Deep Orientation-Guided Gender Recognition from Face Images

Mohamed Selim, Stephan Krauß, Tewodros Amberbir Habtegebrial, Alain Pagani, Didier Stricker

Augmented Vision Department
German Research Center for Artificial Intelligence, DFKI GmbH
Kaiserslautern, Germany
{mohamed.selim, stephan.krauss, tewodros_amberbir.habtegebrial, alain.pagani, didier.stricker}@dfki.de

***Abstract*—In the recent decade, gender recognition and face analysis has been one of the most researched issues in computer vision. Although several solutions have been provided to the problem of gender recognition from face images, nonetheless, it is regarded as a difficult issue. Deep learning has been proven to solve challenging problems. On the other hand, several existing works have proven their ability to accurately predict the head orientation angles. The remaining error in gender prediction models requires novel solutions to try to improve it further. In this work, we present a novel deep learning-based method to predict gender using both the face image and the head orientation angles. We show that the use of head orientation information consistently boosts the accuracy of gender prediction models. We achieve this by increasing the representational power of deep neural networks by introducing a head orientation adapter. It takes the head angles as input and outputs a vector that is used to recalibrate the deep learning neural networks. The proposed method was tested on a large-scale dataset called AutoPOSE, which has sub-millimeter-accurate head orientation angles. We show that using the head orientation adapter consistently boosts the gender prediction models' accuracy, and reduces the error by 20%.**

***Index Terms*—deep learning, gender prediction, face, head orientation**

I. Introduction

Gender recognition is one of the most investigated problems in the last decade [1]. Several contributions have been presented in constrained and unconstrained environments, nevertheless gender recognition is still a challenging problem. On the other hand, head orientation estimation can now achieve high accuracy, as we shown in [2], [3], [4]. Predicting the gender using the face image was usually modeled as a classification problem [5], [6], [7], [8]. In [9], the authors presented a deep learning-based method to predict the gender from the face image. In other words, the predicted gender is a function of the face image only. The presented models learns the gender of the subjects under several changing conditions, like skin color, age, and head orientation. In this paper, we investigate the effect of the head orientation on the gender prediction models. Our aim is to study the effect of appearance changes (due to head orientation variation), on the gender prediction's accuracy, and use the head orientation to improve the accuracy of gender prediction.

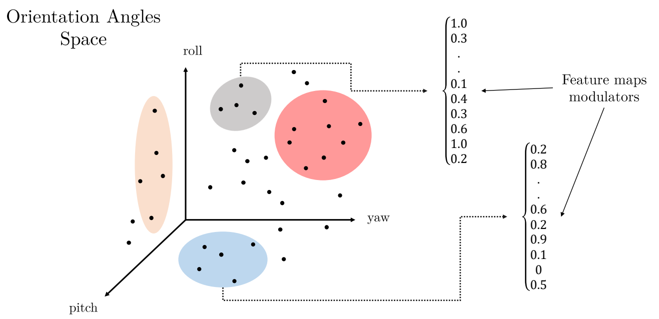

Fig. 1. **Orientation Angles Space.** *The head orientation angles space is divided into partitions where each partition yields feature maps modulators. The feature maps modulators controls the flow of features through the deep neural network model.*

In the recent years, several works aimed at improving the representational power of deep neural networks [10], [11], [12], [13]. Chen et. al. [14] presented a spatial and channel-wise attention network for the purpose of image captioning. The authors showed that employing channel-wise control over the networks feature maps outperformed the visual attention-based image captioning. Hu et. al. [15] won the first place in the ILSVRC 2017 competition with their proposed novel *Squeeze* and *Excitation* network, SENet. The authors introduced an architectural unit designed to improve the representational power of deep learning networks. This was achieved by performing channel-wise feature maps recalibration. Their unit takes the output of a convolutional block as input. The n-channels of the input feature maps are squeezed into n-single numerical values by applying average global pooling. The vector of numerical values is passed through the unit. The unit's output is a n-vector, which is the last fully connected layer. At the end of the module, a sigmoid function is applied, yielding weights between 0 and 1. The excitation part is when the weights are applied to the input feature maps, and then passed to the next layers of the network. In 2019, Wang et. al. [16]

978-1-6654-6694-3/22/$31.00 ©2022 IEEE

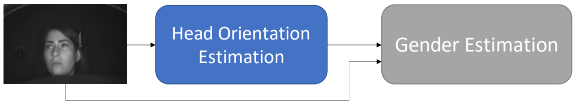

Fig. 2. **Overview.** *The figure depicts the abstract idea of the proposed method. First, The head image is used to predict the head orientation. Afterwards, the face image, along with the predicted head orientation angles are fed into the gender estimation model.*

introduced an effective and efficient object detection system, that can work on different image domains, for example human faces, CT, and satellite images. The authors were able to achieve that by introducing a set of domain adapters to the same deep convolutional network. The aim of the domain adapters is to predict the specific image domain, and based on it, dynamically recalibrate the feature maps that are most effective in such image domain.

This paper studies the problem of gender recognition and its relation to head orientation variations. We show that the accuracy of gender prediction can be boosted given the image and the head orientation angles as input. An overview is shown in figure 2. The proposed method predicts the gender of the subject as a function of the face image and the corresponding head orientation angles. The image is passed to through the deep neural model to generate image features. The head orientation angles are employed to *adapt* the network feature maps, thus boosting the accuracy in gender prediction.

The head orientation can be modeled using the three rotation angles (yaw, pitch, and roll). Given a three dimensional space with axis representing the head angles. A point in the given space represents one possible head orientation. In [9], all images in the datasets were fed into the deep neural networks for training and evaluation. It was learnt by the network to properly predict the gender in any of the given face images. In this paper, we show that the angles space can be divided and separated in a way that supports and adapts the deep neural network for improving the gender prediction accuracy. Figure 1 depicts the head orientation space, and an abstract representation of the possible subdivisions that would yield different modulators to recalibrate the feature maps inside the deep neural network models. Detailed explanation of the proposed method is presented in the following sections.

II. Orientation-Guided Gender Prediction Models

This section introduces the orientation adapter unit, and shows how it can be employed in the gender prediction deep neural network models. The model proposed in [9], the GenderCNN is modified to integrate the orientation adapter. Besides, deeper model, the ResNet-18 is also employed for gender prediction, and the integration of the head orientation in the ResNet-18 is presented.

A. *Orientation Adapter*

The proposed head orientation unit is shown in figure 3. The unit is a Multi Layer Perceptron, MLP. The aim of the orientation adapter is to encode features as a function of the orientation angles. The adapter takes the three rotation angles (yaw, pitch, and roll) as input. The angles are connected to three fully connected layers. The sigmoid function is applied on the last layer to encode the features as numerical values between 0 and 1. The unit's output is then employed in the gender prediction model. Two methods to use the output of the orientation adapter are presented. The first method is a concatenation along with the fully connected layers of the deep gender models. Thus, the network would use the image features encoded in the fully connected layer that is connected to the convolution layer and the orientation features from the adapter, to predict the gender. Another method is to use the output of the orientation adapter as a weighting scale for some feature maps which are generated by the convolution layers. The unit's output layer size must match the number of features maps of the target convolution layer, as they will be multiplied together.

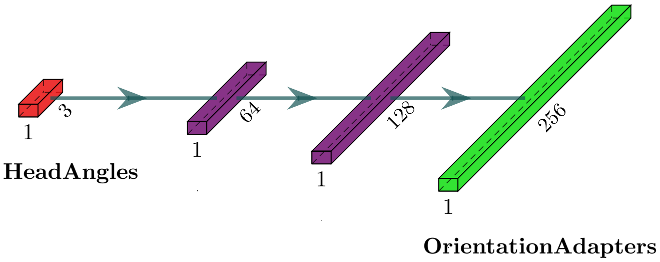

Fig. 3. **Orientation Adapter.** *Orientation adapter network architecture. The adapter takes the 3 head orientation angles as input. They are connected to 2 fully connected layers. The last layer is the output of the network. The output is to be used in other gender prediction models.*

B. *GenderCNN with Orientation Adapter*

In the work presented in [9], the convolutional neural network model GenderCNN was introduced. The model was evaluated on several public still images and videos datasets for gender prediction. The model was efficient and achieved high accuracy. In this work, the model is modified to be employ the proposed orientation adapter unit. The output of the orientation adapter is modulated with the resulting feature maps of the convolutional layer conv2. The orientation adapter takes the head angles as input, and generates weighting scales that are used to control the feature maps in the convolutional part of the model. In this work, the GenderCNN was also modified to take 1-channel images as input, since the AutoPOSE images used are infrared images. The final predicted gender is dependent not only on the input image, but also on the orientation angles. Detailed evaluation of the GenderCNN and orientation-guided GenderCNN are presented in the next section.

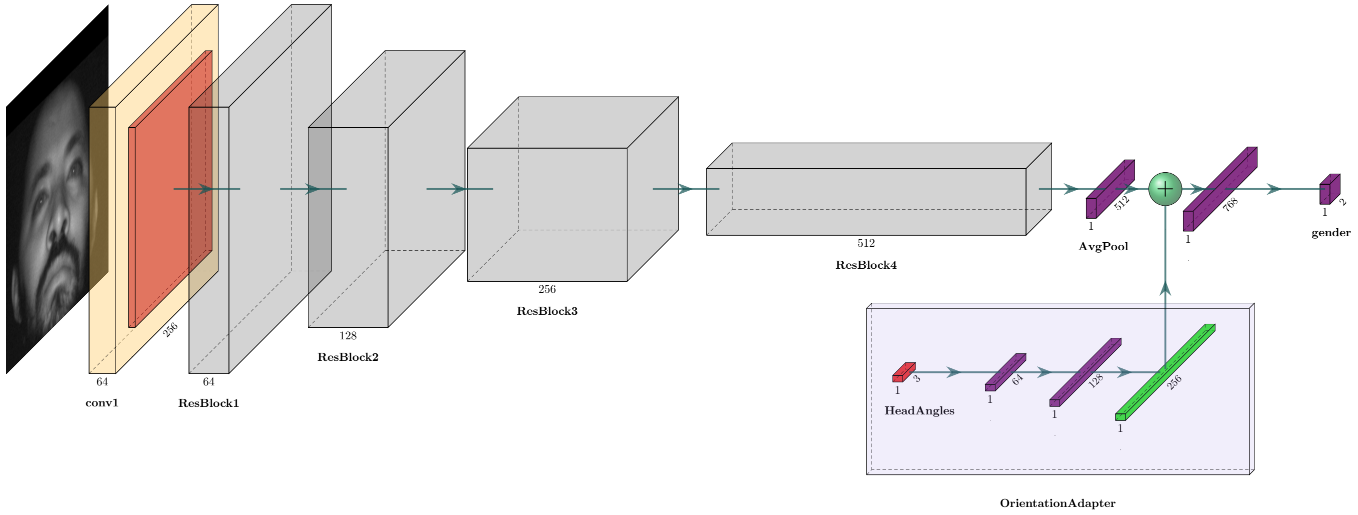

Fig. 4. **Orientation-Guided ResNet-18 - Concatenation.** *An overview of the orientation-guided ResNet-18. The face image is passed through convolution layer and the residual blocks. The head orientation angles are fed into the orientation adapter. The orientation features are concatenated with the image features after the global average pooling. The last layer contains the prediction of the gender.*

C. ResNet-18 with Orientation Adapter

He et. al. [17] introduced the idea of residual learning in 2015, and the authors won the first place in the ImageNet image classification challenge [18]. The authors showed the training error increases if the networks become deeper. The authors introduced the residual learning that elevates the issue and the networks can become deeper while keeping the training error high. More details about residual learning can be found in [17]. The ResNet-18 variant was chosen as backbone for the gender prediction problem. First, the input face image is passed through a convolution layer with 64 filters. Afterwards, the results are passed through 4 residual convolution blocks with different number of filters. After the last residual block, the feature maps are passed through a global averaging pooling layer, where each feature map is represented by one numeric value. The average pooling layer makes the ResNet-18 model independent of the input image dimensions.

The proposed orientation adapter is employed with ResNet-18 model by one of two methods, concatenation or modulation as briefly mentioned before. Figure 4 depicts the concatenation variant and figure 5 depicts the modulation one.

III. Evaluation

This section presents the dataset used, along with the training setup and evaluation results. To have a baseline for the models performance on the gender prediction problem, the GenderCNN models and the ResNet-18 are tested without orientation information. Afterwards, the orientation guided-models are evaluated and compared to the baseline results.

A. Dataset

In order to evaluate the effectiveness of the head orientation on gender prediction accuracy, a dataset with accurate head orientations is required. In this work, we used the AutoPOSE dataset [3]. It groundtruth head orientation labels were acquired using a sub-millimeter motion capturing system. An infrared camera have been used at the dashboard location in a car model. The optical motion capturing system was calibrated using off-shelf commercial software from OptiTrack [19]. The intrinsic of the acquisition cameras were calibrated, and the camera frames were calibration to the motion capturing system using the hand-eye calibration method [20], allowing describing the pose of the head as a rigid body in the coordinate frame of the acquisition cameras. The orientation of the head in 3D space is described using a rotation matrix, which is created from 3 orthogonal axis defining the head coordinate frame, similar to the definition in the DriveAhead dataset [2]. The dataset contains videos of 21 subjects (8 females and 13 males). The balance in gender is important for gender prediction. The amount of images collected was 1.1M images from the dashboard IR camera. Moreover, all frames of the dataset were annotated with information about driver's activity, face accessories (clear glasses, and sunglasses) and face occlusion. More details about the system calibration and the acquired data can be found in the original dataset paper [3]. The dataset is a good candidate for the evaluation of the proposed method as it the number of males and females are similar and the head orientation labels are accurate.

B. Models Training

The participants in the dataset were 8 females and 13 males. The dataset is not perfectly balanced between males

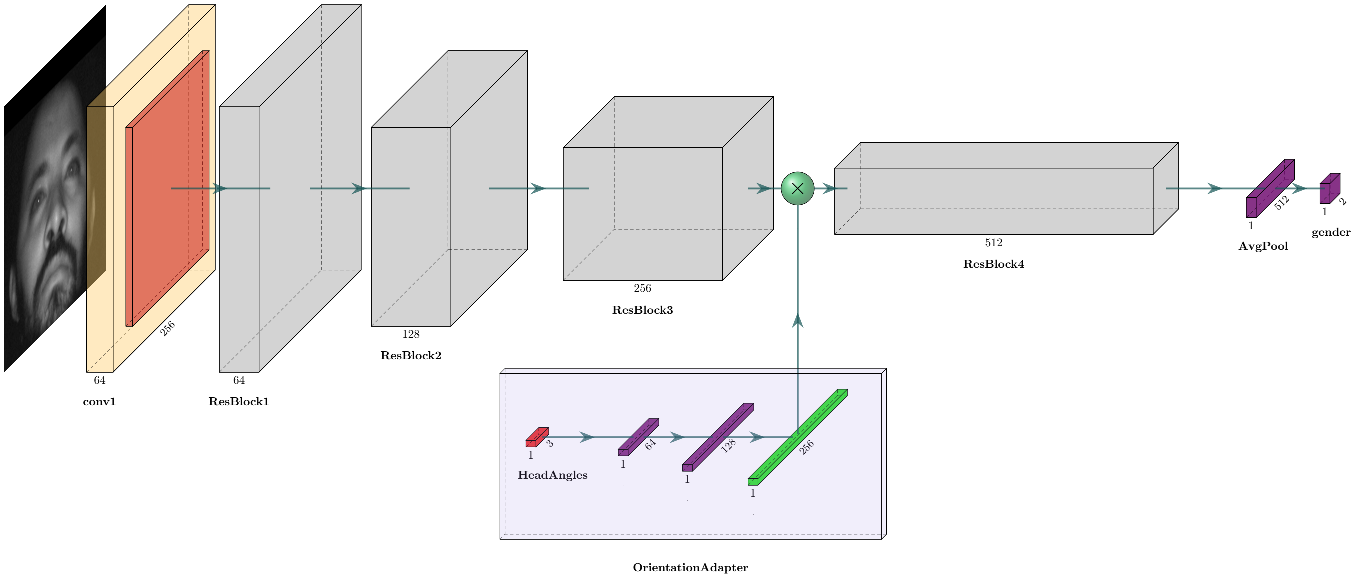

Fig. 5. **Orientation-Guided ResNet-18 - Modulation.** *An overview of the orientation-guided ResNet-18. The face image is passed through convolution layer and the residual blocks till ResBlock3. The head orientation angles are fed into the orientation adapter. The 256 orientation adapters are modulated with the 256 feature maps resulting from ResBlock3, and are fed into ResBlock4. The last layer contains the prediction of the gender.*

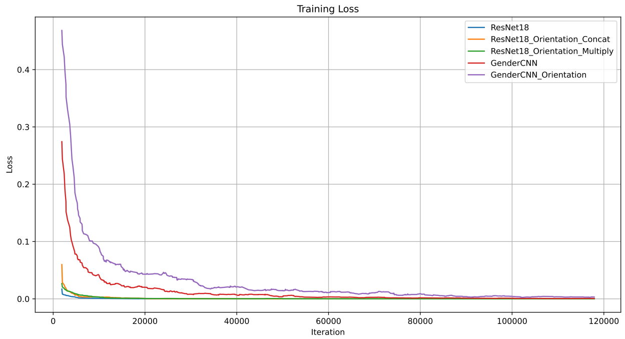

Fig. 6. **Training loss for GenderCNN and ResNet-18 variants.** *Training loss is depicted. ResNet-18 variants can learn faster the GenderCNN variants. The training loss drops much quicker that GenderCNN.*

and females. Using all subject might affect the learning of the variation between the genders. Consequently 4 male subjects were excluded from the training set. The training set consisted of 6 female subjects and 7 male subjects. The evaluation set consisted of 2 male and 2 female subjects. All face images were cropped using the groundtruth face location. The CLAHE method [21] was applied on the cropped face images. The training optimizer used was the Stochastic Gradient Descent, with a fixed learning rate of value 0.001 and momentum of value 0.9. The problem of gender classification is handled as a classification problem. Consequently, the cross entropy loss was employed in the training setup, and it is defined as follows

$$L_{CE}\left(p,y\right)=-\sum_{i=1}^{2}y_i\log\left(p_i\right)$$

where p is the prediction vector and y is the groundtruth.

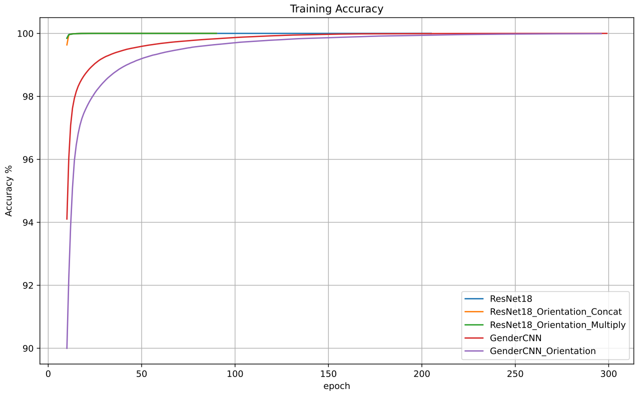

Fig. 7. **Training accuracy for GenderCNN and ResNet-18 variants.** *Training accuracy is depicted. ResNet-18 variants converge faster than the GenderCNN variants.*

Figure 6 shows the training loss. It can be noted that the training loss of the ResNet-18 variants drops considerably faster than the GenderCNN models. GenderCNN without orientation loss drops faster than the GenderCNN with orientation information. This could be due to the more parameters that are being learnt by the orientation adapter unit. Same effect is seen in the ResNet-18 variants. Figure 7 shows the training

Model	Accuracy
GenderCNN	85.5%
GenderCNN with orientation	**90.7%**
ResNet18	98.0%
ResNet18 with orientation (concatenation)	98.2%
ResNet18 with orientation (multiplication)	**98.4%**

TABLE I
Evaluation Results - Orientation-guided gender prediction models. *The table shows the best accuracy achieved by the models. The results show that using the orientation information consistently improved the overall accuracy.*

set accuracy. In general, the ResNet-18 variants can reach accuracy close to 100% much faster than the GenderCNN variants. The GenderCNN variants require much more training epochs to reach convergence.

C. Models Evaluation

This subsection presents the evaluation results of the trained models on the evaluation set. The evaluation set consists of 4 subjects, two females and two males. A summary of the best accuracy results achieved by each model is shown in table I.

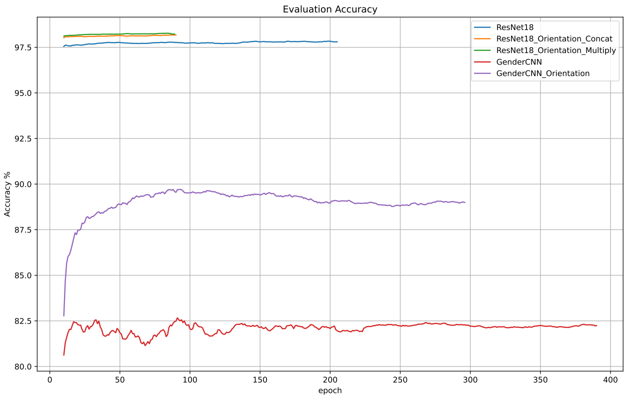

Fig. 8. **Evaluation accuracy - Orientation-guided GenderCNN and ResNet-18 models.** *The graph shows the models accuracy on the evaluation set. GenderCNN: orientation-guided version boosts the accuracy by a big margin. ResNet-18: both orientation-guided variants are better than baseline ResNet-18. Orientation guidance boosts the accuracy on both models.*

1) GenderCNN with Orientation Adapter: Starting with the GenderCNN model, it is important to first check the performance of the model without the orientation adapter. As shown in figure 8, the baseline result for the GenderCNN is on average 82% accurate. One can notice a big difference between the best possible accuracy by the GenderCNN on the AutoPOSE dataset and the other public datasets used in [9]. This can be due to the difference in the data domain. The public datasets consisted of color images with mostly frontal images. On the other hand the AutoPOSE images are IR images, and have a wide variation of head orientations. However, the orientation-guided GenderCNN performs better than the baseline variant. The best accuracy achieved by the orientation-guided GenderCNN is 90.7%. Since the GenderCNN cannot reach higher accuracy on IR images, the ResNet-18 model was employed for further evaluations.

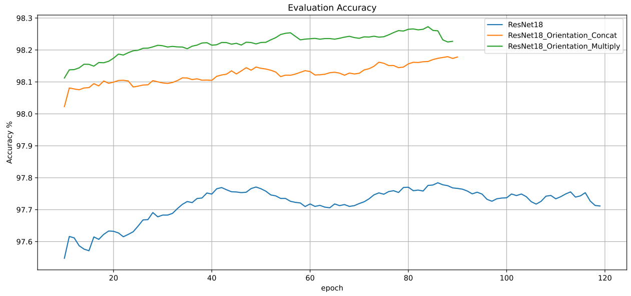

Fig. 9. **Evaluation accuracy - Detailed ResNet-18 variants accuracy.** *The graph shows smoothed accuracy results of the ResNet-18 models. The baseline version with no orientation information achieves at most 98%. Both orientation-guided variants perform consistently better than the baseline version (best result is 98.38%). The error is reduced by 20%.*

2) ResNet-18 with Orientation Adapter: Since the GenderCNN could not achieve high accuracy results, ResNet-18 model was used as backbone for a better gender prediction network. In general as shown in figure 8, all ResNet-18 variants performed considerably better than the GenderCNN baseline and the orientation-guided GenderCNN. The baseline result for the ResNet-18 model could reach at most 98%. Our hypothesis is that in case that part of the 2% error in the accuracy could be related to the orientation variation, then the orientation adapter shall improve the evaluation accuracy. Figure 9 shows part of the last part of the y-axis, where the evaluation accuracy of the baseline ResNet-18, and the two variants of orientation-guidance, concatenation and modulation are depicted. Both orientation-guided variants outperform the baseline ResNet-18 model. Over the whole training and evaluation cycles, predicting the gender using the face image and the angle is consistently better than just using the face image. The concatenation version is not as good as the modulation version. Modulating the orientation adapters vector with the feature maps of Resblock3 achieves the best result, with accuracy if 98.38%.

IV. Conclusion

In this paper, a novel deep learning-based orientation-guided gender prediction method from face image was introduced. A new orientation adapter unit was introduced to be employed along with deep neural networks to boost the accuracy of gender prediction. Two methods were tested for using the orientation features, concatenation with fully connected layers and modulation with feature maps inside the network flow. We show that orientation guidance consistently boosts the gender prediction accuracy on both GenderCNN and ResNet-18 models. The proposed method was evaluated on a large scale and accurate dataset, the AutoPOSE. We also concluded that ResNet-18 variants can predict the gender with higher accuracy compared to GenderCNN. By employing the orientation information in the ResNet-18 model using the orientation adapter, the error in gender prediction was reduced by 20%.

Acknowledgment

This work has been partially funded by the EU project HyperCOG (Grant agreement number 869886).

References

[1] A. Greco, A. Saggese, M. Vento, and V. Vigilante, "A convolutional neural network for gender recognition optimizing the accuracy/speed tradeoff," *IEEE Access*, vol. 8, pp. 130 771–130 781, 2020.

[2] 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*, 2017, pp. 1–10.

[3] 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)*, 2020. [Online]. Available: http://autopose.dfki.de

[4] A. Firintepe, M. Selim, A. Pagani, and D. Stricker, "The more, the merrier? a study on in-car ir-based head pose estimation," in *IEEE Intelligent Vehicles Symposium (IV)*, 2020.

[5] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[6] J. Mansanet, A. Albiol, and R. Paredes, "Local deep neural networks for gender recognition," *Pattern Recognition Letters*, vol. 70, pp. 80–86, 2016.

[7] 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.

[8] G. Levi and T. Hassner, "Age and gender classification using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Workshops*, June 2015. [Online]. Available: http://www.openu.ac.il/home/hassner/projects/cnn_agegender

[9] M. Selim, S. Sundararajan, A. Pagani, and D. Stricker, "Image quality-aware deep networks ensemble for efficient gender recognition in the wild," in *VISAPP 2018 - 13th International Conference on Computer Vision Theory and Applications; VISIGRAPP, Proceedings*, 01 2018, pp. 351–358.

[10] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*. PMLR, 2015, pp. 2048–2057.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[12] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *International Conference on Machine Learning*. PMLR, 2017, pp. 1857–1865.

[13] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 3156–3164.

[14] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, "Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5659–5667.

[15] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[16] X. Wang, Z. Cai, D. Gao, and N. Vasconcelos, "Towards universal object detection by domain attention," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7289–7298.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[18] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2009, pp. 248–255.

[19] "Optitrack." https://optitrack.com/, 2020, online; accessed December-2020.

[20] R. Y. Tsai and R. K. Lenz, "A new technique for fully autonomous and efficient 3d robotics hand/eye calibration," *IEEE Transactions on Robotics and Automation*, vol. 5, no. 3, pp. 345–358, June 1989.

[21] Wikipedia, the free encyclopedia, "daptive Histogram equalization," 2016. [Online]. Available: https://en.wikipedia.org/wiki/Adaptive_histogram_equalization