, smaller networks may be sufficient for IR images. Thus, we settled on our models with fewer layers for ResNetHG and HON.

At first, we compare the results of our trained networks on 64×64 cropped images (Table I).

The results show that the baseline provided based on the POSEidon network in the dataset paper [16] performs best on the 64 pixel images, being the smallest network. ResNetHG18 achieves comparable results on the MAE, STD and RMSE, especially regarding Pitch and Roll. On the BMAE, it shows worse results, suggesting a difference in performance on different orientation ranges of ResNetHG18. With our 128×128 pixel cropped images, we achieve lower error on every metric (Table II).

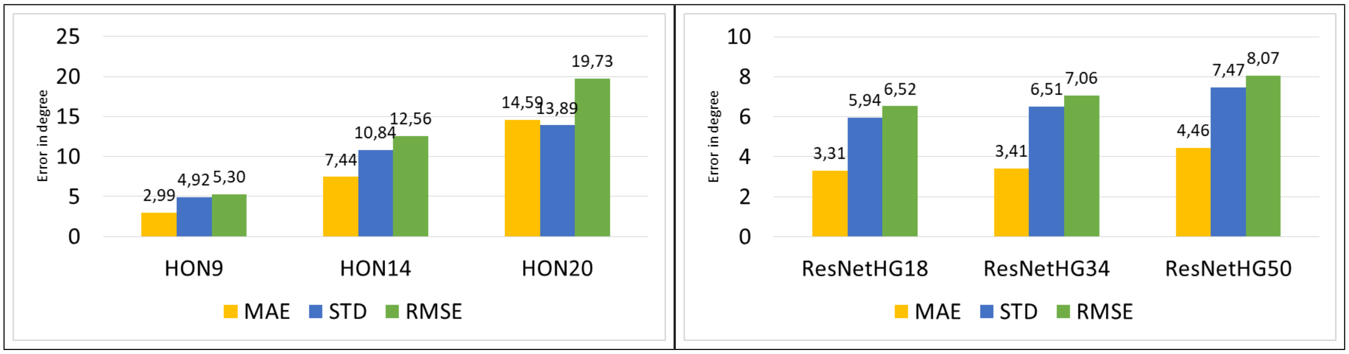
Metric	Model	Pitch	Roll	Yaw	Avg
MAE	POSEidon [16], [5]	**2.96**	**3.16**	**3.99**	**3.37**
	ResNetHG18 (ours)	4.02	3.32	5.20	4.18
	HPN (DriveAHead) [18]	8.18	6.68	13.31	9.39
STD	POSEidon [16], [5]	**4.63**	**3.93**	**7.82**	**5.46**
	ResNetHG18 (ours)	6.25	4.98	11.57	7.60
	HPN (DriveAHead) [18]	10.36	9.62	21.68	13.89
RMSE	POSEidon [16], [5]	**4.73**	**4.55**	**7.98**	**5.97**
	ResNetHG18 (ours)	6.37	5.20	11.58	8.20
	HPN (DriveAHead) [18]	11.25	9.70	21.69	15.18
BMAE	POSEidon [16], [5]	**7.10**	**9.42**	**19.05**	**11.86**
	ResNetHG18 (ours)	12.18	13.58	35.41	20.39
	HPN (DriveAHead) [18]	20.96	23.66	59.69	34.77

TABLE I
RESULTS ON THE 64×64 PIXEL CROPPED IMAGES.

The figure shows that our HON achieves the best results on all metrics and axis, excluding the roll on MAE, where ResNetHG18 produces less error. ResNetHG18 performs comparable to HON on many metrics and axis. We observe a performance loss on the yaw on STD, RMSE and BMAE. The HPN model leads to considerably higher error rates compared to the other methods amongst all axis and metrics and resolutions.

We finally compare the resolution levels and the model performances. As the POSEidon model was trained only on 64×64 pixel images and HON only on 128×128 pixel images, we directly compare them as they are both less deep networks for each resolution level (Table III).

We can observe that HPN performs similarly on both res-



(a) (b)

Fig. 4. Comparison by layer amount for the ResNetHG and HON on average error on three metrics. (a) compares HON with 9, 14 and 20 layers, (b) shows ResNetHG with ResNet-18, ResNet-34 and ResNet-50 as a basis.

Metric	Model	Pitch	Roll	Yaw	Avg
MAE	HON (ours)	**2.68**	2.73	**3.56**	**2.99**
	ResNetHG18 (ours)	3.29	**2.48**	4.15	3.30
	HPN (DriveAHead) [18]	8.32	6.87	13.81	9.67
STD	HON (ours)	**4.21**	**3.55**	**6.99**	**4.92**
	ResNetHG18 (ours)	4.86	3.74	9.23	5.94
	HPN (DriveAHead) [18]	10.36	9.62	21.68	13.89
RMSE	HON (ours)	**4.25**	**3.87**	**7.15**	**5.30**
	ResNetHG18 (ours)	4.98	3.98	9.33	6.52
	HPN (DriveAHead) [18]	11.38	9.81	21.76	15.27
BMAE	HON (ours)	**4.97**	**7.26**	**15.10**	**9.11**
	ResNetHG18 (ours)	8.00	8.18	27.62	14.60
	HPN (DriveAHead) [18]	20.93	23.96	59.4	34.81

TABLE II

RESULTS ON THE 128 × 128 PIXEL CROPPED IMAGES.

Resolution	Model	MAE	STD	RMSE	BMAE
64 × 64	POSEidon [16], [5]	3.37	5.46	5.97	11.86
128 × 128	HON (ours)	**2.99**	**4.92**	**5.30**	**9.11**
64 × 64	ResNetHG18 (ours)	4,18	7.60	8.20	20.39
128 × 128	ResNetHG18 (ours)	3.30	5.94	6.52	14,60
64 × 64	HPN (DriveAHead) [18]	9.39	13.89	15.18	34.77
128 × 128	HPN (DriveAHead) [18]	9.67	13.89	15.27	34.81

TABLE III

DIRECT COMPARISON OF THE METHODS ON DIFFERENT RESOLUTIONS.

olution levels, having the highest error compared to the other networks. For our ResNetHG18 architecture, we see a performance boost on the higher resolution level, as the error declines on all metrics. Especially the error on various degree levels as measured by BMAE and outliers as measured in RMSE declined significantly. Furthermore, in comparison to the POSEidon result provided in the dataset paper, our HON model performs better on all metrics.

On a direct comparison of the models on different resolutions, we show that HON outperforms all other methods, achieving less than 3 degree error on average on the MAE metric. Our ResNetHG18 model achieves comparable, but mostly less accurate results than its VGG-based CNN counterparts like POSEidon and HON with less layers. As we concluded from our ablation study of different depths for our HON and ResNetHG architectures, less layers result in better estimation performance. Thus, the better performance of the CNNs may be caused due to the ResNet-18 foundation of ResNetHG18, containing considerably more amount of layers and parameters.

V. CONCLUSION

In this paper, we conducted a study on IR image-based head pose estimation, in which we compare various deep neural networks of different types and depths. We designed an efficient VGG- and POSEidon-inspired network named HON, a novel network named ResNetHG based on ResNet and Hourglass and used the DriveAHead network for comparison. Based on the AutoPOSE dataset, we extracted 64 × 64 and 128 × 128 pixel cropped head images from the raw data for training, testing and evaluation. In addition, we evaluated different depths within our HON and ResNetHG networks and their effect on the accuracy.

Our evaluation on the AutoPOSE dataset showed that deep neural networks with fewer layers perform better on head orientation regression based on IR images. In addition, we have shown in our experiments on the effects of resolution that higher resolution images lead to lower estimation errors. Finally, our HON and ResNetHG18 architectures outperform the state of the art on IR images with an error with a reduction of the residual error of up to 64% to 74%, depending on the metric.

Future work should be aimed at analyzing additional input resolutions and model depths and benchmarking the models on the DD-Pose dataset for comparison on real-world data.

REFERENCES

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

[2] R. Ranjan, V. M. Patel, and R. Chellappa, "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*, vol. 41, no. 1, pp. 121–135, Jan 2019.

[3] A. Schwarz, M. Haurilet, M. Martinez, and R. Stiefelhagen, "DriveAHead-a large-scale driver head pose dataset," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, July 2017, pp. 1–10.

[4] G. Borghi, M. Venturelli, R. Vezzani, and R. Cucchiara, "POSEidon: Face-From-Depth for Driver Pose Estimation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 5494–5503.

[5] Borghi, Guido and Fabbri, Matteo and Vezzani, Roberto and Calderara, Simone and Cucchiara, Rita, "Face-from-Depth for Head Pose Estimation on Depth Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, pp. 596–609, 2017.

[6] G. Fanelli, J. Gall, and L. Van Gool, "Real time head pose estimation with random regression forests," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2011, pp. 617–624.

[7] G. P. Meyer, S. Gupta, I. Frosio, D. Reddy, and J. Kautz, "Robust model-based 3d head pose estimation," in *Proceedings of the IEEE International Conference on Computer Vision (CVPR)*, December 2015, pp. 3649–3657.

[8] C. Papazov, T. K. Marks, and M. Jones, "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)*, June 2015, pp. 4722–4730.

[9] R. S. Ghiass, O. Arandjelovi, and D. Laurendeau, "Highly Accurate and Fully Automatic Head Pose Estimation from a Low Quality Consumer-Level RGB-D Sensor," in *Proceedings of the 2nd Workshop on Computational Models of Social Interactions: Human-Computer-Media Communication*, New York, NY, USA, 2015, p. 2534.

[10] J. M. D. Barros, B. Mirbach, F. Garcia, K. Varanasi, and D. Stricker, "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)*, March 2018, pp. 2028–2037.

[11] J. Yang, W. Liang, and Y. Jia, "Face pose estimation with combined 2D and 3D HOG features," in *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, Nov 2012, pp. 2492–2495.

[12] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *International Conference on Computer Vision & Pattern Recognition*, vol. 1. San Diego, United States: IEEE Computer Society, Jun 2005, pp. 886–893.

[13] B. Ahn, J. Park, and I. S. Kweon, "Real-Time Head Orientation from a Monocular Camera Using Deep Neural Network," in *Computer Vision – ACCV 2014*. Springer International Publishing, Nov 2014, pp. 82–96.

[14] B. Ahn, D.-G. Choi, J. Park, and I. S. Kweon, "Real-time head pose estimation using multi-task deep neural network," *Robotics and Autonomous Systems*, vol. 103, pp. 1 – 12, 2018.

[15] M. Roth and D. M. Gavrila, "DD-Pose - A large-scale Driver Head Pose Benchmark," in *2019 IEEE Intelligent Vehicles Symposium (IV)*, June 2019, pp. 927–934.

[16] M. Selim, A. Firintepe, A. Pagani, and D. Stricker, "AutoPOSE: Large-scale Automotive Driver Head Pose and Gaze Dataset with Deep Head Orientation Baseline," in *International Conference on Computer Vision Theory and Applications (VISAPP)*, Feb 2020, pp. 599–606. [Online]. Available: http://autopose.dfki.de

[17] "ageitgey/face_recognition." [Online]. Available: https://github.com/ageitgey/face_recognition

[18] A. Schwarz, "Tiefen-basierte Bestimmung der Kopfposition und -orientierung im Fahrzeuginnenraum," Ph.D. dissertation, Karlsruher Institut für Technologie (KIT), 2018.

[19] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *3rd International Conference on Learning Representations, ICLR 2015*, May 2015.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015*, Y. Bengio and Y. LeCun, Eds., May 2015.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2016, pp. 770–778.

[22] A. Newell, K. Yang, and J. Deng, "Stacked Hourglass Networks for Human Pose Estimation," in *Computer Vision – ECCV 2016*, Jan 2016, pp. 483–4990.