

Appendices of WHENet

Yijun Zhou
yijun.zhou1@huawei.com
James Gregson
james.gregson@huawei.com

Huawei Technologies Canada

1 Hyperparameter Studies

Tables 1, 2, 3, 4 and 5 show ablation studies of mean average error for the β and α meta-parameters of WHENet-V and WHENet, tested on the AFLW2000, BIWI datasets and our combined dataset. From this, we selected the best overall performance as $\beta = 2$ and $\alpha = 0.5$ for WHENet-V, $\beta = 1$ and $\alpha = 1$ for WHENe, although performance is not overly sensitive to these choices.

Table 1: WHENet-V MAE vs. α and β on AFLW2000

	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 2$
$\beta = 0.5$	4.984	4.966	5.113
$\beta = 1$	4.946	5.146	4.904
$\beta = 2$	4.834	4.953	5.189

Table 2: WHENet-V MAE vs. α and β on BIWI

	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 2$
$\beta = 0.5$	3.531	3.501	3.554
$\beta = 1$	3.551	3.676	3.626
$\beta = 2$	3.475	3.466	3.513

Table 3: WHENet MAE vs. α and β on AFLW2000

	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 2$
$\beta = 0.5$	5.822	5.624	5.620
$\beta = 1$	5.484	5.424	5.529
$\beta = 2$	5.658	5.414	5.509

Table 4: WHENet MAE vs. α and β on BIWI

	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 2$
$\beta = 0.5$	3.823	3.855	3.880
$\beta = 1$	3.843	3.814	3.786
$\beta = 2$	3.710	4.064	3.935

Table 5: WHENet MAE vs. α and β on our combined dataset

	$\alpha = 0.5$	$\alpha = 1$	$\alpha = 2$
$\beta = 0.5$	8.009	8.287	7.394
$\beta = 1$	7.331	7.655	7.879
$\beta = 2$	7.878	7.457	7.694

Table 6: Comparison results on BIWI dataset with different modality methods. WHENet and WHENet-V are trained on 300W-LP and our combined dataset. The rest of the methods are trained on BIWI where they split the BIWI dataset into testing and training.

	Yaw	Pitch	Roll	MAE
RGB-based				
DeepHeadPose [8]	5.67	5.18	-	-
SSR-Net-MD [10]	4.24	4.35	4.19	4.26
VGG16 [9]	3.91	4.03	3.03	3.66
FSA-Caps-Fusion [10]	2.89	4.29	3.60	3.60
WHENet-V	3.60	4.10	2.73	3.47
WHENet	3.99	4.39	3.06	3.81
RGB+Depth				
DeepHeadPose [8]	5.32	4.76	-	-
Martin [8]	3.6	2.5	2.6	2.9
POSEidon+ [9]	1.7	1.6	1.7	1.6
RGB+Time				
VGG16+RNN [9]	3.14	3.48	2.6	3.07

2 Robustness

A key objective of WHENet is to be robust to adverse imaging conditions as well as occlusions and accessories such as eyewear and hats. Much of the robustness of WHENet can be derived from using a similar network architecture as Hopenet [8] which also performs well due to the CNN architecture.

Figure 1 shows a selection of occluded face images where the subject tried to maintain consistent head pose while blocking areas of their face. The angular predictions are quite stable with angles varying by only 7° in spite of significant occlusions of features (some underlying variation of pose is expected due to subject motion). This suggests the method is learning high-level features rather than specific localized details.

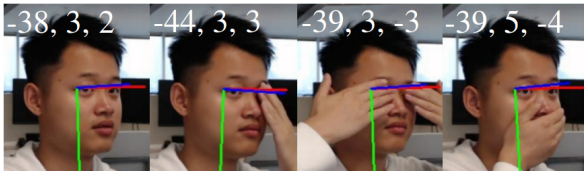


Figure 1: Head pose estimation with occlusion. Subjects asked to remain still while covering different regions of their face. Predicted deviations are within 7° of the unoccluded view (left). Some amount of deviation is expected due to slight subject motions.

We also evaluated the effect of resolution. Figure 2 illustrates qualitatively that prediction

accuracy is not seriously degraded by aggressive downsampling of up to 16X.

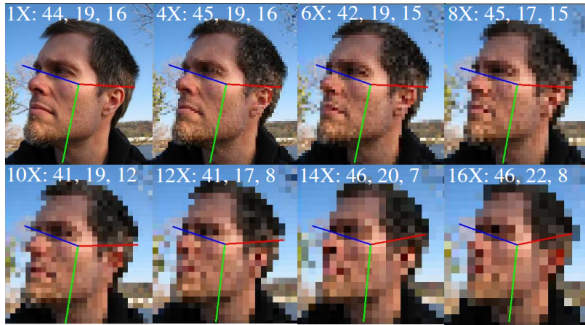


Figure 2: Downsampling factor vs. yaw, pitch & roll. Ground-truth values were 47.6, 22.0, 18.8. Images were downsampled by indicated amount and then resized to their original size using nearest-neighbor interpolation before being supplied to WHENet. Head pose predictions remain relatively stable event when images are aggressively downsampled by up to 16X. Original image from [10]

We carried out this test in aggregate on the AFLW2000 dataset. The results are shown in Figure 3 and compared to Hopenet [8] and FSANet [11]. We list the smallest reported errors for Hopenet among the four training strategies in [8] and thank the authors for providing this data.

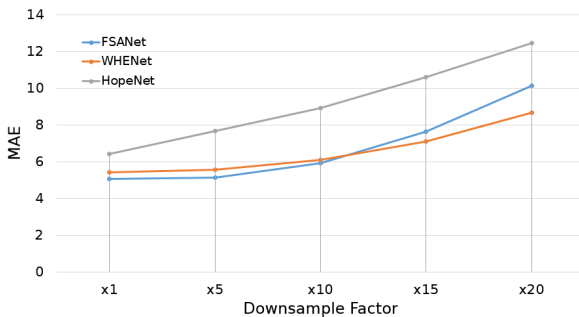


Figure 3: Effect of downsampling factor on MAE. WHENet (orange) shows consistent improvement over the already impressive Hopenet [8] and FSANet [11] performance (grey and black). For Hopenet we plot the minimum (best) value at each downsampling factor among all training strategies reported in [8]

In summation, full-range WHENet targets a task that is outside the scope of the existing state-of-the-art using a faster and significantly smaller network. In spite of this, it meets or beats state-of-the-art performance for the restricted case of HPE for frontal-to-profile views when evaluated on two datasets that were not used during training.

3 Applications

Here we show qualitative examples of WHENet applied to several applications that demonstrate how HPE can integrate with real-world systems and how our training strategy allows the method to generalize to low-resolution and low-quality data that was not present during training.

Figure 4 shows using a pose detector based on Lightweight OpenPose [9] code to detect pose keypoints while using WHENet to predict head pose. Frequently pose-estimations do not estimate sufficient keypoints for accurate HPE but by incorporating a full-range HPE method such as WHENet, such limitations may be overcome. This could be used, for example, in sports broadcasting or by coaching staff to estimate participants fields of views and situational awareness when analyzing plays.

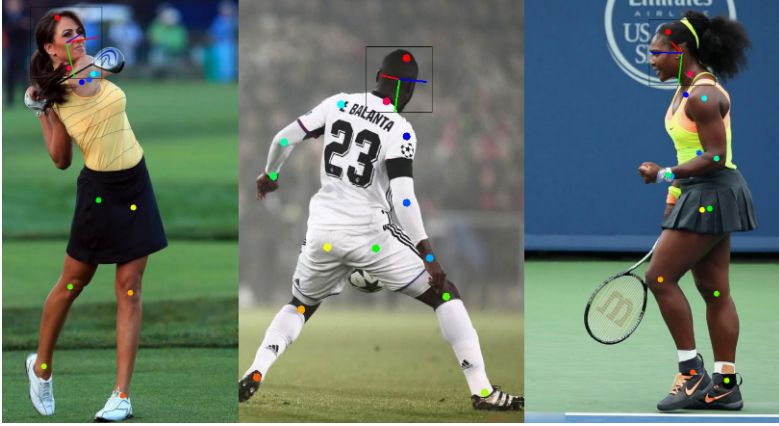


Figure 4: WHENet applied to head crops generated from keypoint predictions from [9], keypoints shown as dots, illustrating how HPE can be integrated with full-body pose estimation methods. Images from [9]

Figure 5 depicts a hypothetical driver-attention module where drivers are considered attentive with camera-relative yaw $< 30^\circ$ and inattentive otherwise. The extension to full-range could extend this to predicting blind spots during other activities such as reversing without requiring additional hardware.

References

- [1] Shabnam Abtahi, Mona Omidyeganeh, Shervin Shirmohammadi, and Behnoosh Hariiri. Yawdd: A yawning detection dataset. In *Proceedings of the 5th ACM Multimedia Systems Conference*, pages 24–28. ACM, 2014.
- [2] Guido Borghi, Matteo Fabbri, Roberto Vezzani, Rita Cucchiara, et al. Face-from-depth for head pose estimation on depth images. *IEEE transactions on pattern analysis and machine intelligence*, 2018.
- [3] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets

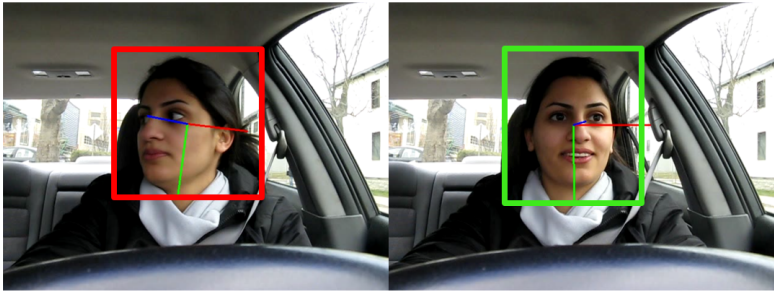


Figure 5: Applications to autonomous driving and driver assistance. Left: Green boxes indicate yaws $< \pm 45^\circ$ and potential awareness of vehicle, red boxes indicate probable inattention. This example highlights the need for efficient and low-resolution approaches to HPE with 6 total low-resolution detections. Here low-quality pose-estimates yield poor cropping regions but WHENet successfully generalizes despite having no comparable training data. Images from [1]. Right: WHENet is used to monitor driver attention, marking the driver as inattentive (red) when yaw exceeds 30° and attentive (green) otherwise. Images from [10]

robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11): 1231–1237, 2013.

- [4] Jinwei Gu, Xiaodong Yang, Shalini De Mello, and Jan Kautz. Dynamic facial analysis: From bayesian filtering to recurrent neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1548–1557, 2017.
- [5] Manuel Martin, Florian Van De Camp, and Rainer Stiefelhagen. Real time head model creation and head pose estimation on consumer depth cameras. In *2014 2nd International Conference on 3D Vision*, volume 1, pages 641–648. IEEE, 2014.
- [6] Sankha S Mukherjee and Neil Martin Robertson. Deep head pose: Gaze-direction estimation in multimodal video. *IEEE Transactions on Multimedia*, 17(11):2094–2107, 2015.
- [7] Daniil Osokin. Real-time 2d multi-person pose estimation on cpu: Lightweight open-pose. *arXiv preprint arXiv:1811.12004*, 2018.
- [8] Nataniel Ruiz, Eunji Chong, and James M Rehg. Fine-grained head pose estimation without keypoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2074–2083, 2018.
- [9] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipai Zhou, Guosen Lin, Yanwei Fu, et al. Ai challenger: a large-scale dataset for going deeper in image understanding. *arXiv preprint arXiv:1711.06475*, 2017.
- [10] Tsun-Yi Yang, Yi-Hsuan Huang, Yen-Yu Lin, Pi-Cheng Hsiu, and Yung-Yu Chuang. Ssr-net: A compact soft stagewise regression network for age estimation. In *IJCAI*, volume 5, page 7, 2018.
- [11] Tsun-Yi Yang, Yi-Ting Chen, Yen-Yu Lin, and Yung-Yu Chuang. Fsa-net: Learning fine-grained structure aggregation for head pose estimation from a single image. In

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1087–1096, 2019.

- [12] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016.