Structured Aleatoric Uncertainty in Human Pose Estimation

Nitesh B. Gundavarapu[1] Divyansh Srivastava[1] Rahul Mitra[1] Abhishek Sharma[2]
Arjun Jain[1]
[1]Indian Institute of Technology, Bombay [2]Axogyan AI, Bangalore

Abstract

Human pose estimation from monocular images exhibits an inherent uncertainty through self-occlusions and inter-person occlusions, aside from typical sources of uncertainty. Recently, there has been an increased focus in modelling uncertainty in supervised machine learning tasks. In line with this trend, we propose a novel formulation to capture aleatoric uncertainty in human pose using a multivariate Gaussian distribution over all the joints of human body and show that this improves generalization in 2D human pose estimation by implicitly suppressing the gradients from uncertain joints. Further, we develop a novel method to triangulate 3D human pose from predicted 2D poses, under the predicted uncertainty, that out-performs the baselines by over 10.8% *and provide a multi-view inference benchmark for 3D human pose estimation on Human 3.6M dataset.*

1. Introduction

Human pose estimation has evolved over the recent years by virtue of complex CNNs [9, 10] and availability of large scale datasets [6, 1]. However, monocular images of humans are prone to occlusions from the self, another person or objects. In all these cases, the corresponding posterior for human pose, conditioned on the image, is inherently stochastic in nature. So it is imperative for the neural network to return a distribution on the expected joint locations as opposed to confidently localizing to unique locations. In Figure 1, we demonstrate few such cases where predicting to one specific location is not appropriate.

Uncertainty in neural network predictions are typically either *epistemic* or *aleatoric* in nature [2, 7, 3], where the former captures the uncertainty from the model while the latter captures uncertainty of the input data. The kind of uncertainty we wish to capture is *aleatoric* and cannot be reduced by collecting more data as opposed to *epistemic* uncertainty. Since the human pose is anthropomorphically constrained, we consider the joint locations as correlated stochastic variables and propose to predict a multi-variate

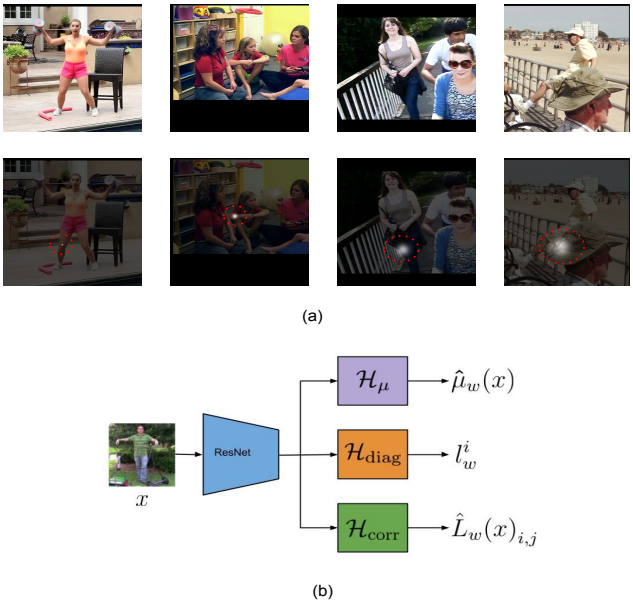

Figure 1: (a) Images with increasing levels of predicted uncertainty on **Right Knee** joint on MPII validation set (best viewed when zoomed in). Red circles are visual aids while the white region represents actual Gaussian field. Images towards the right have severe occlusions and consequently, our network assigns higher uncertainty. (b) Network architecture

Gaussian distribution over the joint locations. To the best of our knowledge, this is the first work that explicitly captures uncertainty in human pose estimation.

In Section 2, we detail our problem formulation. Unlike [2], we are able to accomplish end-to-end training to jointly predict a full rank covariance matrix along with the expected joint locations. In Section 3, we empirically and qualitatively verify the significance of the predicted uncertainty on MPII dataset [1]. Our joint prediction scheme also helps in generalization i.e. our model outperforms the baseline without uncertainty by a significant margin. The intuition is that network learns to reduce the the contribution from occluded/noisy joints' locations by increasing their corresponding uncertainty when computing the loss. Section 4 provides the implementation details.

In Section 5, we propose to exploit the predicted uncertainties to weigh the residuals while triangulating 3D pose from predicted 2D locations in multi-view images. We demonstrate an out-standing improvement of over 10.8% over the baseline method through this simple improvisation on the Human 3.6M dataset [6] and provide a strong benchmark for multi-view 3D pose estimation.

2. Problem Formulation

We model joint locations as a multivariate Gaussian random variable $y \in \mathcal{Y}$, $\mathcal{Y} \subset \mathcal{R}^{n \times k}$ conditioned over image $x \in \mathcal{X}$, $\mathcal{X} \subset \mathcal{R}^{h \times w \times 3}$ where n is total number of joints, k is 2, 3 for 2D and 3D respectively, w is image width and h is image height. The posterior probability of y conditioned on x is given by Eq. 1

$$p(y|x) = N(\mu(x), \Sigma(x)) \tag{1}$$

Given an Image $x \in \mathcal{X}$, a neural network with parameters w can be used to estimate $\hat{\mu}_w(x)$ and $\hat{\Sigma}_w(x)$. A negative logarithm of maximum likelihood (shown in Eq. 2) is minimized with respect to w to train this network.

$$\begin{aligned}\mathcal{L} = \arg\min_w \; & log(|\hat{\Sigma}_w(x)|) \\ & + (y - \hat{\mu}_w(x))^T (\hat{\Sigma}_w(x))^{-1} (y - \hat{\mu}_w(x))\end{aligned} \tag{2}$$

If $\hat{\Sigma}_w(x)$ is directly estimated using a neural network, it needs to be inverted which could result in numerical instability. Hence, it is more practical to estimate $\hat{\Psi}_w(x) = \hat{\Sigma}_w(x)^{-1}$ called *Precision* matrix. The *Precision* matrix is symmetric and positive definite with $(s^2 - s)/2 + s$ unique parameters where $s = n \times k$. Hence, we represent *Precision* matrix with it's Cholesky decomposition (shown in Eq. 3) and restrict diagonals terms to be positive to ensure a unique decomposition and positive definiteness for *Precision* matrix.

$$\hat{\Psi}_w(x) = \hat{L}_w(x) \hat{L}_w(x)^T \tag{3}$$

where L is a lower triangular matrix. To ensure positive diagonal terms in L, our networks outputs $l_w^i = -log(\hat{L}_w(x)_{ii})$. The modified loss function is given in Eq. 4

$$\begin{aligned}\mathcal{L} = \arg\min_w \; & 2 * \sum_{n=1}^{s} l_w^i \\ & + ||(y - \hat{\mu}_w(x))^T (\hat{L}_w(x))||^2\end{aligned} \tag{4}$$

3. 2D Human Pose Estimation

In this section, we use the uncertainty formulation from Section 2. to model the uncertainty in 2D human pose estimation task under two settings i) Diagonal Covariance Matrix, herewith called *diagonal* iii) Full Convariance Matrix, herewith called *full*.

3.1. Significance Of Uncertainty

To understand the significance of the predicted uncertainty, we sort the images using i) Entropy $log(|\hat{\Sigma}_w(x)|)$, ii) Variance on right knee iii) Variance on left elbow. We observed that the images with highest predicted uncertainties in each of the cases have heavy occlusions or have multiple people cluttered together in the image. Images with low predicted uncertainty have the corresponding joints clearly visible. Qualitative results are presented in Figure 1. Table. 1 shows the PCKh for top and bottom K images sorted based on uncertainty.

Joint Type	Top K PCKh			Bot. K PCKh		
	K=1	50	500	K=1	50	500
All joints	0.1	0.13	0.16	1.0	1.0	0.99
R.Knee	0.0	0.22	0.49	1.0	1.0	0.99
L.Elbow	0.0	0.38	0.55	1.0	1.0	0.99

Table 1: PCKh values of the top K and bottom K images from MPII validation set sorted in descending order of uncertainty.

To isolate the effect of occlusion on uncertainty, we add synthetic occlusions by placing black square patches of increasing size on the left elbow joint. The corresponding results are presented in Fig . 2 and Table. 2. We can observe a clear increase in uncertainty with occlusion size. Further, we can observe uncertainty propagation through the covariance terms, with marked increase in adjacent joints. This is in line with the anthropometric constraints.

Occlusion	Avg. L.El. Standard Deviation	
	diagonal (in px)	*full (in px)*
0px	9.99	8.96
10px × 10px	10.69	9.91
20px × 20px	11.93	11.3
30px × 30px	13.36	13.06
40px × 40px	14.79	14.43

Table 2: Average standard deviation due to synthetic occlusion of Left Elbow on MPII validation set. We observe an increasing trend for both the trained models with increase in occlusion

4. Network Architecture and Training

We use a ResNet-50 [5] backend with three heads respectively for expectation ($\hat{\mu}_w$), diagonal (l_w^i) and off-diagonal ($\hat{L}_w$) terms. The backend and expectation head is exactly same as used in [10]. We use additional average pooling heads for diagonal and off-diagonal predictions, and further two hidden layers of size 2048 with ReLU activations, batch normalization and dropout for the off-diagonal predictions. Schematic diagram is presented in Figure 1. Training procedure is divided into three stages - In first stage, we train

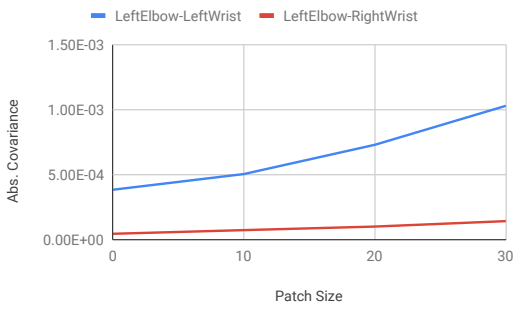

Figure 2: Comparison of average of absolute covariance over MPII dataset with increasing occlusion on left elbow. We observe that absolute covariance increases for left wrist but remains flat for right wrist implying propagation of uncertainity between directly connected joints

using only L1 loss on expectation branch, followed by joint training of expectation and diagonal branches by assuming a diagonal covariance matrix. Finally, we train for all three branches using loss from Eq. 4. We use the Adam optimizer [8] with an initial learning rate of 1e-3 and an L2 weight regularization of 1e-5.

4.1. Generalization

Starting from the architecture in [10] trained using an L1 Loss, we observed that the generalization improves after adding an uncertainty based loss as defined in Eq. 4. Table. 3 demonstrates this observation on MPII validation set, as is the common practice for ablation studies [10, 9].

Method	PCKh@0.5	PCKh@0.1
Sun[10] - Direct	84.6	25
Diagonal	85.6	27.65
Full	**86.2**	**28.4**

Table 3: Comparing PCKh of *diagonal* covariance and *full* covariance to baseline on MPII validation set. Direct refers to direct regression. Note that our objective is not to compare against state-of-the-art, but to show the improved generalization achieved by learning to predict uncertainty.

5. Triangulating 3D pose

In this section, we utilize our learned uncertainty measure to improve accuracy of triangulating 3D joint positions from their 2D predictions in multi-view images [4]. Noisy 2D predictions are geometrically inconsistent and reduce the triangulation accuracy. To this end, we compute confidence of prediction of every 2D joint position from the modelled uncertainty and assign weights to its corresponding least square residuals according to the computed confidence. A higher weight is assigned to a joint when its uncertainty is lower. This is in coherence with our observations in Section 3.1, where larger values uncertainty are captured when joints are incorrectly predicted or occluded. In Eq. 5, we present our modified triangulation objective to obtain the optimum 3d position of the i^{th} joint p^{i*} with $i = 1 \dots N$ with N as the no. joints in the human pose. Let $j = 1 \dots M$ with M as the number of views.

$$p^{i*} = \arg\min_{p^i} \sum_{j=1}^{M} W^{ij} * \|x^{ij} - \pi^j(p^i)\|_2$$
$$W^{ij} = (1/e^{2l^{ij}}) \; / \; \max_k (1/e^{2l^{ik}}) \tag{5}$$

where x^{ij} is the predicted 2D location for i^{th} joint in j^{th} view, W^{ij} is the corresponding weight applied to its residual and $l^{ij} = -log(\hat{L}^j_{ii})$ is the predicted diagonal term. The projection from 3D to 2D using camera intrinsic and extrinsic parameters in the j^{th} view is denoted by function π^j.

We justify our claims by performing 3D triangulation on the popular Human3.6M [6] dataset with $(M = 3, 4)$ views and $(N = 16)$ joints per skeleton. Corresponding results are presented in Table . 4. For 2D joints locations and its uncertainty, we use the the predictions of the pre-trained model used in Section 4 fined tuned on Human3.6M. We observe a highest gain of $\mathbf{10.8\%}$ when using 4 views.

Method	MPJPE(mm)	
	3 views	4 views
without-weights	40.50	36.80
with-weights	**39.20**	**32.72**

Table 4: Comparison of triangulated skeleton obtained by our modified formulation ('with') against the vanilla formulation ('without'). We perform two experiments, triangulating from all the 4 views and taking a subsets of 3 views and report MPJPE values. It is evident that weighing the prediction based on confidence improves the triangulation performance.

6. Conclusion

In this paper we model joints locations using a multivariate Gaussian to capture *aleatoric* uncertainty in human pose estimation. We improve upon the existing 2D human pose estimation baselines by obtaining a better fit to the data on MPII dataset. Further, we show our predicted uncertainty improves 3D pose triangulation from multi-view images by suppressing contributions of occluded or noisy residuals.

In future, we plan to extend this framework to different state-of-the-art human pose architectures and use the predicted uncertainty in motion tracking and future frame prediction problems.

References

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[2] G. Dorta, S. Vicente, L. Agapito, N. D. F. Campbell, and I. Simpson. Structured uncertainty prediction networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[3] Y. Gal and Z. Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML-16)*, pages 1050–1059, 2016.

[4] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, New York, NY, USA, 2 edition, 2003.

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[6] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.

[7] A. Kendall and Y. 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.

[8] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[9] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *The European Conference on Computer Vision (ECCV)*, September 2016.

[10] X. Sun, B. Xiao, F. Wei, S. Liang, and Y. Wei. Integral human pose regression. In *The European Conference on Computer Vision (ECCV)*, September 2018.