Informed MCMC with Bayesian Neural Networks for Facial Image Analysis

Adam Kortylewski[*], **Mario Wieser**[*], **Andreas Morel-Forster**[*], **Aleksander Wieczorek, Sonali Parbhoo, Volker Roth, Thomas Vetter**
Department of Mathematics and Computer Science
University of Basel

1 Introduction

Motivation. Computer vision tasks are difficult because of the large variability in the data that is induced by changes in light, background, partial occlusion as well as the varying pose, texture and shape of objects. Generative approaches to computer vision allow us to overcome this difficulty by explicitly modeling the physical image formation process. Such models can produce images $R(y)$ using a deterministic rendering engine R and a set of parameters y that define the scene in terms of e.g. light sources and object properties. The analysis of an observed image x is then performed via Bayesian inference of the posterior distribution $p(y|x)$.

Problem. This conceptually simple approach tends to fail in practice because of several difficulties stemming from sampling the posterior distribution [4]: high-dimensionality and multi-modality of the posterior distribution $p(y|x)$ as well as expensive simulation of the rendering process. Sampling $p(y|x)$ is typically performed with a Markov Chain Monte Carlo algorithm, such as Metropolis-Hastings [9, 4]. The general idea is to sequentially generate samples from the posterior distribution of y by performing the following two steps:

1. Generate a new point from the proposal distribution: $y_{t+1} \sim Q(\cdot|y_t)$.
2. Accept the new point with the acceptance probability:
   $A(y_{t+1}, y_t) = \min\left(1, \frac{p(y_{t+1})Q(y_t|y_{t+1})}{p(y_t)Q(y_{t+1}|y_t)}\right)$.

The main difficulty of MCMC in a computer vision context is how to choose the proposal distribution accurately so that maxima of the posterior are explored early and the Markov chain quickly converges to a valid image interpretation.

Contribution. In this work, we propose to use a Bayesian Neural Network for estimating an *image dependent* proposal distribution $Q(\cdot|x)$. Compared to a standard Gaussian random walk proposal, this will accelerate the sampler in finding regions of the posterior with high value. In this way, we can significantly reduce the number of samples needed to perform facial image analysis.

2 Methodology

Generative Face Model. In the context of facial image analysis, the 3D Morphable Model (3DMM) [2] is commonly used as prior for the 3D face geometry, color as well as the computer graphics parameters needed for the rendering process. Schönborn et al. [9] proposed using the Metropolis Hastings algorithm to estimate the posterior over the model parameters y:

$$p(y|x) \sim p(x|y)p(y). \tag{1}$$

[*]Equal contribution.

Third workshop on Bayesian Deep Learning (NeurIPS 2018), Montréal, Canada.

The likelihood $p(x|y)$ measures the similarity between the target image x and the rendered image $R(y)$ assuming pixel-wise independence. Given a posterior estimate, we can perform a multitude of facial image analysis tasks, such as face recognition [1], 3D face reconstruction [9] or face manipulation [10].

Informed Sampler. A key component of an MCMC sampler is the proposal distribution $Q(y)$. In the context of computer vision, $Q(y)$ needs to be carefully tuned in order to explore the posteriors maxima in a reasonable time. In order to overcome this limitation Jampani et al. [4] propose to combine a local Gaussian random walk proposal Q_L with an image dependent, global proposal distribution Q_I:

$$y_{t+1} \sim \alpha Q_L(\cdot|y_t) + (1-\alpha)Q_I(\cdot|x). \tag{2}$$

The global proposal distribution is estimated discriminatively based on the input image. In [4] the authors propose to use manually designed image features and a kernel density estimate for estimating $Q_I(\cdot|x)$. We instead propose to learn this distribution from data using Bayesian Neural Networks.

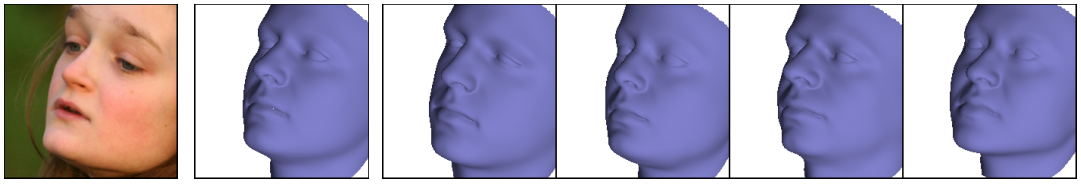
Bayesian Neural Networks. A Bayesian Neural Network (BNN) estimates, in contrast to traditional Neural Networks, not only a point estimate but also the corresponding uncertainties. In [5], Kendall and Gal describe model (Epistemic) and data (Heteroscedastic Aleatoric) uncertainties to be crucial for computer vision tasks and introduce an approach to unify both uncertainties within a BNN. We build upon this approach and estimate our global distribution $Q_I(\cdot|x)$ with a BNN which is in turn used to inform the MCMC sampler. In doing so, we place a prior distribution over the neural network weights W to capture the model uncertainty. We estimate the posterior distribution of W during training using Bayesian inference given our training data $X = \{ x_1, \ldots, x_N\}$ and $Y\{ y_1, \ldots, y_N\}$:

$$p(W \mid X, Y) = \frac{p(X, Y \mid W)p(W)}{\int p(Y \mid X, W)p(W)dW} \tag{3}$$

Subsequently, we formulate our data uncertainty in terms of a Gaussian likelihood $p(y \mid f^W(x)) = \mathcal{N}(f^W(x), \sigma^2)$ because the 3DMM parameters Y are continuous values. Here, the mean is denoted as our model output $f^W(x)$ and σ^2 defines the corresponding variance. Having estimated both uncertainties, we combine these uncertainties as described in [5] to obtain our informed proposal distribution $Q_I(\cdot|x)$.

3 Experiments

Datasets and setup. We train our BNN on synthetic data in similar to as proposed by Kim et al. [6]. We train an AlexNet [7] architecture using 300K synthetically generated face images with corresponding 3DMM parameters $\{X, Y\}$. The code and data used for our experiments will be made available [2]. For testing, we use a sample of 150 face images from the CMU-Multipie face dataset [3], sampled from Session-01 using the frontal and 30° cameras.



(a) (b) (c)

Figure 1: Uncertain 3D face reconstruction with BNNs. We illustrate the joint pose and shape distribution for simplicity. Note that our model predicts a joint distribution over all 3DMM parameters. (a) The test image from the AFLW dataset [8]. (b) The mean prediction of our BNN. (c) Samples from the joint prediction uncertainty in head pose and 3D shape. Notably, the variability in the head pose is low, whereas the remaining variability in the shape is comparably high (e.g. the nose region).

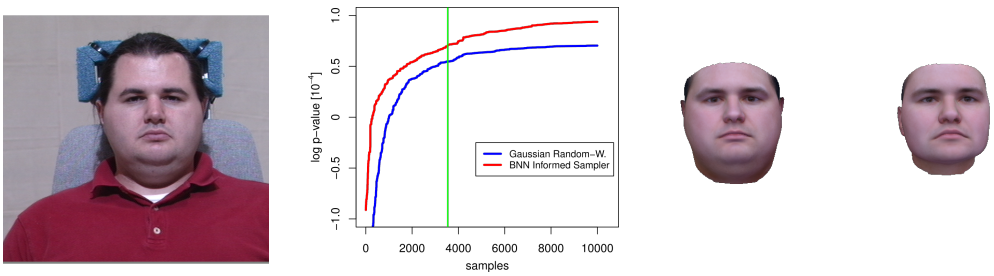
Qualitative results. Figure 1 illustrates the prediction uncertainty of our model given the test image in Figure 1a. Note that the mean prediction (Figure 1b) has a correct head pose and a similar 3D face shape as the face in the test image. The prediction uncertainty of our model (Figure 1c) is visualized

[2] https://github.com/unibas-gravis/bnn-informed-face-sampler

by sampling from the normal distribution defined by the mean prediction and the joint uncertainty estimated as described in the previous section. Note the remaining variability in the head pose is low, whereas it is large in the shape, e.g. in the nose region. This observation is reasonable as the 3D head pose can in principle be estimated from the 2D spatial configuration of a few facial features, whereas the estimation of the 3D face geometry from a single monocular image is ill-posed.

Quantitative results. We integrate the uncertain prediction of our BNN into Markov Chain Monte Carlo (MCMC) sampling as an informed proposal distribution $Q_I(\cdot|x)$. In Figure 2 we compare our BNN-informed sampling to one with an uninformed block-wise Gaussian proposal distribution when applied to the test image in Figure 2a. When plotting the maximal unnormalized posterior over runs of 10000 samples we can observe that the proposed BNN-informed MCMC (red curve) explores samples with high posterior values earlier compared to the uninformed sampler (blue curve). From the plot, we can also see that the BNN-informed sampler reaches the maximal posterior value of the uninformed sampler already after about 3500 samples (green vertical line). Overall this results in a better image interpretation (Figure 2c) compared to the uninformed sampler (Figure 2d).

We test the significance of our result by evaluating the informed and uninformed samplers over a population of 150 face images from the CMU-Multipie dataset. We compare the maximal posterior observed with a ranked Friedmann test and obtain a p-value of 5.7×10^{-5}. This result highlights the superiority of our approach over an uninformed sampler in terms of exploring the maxima of the posterior within a fixed frame of samples.



(a) (b) (c) (d)

Figure 2: Comparison of uninformed and BNN-informed MCMC. (a) The test image. (b) Maximal observed posterior over 10K samples for both MCMC approaches. Our informed sampler (red curve) explores high posterior values faster than the uninformed sampler (blue curve). It also reaches the same maximal posterior value after already about 3500 samples (green vertical line). Therefore it can obtain a better image interpretation (c) compared to an uninformed sampler (d) within a fixed frame of 10K samples.

4 Discussion

We have presented a novel approach to inform MCMC sampling with Bayesian Neural Networks. In our experiments we demonstrate that:

BNNs allow for the estimation of an image-dependent proposal distribution. Our qualitative results indicate that the BNN estimate is a meaningful measure of the uncertainty in the 3D face reconstruction process (Figure 1).

BNN-Informed MCMC significantly improves the exploration of maximal posterior regions compared to an uninformed Gaussian random walk. An extensive evaluation of our approach on a population of 150 face images demonstrated a highly significant improvement in terms of the observed the face reconstruction quality (Figure 2).

Acknowledgment

A.K. is supported by a Novartis University of Basel Excellence Scholarship for Life Sciences. M.W., A.W. and S.P are partially supported by the Swiss National Science Foundation (SNF), SystemsX.ch and the National Center of Competence in Research MARVEL. We gratefully acknowledge the support of NVIDIA with the donation of a Titan Xp.

References

[1] Volker Blanz and Thomas Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Transactions on pattern analysis and machine intelligence*, 25(9):1063–1074, 2003.

[2] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schönborn, and Thomas Vetter. Morphable face models-an open framework. In *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*, pages 75–82. IEEE, 2018.

[3] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-PIE. *Image and Vision Computing*, 28(5):807–813, May 2010.

[4] Varun Jampani, Sebastian Nowozin, Matthew Loper, and Peter V Gehler. The informed sampler: A discriminative approach to bayesian inference in generative computer vision models. *Computer Vision and Image Understanding*, 136:32–44, 2015.

[5] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5574–5584. Curran Associates, Inc., 2017.

[6] Hyeongwoo Kim, Michael Zollhöfer, Ayush Tewari, Justus Thies, Christian Richardt, and Christian Theobalt. Inversefacenet: Deep single-shot inverse face rendering from a single image. *arXiv preprint arXiv:1703.10956*, 2017.

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[8] Peter M. Roth Martin Koestinger, Paul Wohlhart and Horst Bischof. Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization. In *Proc. First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.

[9] Sandro Schönborn, Bernhard Egger, Andreas Morel-Forster, and Thomas Vetter. Markov chain monte carlo for automated face image analysis. *International Journal of Computer Vision*, 123(2):160–183, 2017.

[10] Mirella Walker and Thomas Vetter. Portraits made to measure: Manipulating social judgments about individuals with a statistical face model. *Journal of Vision*, 9(11):12–12, 2009.