HHP-Net: A light Heteroscedastic neural network for Head Pose estimation with uncertainty Supplementary material

Giorgio Cantarini[1, 2], Federico Figari Tomenotti[1], Nicoletta Noceti[1], and Francesca Odone[1]

[1]MaLGa-DIBRIS, Università degli Studi di Genova, via Dodecaneso 35, 16146-IT Genova,Italy
[2]IMAVIS srl, via Trento 5/2, 16145-IT Genova,Italy
giorgio.cantarini@imavis.com, federico.figaritomenotti@edu.unige.it, nicoletta.noceti, francesca.odone@unige.it

1. Derivation of our loss function

In this section we report the derivation of the loss function presented in Section 3 of the main paper. We target scenarios where uncertainty may be due to data noise and varying on different inputs. Considering a typical regression problem where we want to estimate a function f_ω from the input x_i to the output y_i, we can formalize the setting as

$$y_i = f_\omega(x_i) + \epsilon(x_i)$$

where the output can be seen as the sum between a function $f_\omega(x_i)$ and $\epsilon(x_i)$ that is the noise that depends on the input x_i [4].

To quantify the uncertainty, the model is trained to learn a function that estimates both the mean and the variance of a target distribution using a maximum-likelihood formulation of a neural network [1, 3, 6]: in order to do that, we need to assume that the errors are normally distributed $\epsilon(x_i) \sim \mathcal{N}(0, \sigma(x_i)^2)$.

The likelihood for each point x_i is:

$$p\left(y_i | x_i; \omega\right) = \mathcal{N}\left(f_\omega\left(x_i\right), \sigma\left(x_i\right)^2\right)$$
$$= \frac{1}{\sqrt{2\pi\sigma\left(x_i\right)^2}} \exp\left[-\frac{\left(y_i - f_\omega\left(x_i\right)\right)^2}{2\sigma\left(x_i\right)^2}\right]$$

where y_i is the mean of this distribution and $\sigma(x_i)^2$ is the variance.

The neural network architecture should be modified to include an additional term to the output layer, to predict the variance (or the logarithm of it): this latter quantifies the uncertainty associated with the prediction based on the noise in the training samples (the uncertainty is a function of the input e.g. if the noise is uniform over all the input values, the uncertainty should be constant).

Applying the logarithm to both sides we obtain:

$$\log p\left(y_i | x_i; \omega\right)$$
$$= -\frac{\left(y_i - f_\omega\left(x_i\right)\right)^2}{2\sigma\left(x_i\right)^2} - \frac{1}{2}\log\sigma\left(x_i\right)^2 - \frac{1}{2}\log(2\pi)$$

that is the log likelihood we want to maximize (the last term is ignored in the following being a constant).

Maximizing the log likelihood is equivalent to minimizing the negative log likelihood, and therefore we rewrite the minimization problem as:

$$\min_\omega -\frac{1}{N}\sum_{i=1}^{N} \log p\left(y_i | x_i; \omega\right)$$

Finally the objective we want to minimize over all x_i becomes:

$$\frac{1}{N}\sum_{i=1}^{N} \frac{1}{2\sigma\left(x_i\right)^2}\left\|y_i - f_\omega\left(x_i\right)\right\|^2 + \frac{1}{2}\log\sigma\left(x_i\right)^2$$

In order to solve possible numerical issues the objective is modified in this way:

$$\sum_{i=1}^{N} \frac{1}{2}\exp\left(-s_i\right)\left\|y_i - f_\omega\left(x_i\right)\right\|^2 + \frac{1}{2}s_i$$

where $s_i = \log\sigma(x_i)^2$: in this way potential divisions by zero are avoided [2].

Lastly we extend this objective for a multi-ouput regression model for training our network obtaining the objective proposed in Section 3 of the main paper.

Table 1. Comparison among different losses (see text). **All errors are expressed in degrees** (°): err$_y$= yaw error, err$_p$=pitch error, err$_r$= roll error, MAE = Mean Absolute Error.

Train	Val	Loss	err$_y$	err$_p$	err$_r$	MAE
BIWI	BIWI	MSE	2.90	4.80	3.34	3.70
BIWI	BIWI	COMB	3.15	4.85	3.40	3.80
BIWI	BIWI	**UNC**	3.04	4.79	3.21	**3.68**
300WLP	BIWI	MSE	4.75	6.65	4.45	5.28
300WLP	BIWI	COMB	4.67	8.08	4.87	5.88
300WLP	BIWI	**UNC**	4.14	7.00	4.40	**5.18**
300WLP	AFLW2000	MSE	5.72	10.41	8.08	8.07
300WLP	AFLW2000	COMB	5.55	10.39	8.18	8.04
300WLP	AFLW2000	**UNC**	5.26	10.12	7.73	**7.70**
AFLW	AFLW2000	MSE	7.60	6.43	4.76	6.26
AFLW	AFLW2000	COMB	7.31	6.55	4.68	6.18
AFLW	AFLW2000	**UNC**	7.40	6.63	4.47	**6.16**

2. Removing the uncertainty: an ablation study

In this section we provide more details about the ablation study discussed in Sec. 4.3 of the main paper where we consider two variations of our method:

MSE: we directly regress the three angles adopting a loss computed as the sum of the Mean Squared Error (MSE) on each angle:

$$\mathcal{L}_{MSE} = \sum_{i \in \{y,p,r\}} \left\| q_i - f_i\left(\mathbf{x}_1, \mathbf{x}_2, \mathbf{c}\right) \right\|^2 . \quad (1)$$

where $\mathbf{q} = [y, p, r]$ (y=way, p=pitch and r=roll).

COMB: we employ an alternative loss function $\mathcal{L}_{COMB}$ proposed in [5] which has been proved to be very successful on the same estimation task. The loss allows to jointly solve a regression and classification tasks, and it can be formalized as follows $\mathcal{L}_{COMB}$:

$$\sum_{i \in \{y,p,r\}} \sum_{j} -q_j \log\left(f_j\right) + \alpha * \left\| q_i - f_i\left(\mathbf{x}_1, \mathbf{x}_2, \mathbf{c}\right) \right\|^2 \quad (2)$$

combining the cross entropy loss, computed between the binned angles, and the MSE loss, computed between the scalar angles; α is an hyperparameter that controls the weight of the regression loss (in the experiment we set $\alpha = 1$).

In Tab. 1 we extend Table 1 of the main paper and report the angular errors we obtain with the three different losses. As it can be observed, learning the angles associated with the uncertainty provides the best average performance, showing the benefit of the uncertainty not only in terms of interpretability of the model but also as a way to improve its effectiveness.

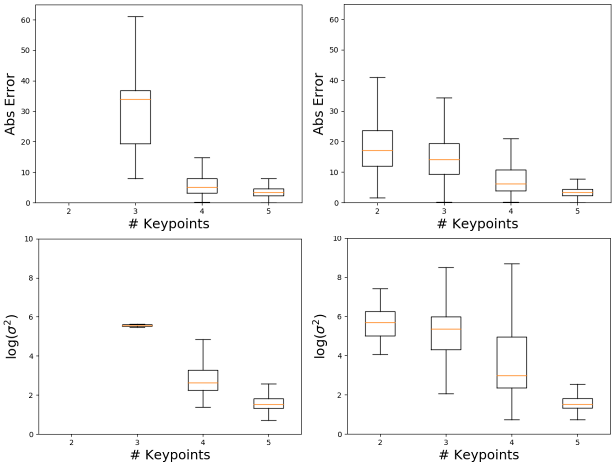

Figure 1. Performance of our method (top row: mean angular error, bottom row: uncertainty) with respect to the number of input points, considering the output of Open Pose (above) and randomly dropping points from the input (below). We used a model trained on 300W-LP and tested on the whole BIWI (plots refer to the latter).

3. On the number of keypoints

In this section we provide an extended version of the assessment we discuss in Sec. 4 of the main paper with the aim of observing the influence of the quality and quantity of input semantic features on the final head pose estimate. For the sake of the discussion, we report here plots and comments already included in the main document.

In Fig. 1, first column, we analyse the performance of our method in terms of uncertainty values (bottom-left) and absolute angular error (top-left) as we group the input data according to the number of keypoints provided by the Open Pose detector. When only 3 keypoints are available the uncertainty is rather high on average. Increasing the number of points it is progressively reduced, with a similar trend shown by the error. This confirms the intuition that the more input points the method has, the higher is its confidence in the prediction, which is more reliable and accurate.

Since the worst case scenario corresponds to having at least three keypoints, we randomly dropped points from the input to evaluate the behavior of the method in more challenging scenarios. The results are shown in the second column of Fig. 1. When points are randomly dropped, we only consider samples with more than two points.

When all the 5 keypoints are available, the uncertainty is compactly lower (confirming what already observed in the previous experiment) as the method can rely on a more comprehensive representation of the input. In the intermediate cases – where we may have 2, 3, or 4 keypoints available in input – the uncertainty progressively decreases, but we also have a higher degree of variability, as some keypoints configurations are more significant than others

Table 2. Comparison among models with different sizes (Protocol P1: 300W-LP train, BIWI test). α = neurons reduction factor (see text), MAE = Mean Absolute Error

α	**MAE**	**Mult-adds**	**Parameters**	**MB**
1	5.18	93000	94000	0.385
0.6	5.43	37000	37000	0.158
0.2	5.54	6000	6000	0.032

and thus the amount of information they provide to the model may be uneven reflecting the concept that the noise could be different for each input sample. With respect to the plots in the first column of Fig. 1, the box plots at right show a higher standard deviation since randomly dropping points from the input we simulate a higher variability in the input configurations with respect to the ones usually provided by Open Pose and from the datasets we used.

4. Model size and parameters

In this section we show the robustness of our method with respect to reductions of size, that may be needed when the available computational resources are very limited. More specifically, we analyse how the performance changes as we reduce the size of the model. We choose 300W-LP training and BIWI test (protocol P1) for its larger training and test sets and decrease the number of neurons in the fully connected layers so the backbone remains the same proposed in the paper, while its size decreases. Given a reduction factor $\alpha \in (0, 1)$, we obtain a “reduced” version of our architecture by multiplying the original number of neurons in each layer (250, 200 and 150 in, respectively, the first, second and third layer) by α.

By varying α in the range $(0, 1)$ we reduce the model size (the number of parameters) and thus also the number of sum and multiplication operations. Tab. 2 compares our baseline ($\alpha = 1$) with two reduced models (overall size in MB up to $10\times$ smaller) causing a very limited degradation in the Mean Absolute Error (below 1 degree). This experiment highlights the possibility of further reducing the size of the architecture, with a very limited performance loss, if required by the system.

References

[1] Wray L. Buntine and A. Weigend. Bayesian backpropagation. *Complex Syst.*, 5, 1991.

[2] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[3] David J. C. MacKay. A practical bayesian framework for backpropagation networks. *Neural Comput.*, 4(3):448–472, May 1992.

[4] D.A. Nix and A.S. Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN’94)*, volume 1, pages 55–60 vol.1, 1994.

[5] Nataniel Ruiz, Eunji Chong, and James M. Rehg. Fine-grained head pose estimation without keypoints. In *CVPR Workshops*, 2018.

[6] D. Rumelhart, R. Durbin, R. Golden, and Y. Chauvin. Backpropagation: the basic theory. 1995.