Active Learning for Bayesian 3D Hand Pose Estimation

Razvan Caramalau  Binod Bhattarai  Tae-Kyun Kim
Imperial College London
{r.caramalau18, b.bhattarai, tk.kim}@imperial.ac.uk

Abstract

We propose a Bayesian approximation to a deep learning architecture for 3D hand pose estimation. Through this framework, we explore and analyse the two types of uncertainties that are influenced either by data or by the learning capability. Furthermore, we draw comparisons against the standard estimator over three popular benchmarks. The first contribution lies in outperforming the baseline while in the second part we address the active learning application. We also show that with a newly proposed acquisition function, our Bayesian 3D hand pose estimator obtains lowest errors with the least amount of data. The underlying code is publicly available at: https://github.com/razvancaramalau/al_bhpe

1. Introduction

Hand Pose Estimation (HPE) is an active and important research topic. With the recent advancement in deep learning [19, 18, 3, 30, 17, 33, 28], pose estimation has become one of the key ingredients in Robotics, Augmented Reality (AR)/ Virtual Reality (VR), Human-Computer Interaction (HCI) and to mention but a few. In this work, we address the problem of Bayesian 3D hand pose estimation in the active learning setting. We consider a scenario where the hand skeleton is represented by volumetric information, captured by depth cameras, and where is a limited budget to annotate the hand skeleton. The downstream task is to estimate 3D coordinates of the pre-defined key locations of the hands. The advantage of depth-based representation being illumination and colour invariant motivated development of multiple large-scale depth-based benchmarks [37, 28, 29] and several international challenges [36, 35, 2].

Most of the success stories on 3D-HPE. [36, 35] are due to the models with a large number of learnable parameters and a large-scale of annotated data (in the order of $10^6$) such as BigHand2.2 [37]. To annotate such a dataset requires a lot of effort, is expensive and also time-consuming. It is essential to develop a methodology that identifies a small sub-set of the most influencing and discriminative examples to annotate. Active Learning (AL) frameworks [26, 9] have widely been used for a similar purpose.

The active learning framework is a well-studied research domain applied for several tasks such as image classification [9, 4, 21], semantic segmentation [26, 25], human pose estimation [1, 34]. However, it has not yet been addressed for the 3D HPE. In this paper, we systematically adapt the classical architecture of DeepPrior [19] in a Bayesian Convolutional Neural Networks, i.e. Bayesian DeepPrior. Besides, we present a novel AL selection method, the key component of this framework, optimised for the Bayesian learner. Finally, we evaluate it on three challenging benchmarks for 3D HPE.

AL frameworks primarily consist of two major components: learner and sampling technique. With the increasing trend of deep learning algorithms usage, the learner is approximated by a large-scale standard CNN, DeepPrior [19]. However, these frameworks ignore modelling two important uncertainties incurred due to noisy acquired data (aleatoric) and due to model’s lack of knowledge (epistemic). Modelling these uncertainties on discriminative models for semantic segmentation [12] and depth regression [13], has proven to be effective. As AL frameworks are principally designed to select the most influencing and discriminative examples, it is crucial to model both the uncertainties. To this end, we propose to approximate the learner by replacing standard DeepPrior with its
Bayesian adapted version similarly to [12]. To the best of our knowledge, this is the first work to employ a Bayesian 3D-HPE as a learner in an AL framework. The sampling technique is another important component to determine the fate of the AL framework. Existing acquisition functions such as Coreset [23] are widely and successfully been applied. However, the major limitations of this method consists in relying only on the fixed mean posterior probability while ignoring its epistemic variance. This mean value does not describe fully the complete characteristics of the predicted skeleton. Hence, we propose a novel sampling technique called CKE (a combination between Core-Set and epistemic uncertainty) which models both the upper and lower bound of the epistemic variance of the averaged predicted skeleton. Figure 1 shows some of the hand’s depth images and their skeleton annotations together with their corresponding uncertainties (represented as circle radii). We describe more details about our idea in Section 5.

To summarise the contributions:

- We formulate the 3D Hand Pose estimation problem under the active learning framework.
- We propose to approximate in a Bayesian fashion the DeepPrior 3D-HPE.
- Proposed a novel AL sampling technique incorporating both the predicted skeleton and its epistemic variance.
- We systematically evaluated active learning for both standard and Bayesian DeepPrior on three challenging 3D-HPE benchmarks: BigHand2.2, ICVL and NYU.
- The proposed method consistently outperforms the counter-part competitive baselines.

2. Related Work

3D Hand Pose Estimation. The first comprehensive review in hand pose estimation that established the current taxonomy is published in [6]. Together with the deep learning popularity and easy access to depth camera sensors, 3D HPE has gained a deep interest in the computer vision community. In terms of methodology, theoretical approaches have investigated 2D or 3D representations, detection-based [17], hierarchical and/or structured models [33, 19, 28] over single or multi-stage processes. Earlier state-of-the-art methods favoured combination of discriminative (random forest [28] or CNN-based [19]) and generative solutions [3] as in [33, 18]. Due to the past concepts from [5], volumetric representations [17] managed to outstand by compensating from the high non-linearity in direct regressions. These recent works [17, 30, 32] have obtained impressive results with average 3D joint errors below 10 mm on datasets like NYU [29], ICVL [28] or BigHand2.2 [37]. As we are tackling uncertainty exploration and data representatives of the model, for simplicity and efficient analysis we deploy a standard DeepPrior [19] architecture. Even though the accuracy is lower than current arts due to relatively lesser model parameters, our method is generic, the insights obtained can be easily transferred to deeper and generative models.

Bayesian Deep Learning. Since 2016, there have been several investigations in quantifying and representing uncertainty in CNNs. A novel approach was to approximate variational inference in a Bayesian implementation [7] for image classification [9]. Furthermore, the types of uncertainties (aleatoric or epistemic) and the ways to employ these uncertainties in both regression and classification tasks are presented in [13, 14]. The novel approximation of Bayesian [8] for deep learning by using the Dropout layers [27] reduced the computational complexities of the naive Bayesian Neural Network implementation. An analysis of uncertainty for the active learning framework is presented in [33]. However, the evaluation is conducted on small scale dataset such as MNIST [16] classification. In contrast to this, our work analyses a more difficult and large-scale experimental setup. Another recent work on image classification [4] proposed ensemble-based active learning and demonstrates outperforming the Dropout Bayesian approximation [8]. Again the experiments are constrained on small scale scenarios. On these premises, we further analyse the exploitation of both data and model-dependent uncertainties similar to [13], but for the large-scale and more challenging problem i.e. 3D-HPE. Furthermore, this will integrate the risk measurement concerns presented in [20] through the aleatoric uncertainty.

Active Learning Methods. This branch of machine learning was explored in the need of informative datasets that are in most cases model-dependent. A survey of different standard active learning schemes is in [24]. With the advances in deep neural networks, the research community began to integrate the classical approaches despite the lack of integration in online training (essential part of the active learner methodology). The most common scenario used, pool-based sampling, limits the live model refinement by performing offline training [22, 10, 34, 26]. On the other hand, pool-based active learning opened a new direction of research for deep neural networks [9]. This has evaluated with what percentage of the training set the model can achieve the same accuracy than the entire one. In terms of pose estimation, a practical application of active learning for hands has been conducted in [11], where KD-trees are applied to guide the camera movement to a more informative viewpoint. A theoretical approach has been explored in [11] over the human pose estimation problem by actively collecting unlabelled data from the heat-map output of the Convolutional Pose Machines (CPMs). However, our methodology is driven independently from the output as
it relies directly on the predicted hand skeleton.

3. Method

We start this section with formulating the 3D-HPE problem as a Bayesian approximation inspired from [12, 13] for semantic segmentation and depth regression. We adopt such a framework (similarly to Kendall et. al. in [13]) to model the two uncertainties: aleatoric and epistemic. These two statistics characterise the important, but complementary aspect of the model when trained from data. In particular, aleatoric uncertainty captures uncertainty due to noisy training examples which can not be eradicated from the model even if there is plenty of the training examples. Whilst, epistemic uncertainty quantifies the ignorant aspect of the model parameters which can be addressed with the availability of training examples. For an AL application, the aleatoric uncertainty plays a key informative role, to avoid annotating hard or noisy samples when acquiring new data. Besides, the epistemic variance helps the AL sampling method in indicating relevant unseen data for the learner. In the second part, given the parameters inferred from the Bayesian hand pose estimator, we further investigate an acquisition method, commonly known as a sampler in the active learning scheme, by enclosing the uncertainty variances. Hence, our contribution relies on the analysis of the Bayesian 3D-HPE approximation (comprises the learner component in AL), together with a newly proposed selection mechanism. Figure 2 summarises the proposed pipeline.

3.1. Bayesian 3D Hand Pose Estimator

3.1.1 3D Hand Pose Estimator

In classification tasks, uncertainty is estimated by the posterior probability of a class. However, the 3D hand pose estimation, a regression problem, maps hand (depth) images to 3D co-ordinates of the hand joint locations. We have a scenario where hand depth-image and representative co-ordinates describe its skeleton are given. We describe the hand skeleton following [6] in 21 joints as 1 root palm, 5 metacarpophalangeals (MCP), 5 proximal interphalangeals (PIP), 5 distal interphalangeals (DIP) and 5 fingertips. With this information, we shortly describe a method to regress the skeleton co-ordinates along with modelling the aforementioned uncertainties.

In order to topologically evaluate the regression uncertainties, we deploy a standard DeepPrior [19] architecture. This comprises of a convolutional feature extractor and a dense regression. Given $x \in \mathbb{R}^{w \times h}$, a 2D cropped hand image with $w$ width and $h$ height is inferred through the CNN, together with its corresponding ground-truth $y \in \mathbb{R}^{21 \times 3}$. Structurally, the feature extractor is composed of three block groups of convolution, max pool and LeakyReLU activation. The flatten output of the feature extractor is regressed in the end by 2 dense layers (see Figure 2). For simplicity, we exclude the final PCA layer from the initial design of [19] and directly predict the normalised UVD joint coordinates $\hat{y}$ (description of the pipeline in Fig. 2).

Finally, we optimise the parameters by Stochastic Gradient Descent (SGD) to minimise the mean squared error. The objective function is as below:

$$\mathcal{L}(x, y; \Theta) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{1}{K} \sum_{k=1}^{K} \|y_{i,k} - \hat{y}_{i,k}\|^2 \right),$$

where, $n$ is the data size, $\Theta$ represent model parameters and $K$ is the number of joints.

3.1.2 Bayesian DeepPrior

Here, we deep dive the Bayesian DeepPrior into details. As we stated before, unlike in classification model, it is not that straight forward to model the uncertainties in regression model. In order to make the DeepPrior a probabilistic model, we introduced Dropout [27] on its layers similar to that of [8]. As in [13], we also propose to estimate and investigate aleatoric and epistemic uncertainties for our task.

Bayesian Neural Networks (BNNs) sets a Normal distribution $\Theta \sim \mathcal{N}(0, I)$ of its weight parameters as prior. To approximate the posterior distribution over the weights $f(\theta|x, y)$, we minimise the Kullback-Leibler (KL) divergence of a variational inference distribution $q(\theta)$ and its posterior: KL$(q(\theta) || f(\theta|x, y))$. To minimise the KL divergence loss, we apply Monte Carlo Dropout (MCD) during the variational inference and we minimise the mean squared error of joint location’s prediction. This minimisation is equivalent to the KL divergence minimisation. For more details, we suggest readers to refer [13]. Once we keep Dropout active over the entire $\Theta$ we can obtain a Bayesian approximation of the posterior’s mean and variance.

As stated in [20], the uncertainty variance of the BNN consists of the Bayes risk rather than a systematic uncertainty. Therefore, we learn together with the outputs of the final dense layer the aleatoric uncertainty that balances the mean squared error objective function during training. Specifically, we allocate an aleatoric variance for each hand joint coordinate. Hence, the new objective function to model such uncertainties is defined as in [13]:

$$\mathcal{L}_B(x, y; \Theta) = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{1}{K} \sum_{k=1}^{K} e^{-\alpha_{i,k}} \|y_{i,k} - \hat{y}_{i,k}\|^2 + \frac{1}{2} \alpha_{i,k} \right),$$

with the logarithmic variance $\alpha_{i,k} = \log \hat{\sigma}_{al}^2$ and $\hat{\sigma}_{al}^2$, the aleatoric variance. Thus, the learnt numerically-stable logarithmic tracks the noise present in the data.
Figure 2. Schematic diagram showing the end-to-end pipeline of the proposed method. The depth image of the hand is pre-processed and fed to the Bayesian DeepPrior Network. This network is trained to minimise objective function given in Equation 2. The aleatoric and epistemic variances of a sample that we obtain from $M$ number of Monte Carlo Dropouts are passed on to the selection criteria for querying unlabelled examples. The selection method is highlighted in the Algorithm 1. The switches in the Figure are present to indicate the activation of Dropout in different Bayesian approximations.

To summarise, by applying Monte Carlo Dropout (MCD) we obtain a Bayesian DeepPrior that generates a mean value for each joint coordinates. After variational inferences we also evaluate its epistemic and learnt aleatoric variances. For $M$ times passes of a sample, the epistemic uncertainty is estimated as below:

$$\hat{\sigma}_{ep}^2 \approx \frac{1}{M} \sum_{m=1}^{M} \hat{y}_m^2 - \left( \frac{1}{M} \sum_{m=1}^{M} \hat{y}_m \right)^2$$  \hspace{1cm} (3)

Finally, the combined variances for a predicted skeleton $\hat{y}$ can be expressed as:

$$\hat{\sigma}(\hat{y}) \approx \frac{1}{M} \sum_{m=1}^{M} \hat{y}_m^2 - \left( \frac{1}{M} \sum_{m=1}^{M} \hat{y}_m \right)^2 + \frac{1}{M} \sum_{m=1}^{M} \sigma^2_{al_m}$$  \hspace{1cm} (4)

Further in our experiments section 4.2 we present the study on number of $M$ vs the mean joint error stabilisation.

From the architectural perspective, the Bayesian DeepPrior contains Dropout layers after every convolutional and dense layer. In practice [12] [13], it has been shown that this model may suffer from strong regularisation. Thus, similarly to [12], we propose different variants where Dropout is applied to designated locations. These probabilistic variants of Bayesian DeepPrior are simulated through switches in the proposed pipeline accordingly:

- **A** - only to all convolutional layers;
- **B** - centrally to the last convolutional layer and the first dense layer;
- **C** - throughout the entire DeepPrior architecture.

To fine-tune the design of the Bayesian DeepPrior architecture, we perform cross-validation study on the NYU Hand dataset [29] in section 4.2.

### 3.2. Active Learning Framework

In this section, we briefly describe the active learning process for deep learning together with the proposed acquisition function adapted for the Bayesian DeepPrior.

**Pool-based Active Learning Strategy.** As active learning has gained stronger interest in deep learning, a pool-based scenario has become a standard methodology to overcome the data-greedy models with their slow training process [1]. Therefore, the pool-based active learning considers a scenario with an initial annotated set $s^0$ and an available unlabelled dataset $U_{pool}$. The goal is to find the least amount of $L$ annotated subsets $s^1, s^2 \ldots s^L \subset U_{pool}$ so that we achieve the targeted mean squared joint error. Given an acquisition function $A$, this can be summarised under the following equation:

$$\min_{L} \min_{s^1, s^2 \ldots s^L \subset U_{pool}} A(L; s^1, s^2 \ldots s^L).$$  \hspace{1cm} (5)

In the next section, we analyse and propose a function $A$ suitable for our Bayesian DeepPrior architecture.

**Combination of Geometric and Uncertainty Acquisition Function.** Acquisition functions have been extensively developed for classification tasks [10] [21] due to its probabilistic output. However, in regressions, we either have to
Step 2. Find maximum distance between minimas.
Step 3. Select the corresponding unlabelled point.

CoreSet vs CKE:

Moreover, the estimated 3D coordinates are averaged of posterior distribution, we define the selection under the

Algorithm 1 CKE

1: Input: labelled set \( x_j \in s^0 \), unlabeled pool \( x_i \in U_{\text{pool}} \), query budget \( B \), corresponding epistemic variances \( \hat{\sigma}_{\text{ep}}, \sigma_{\text{ep}} \).
2: Initialise \( s \subseteq U_{\text{pool}} \).
3: repeat
4: \( \Delta d_{\text{ub}} = \min_{j \in s^0} \Delta(\hat{y}_i, \hat{y}_j) \)
5: \( \arg_{\text{ub}} = \arg \min_j \Delta(\hat{y}_i, \hat{y}_j) \)
6: \( b = \arg \max_i \Delta d_{\text{ub}}(\arg_{\text{ub}}) \)
7: \( s_{\text{ublb}} = s^0 \cup \{ b \} \)
8: until \( s_{\text{ublb}} = B + s^0 \)
9: Return: \( B_{\text{ublb}} = s_{\text{ublb}} \setminus s^0 \)

note that CKE benefits of the Bayesian model uncertainty \( \hat{\sigma}_{\text{ep}} \) when minimising the global geometric cover \( \gamma_s \). In this manner, our proposed solution identifies hand poses furthest from the labelled centres and with the highest epistemic uncertainty deviation. The 3D MCD outputs are also more robust to noise due to their learnt aleatoric uncertainties. On these premises, we consider that our active learning method adapted to the Bayesian DeepPrior is superior to the standard approach.

4. Experiments

4.1. 3D Hand Pose Estimation Datasets

ICVL [28]: This is one of the earliest depth-based hand datasets. It consists of a total number of 17,604 (16,008 train, 1596 test) from 10 subjects with 16 joints annotations.

NYU [29]: Created from 2 subjects, NYU has 72,757 training images and an 8,252 testing set. The hand skeleton consists of 36 3D coordinates.

BigHand2.2 [37]: The largest benchmark to date, it consists of 2.2M frames from 10 different hand-models. For practicality, we decide to uniformly sub-sample every 10 frames...
Bayesian DeepPrior variants | A | B | C
---|---|---|---
NYU Testing MSE [mm] | 22.48 | 22.94 | 22.65

Table 1. Ablation evaluation of Bayesian DeepPrior

due to the insignificant accuracy gain over the entire set as analysed in [37]. Hence, we train our model with 251,796 frames and test 21 hand key-points on 39,099 frames.

For all the datasets, we standardise the number of joints to 21 3D locations. We use pre-trained UNet [22] to detect hand and centre cropped it to the dimension of 128×128.

### 4.2. Bayesian DeepPrior Architecture

**Ablation Studies on Architecture.** To approximate a Bayesian Neural Network using MCD as described in Section 3.1.2, we proposed three probabilistic variants by dropping out different combinations of layers. We applied these three strategies: A - convolutional feature extractor; B - centrally, on the last convolution and first dense layer; C - throughout the entire DeepPrior and evaluated the performance on NYU data set. The comparison of the performance on these configurations is summarised in Table 1. Accordingly, when applying MCD only on the feature extractor (A), it yields the best performance. This is because in C Dropout brings too much regularisation, while B locks low-level features in the first convolutional layers. The trend we observe is similar to one reported in the Bayesian SegNet [12]. We deploy the Adam [15] optimiser with a learning rate of 10^{-3}, a batch size of 128 and a total number of $M = 70$ MCDs. These parameters remain constant for all three setups. Throughout all of our upcoming experiments, we maintain the Bayesian DeepPrior variant where Dropout is present after the convolutional layers.

After identifying the optimal configuration of Dropouts, we played with the number of variational inferences $M$ under the same hyper-parameters setting. We observed that increasing the value of $M$ does not impact the performance, however, adds computational complexity. Lowering the value of $M$ to 40 does not impact the performance. Hence, we keep this value for the upcoming experiments.

**Bayesian DeepPrior vs Standard DeepPrior** To demonstrate the effectiveness of our Bayesian DeepPrior approach, we conduct a quantitative comparison against the standard DeepPrior baseline proposed in [19]. We evaluate the average MSE for both train and test sets of the three compared benchmarks. The performance comparison is summarised in Table 2.

The Bayesian DeepPrior brings a clear advantage over Standard by yielding the lower testing error on all the three benchmarks. This shows how effectively we can generalise overall testing sets, while the standard DeepPrior over-fits the training sets. In addition to the performance improvement, the Bayesian DeepPrior also generates uncertainty metric for the hand poses.

| Hand Dataset | Bayesian DeepPrior | DeepPrior |
|--------------|--------------------|-----------|
| ICVL | train | 7.5261 | 7.1633 |
| | test | 10.0988 | 10.6233 |
| NYU | train | 7.7566 | 4.7034 |
| | test | 22.4838 | 21.4649 |
| BigHand2.2 | train | 7.7566 | 12.7731 |
| | test | 21.4649 | 22.65 |

Table 2. Bayesian DeepPrior vs DeepPrior - averaged MSE [mm]

| Hand Dataset | aleatoric var | epistemic var |
|--------------|---------------|---------------|
| ICVL | train | 1.057 | 0.004 |
| | test | 1.016 | 0.0039 |
| NYU | train | 1.169 | 0.0142 |
| | test | 1.512 | 0.0066 |
| BigHand2.2 | train | 1.964 | 0.0132 |
| | test | 2.1594 | 0.0174 |

Table 3. Averaged epistemic and aleatoric variances at $10^{-2}$ on ICVL, NYU and BigHand2.2

**Epistemic and Aleatoric Uncertainties** For every coordinate of the hand skeleton, our model estimates their aleatoric and epistemic deviation. Table 3 enlists the average of both uncertainties for all the three benchmarks. The values are in the range of 0 to 1. Please note, we predict the normalised UVD coordinates of the hand joints.

As deducted in Kendall et al. [13], we consistently obtain higher aleatoric variances than the epistemic ones. Hence, the noise present in the hand data tends to overcome the model’s learning capability. Apart from BigHand2.2, the testing epistemic variance seems lower than on the training set. Furthermore, we notice an increased aleatoric variance on BigHand2.2 compared to the other sets. High noise during annotation could be the reason for it.

### 4.3. Active Learning Evaluation

Active learning has shown to be an effective tool [23] in acquiring representative data for a learning model. Given the proposed Bayesian DeepPrior architecture, we presented in Section 3.2 the CKE query method. While our pipeline gets an advantage from reducing the noisy samples through learnt aleatoric variances, we gather both geometric and epistemic information regarding new unlabelled poses. We shortly describe the active learning selection baseline used for both Bayesian and standard DeepPrior architecture:

**Random sampling:** The typical approach of annotating data just by uniformly sampling the unlabelled pool $U_{pool}$.

**Uncertainty sampling:** Although for 3D HPE regression there is no confidence measurement like in the classification tasks, we apply this method only on the evaluated epistemic uncertainties. The unlabelled examples are inferred through the Bayesian DeepPrior and their epistemic variances result after 40 MCDs. For each predicted skeleton, we sum all their corresponding epistemic deviations and we select to
annotate the topmost uncertain hand poses.

**CoreSet:** It is one of the current state-of-the-art geometric acquisition function. Also, it has been widely used due to its task-agnostic properties relying either on the model’s extracted features or on the output space. In our evaluation, we apply CoreSet directly on the predicted skeletons of both labelled and unlabelled sets.

**CKE:** This is our proposed method that extends the CoreSet functionality by including the epistemic deviations in the risk minimisation between the loss of the newly selected samples and the loss of the available labelled set.

### 4.3.1 Quantitative Results of Selecting Methods

In this part, we evaluate on three hand pose datasets: ICVL, NYU and BigHand2.2, under the pool-based scenario. We perform a quantitative comparison between the random sampling and CoreSet on the standard DeepPrior learner followed by CKE, our proposed selection method, on the Bayesian 3D HPE. We follow the standard pool-based protocol, where there is a large pool, $U_{pool}$, of unlabelled data. A small size initial set seed annotations $s^0$ are made available for the first offline training of the target model. After this, we follow the same practice as in \[4\] and randomly create a smaller pool $s \subset U_{pool}$. Thus, we efficiently deploy the selection methods on $s$ under the budget $B$. We repeat this systematically over 10 stages. As a quantitative metric, we quantify the performance in averaged MSE. Also, due to the variation in the initial selected set $s^0$ and new random subsets $s$, we average our results over 5 trials and while also computing the standard deviation.

For NYU and BigHand2.2, we set a budget ($B$) as well as seed annotations $s^0$ equal to 1000 samples. Whereas, for ICVL, we set it 100 due to its smaller size. The intermediate subset $s$ size used for active learning selection is set to 10% of the entire $U_{pool}$ which results in 20,000 of BigHand2.2; 7,276 of NYU; and 1,601 of ICVL. In our CKE query function, we identified the most informative samples when the uncertainty influence parameter was set to $\eta = 0.3$.

Figure 4 compares the performance of the proposed method with the default baseline (random sampling) and one of the state-of-the-art methods (CoreSet) in three different challenging data sets. Referring to the same Figure, we can clearly see that the proposed CKE consistently outperforms the existing state-of-the-art. Specifically, we achieve after 10 AL passes the lowest averaged MSE on every benchmark accordingly: 12.17 mm for ICVL, 23.48 mm for NYU and 25.21 mm for BigHand2.2. This demonstrates how effective is the Bayesian AL in obtaining top accuracy with fractions of the entire datasets.

Similarly, Figure 5 illustrates performance comparison between the different existing selection method with our proposed sampling technique CKE when applied with the proposed MCD learner i.e. Bayesian DeepPrior. Comparing the performance of CoreSet when the learner is DeepPrior (Figure 4) vs when the learner is Bayesian DeepPrior (Figure 5), we observe the improvement of the averaged MSE on every dataset. This trend is followed even in random sampling technique. Hence, this demonstrates that modeling aleatoric and epistemic uncertainties while training learner in the AL framework is effective.
4.3.2 Qualitative comparison

We also evaluated the sampling methods qualitatively. We compared the baselines (random and CoreSet) against our proposed CKE function. To this end, we extract the predicted 3D hand skeletons on NYU test set and track their structural representation in the first two initial stages as well as in the last two AL stages. In the left part of Figure 6, we can observe that CKE under the Bayesian DeepPrior learner generates 3D hand poses quite closer to the ground truth from the early stages. Whereas, the other two methods failed to do so. These characteristics are equally visible even in highly articulated poses as shown in the last row.

From a qualitative perspective, we also evaluate the relevance of the epistemic and aleatoric uncertainty values in the context of 3D joint error locations. The last two columns in the Figure 6 depict these uncertainties. The extended variance present at the fingertips (proportional to the radius of the circle) can be interpreted as high acquisition noise. Moreover, we can observe that for occluded poses (middle and last row), the bigger circle indicates that the model is less confident in such cases. This happens when there are not sufficient training examples in such extreme poses and occlusions. To overcome this, a more powerful estimator trained on annotated examples with such extreme poses and occlusion needs to be deployed.

5. Conclusions

We have elaborated the first work of AL applied to the 3D hand pose estimation task. We successfully approximated the 3D-HPE DeepPrior as a BNN while deriving its model and data-dependent uncertainties. We have shown through qualitative and quantitative evaluation how the two uncertainties play a critical role in the AL scheme. Furthermore, by combining the geometric sampling from CoreSet, we proposed a sampling technique, CKE, suitable for Bayesian DeepPrior infrastructure. Under the pool-based scenario, we achieved the lowest 3D joint errors with the least amount of data for three well-know datasets. To conclude, this work demonstrates the lack of representativeness and redundancy that can be present when gathering a 3D hand dataset. Therefore, a Bayesian approximation together with the CKE acquisition method may help in building a holistic and model-refined dataset while saving a considerable annotation time.

Acknowledgement

This work is partially supported by Huawei Technologies Co. and by EPSRC Programme Grant FACER2VM (EP/N007743/1).
References

[1] Active Learning for Human Pose Estimation. In ICCV, 2017.

[2] Anil Armagan, Guillermo Garcia-Hernando, Seungryul Baek, Shreyas Hampali, Mahdi Rad, Zhaohui Zhang, Shipeng Xie, MingXiu Chen, Boshen Zhang, Fu Xiong, Yang Xiao, Zhiguo Cao, Junsong Yuan, Pengfei Ren, Weitong Huang, Haifeng Sun, Marek Hrtz, Jakub Kanis, Zdenek Krhoul, Qingfu Wan, Shile Li, Linlin Yang, Donghui Lee, Angela Yao, Weigu Zhou, Sijia Mei, Yunhui Liu, Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Philippe Weinzaepfel, Romain Brégier, Grégory Roze, Vincent Lepetit, and Tae-Kyun Kim. Measuring generalisation to unseen viewpoints, articulations, shapes and objects for 3d hand pose estimation under hand-object interaction. In ECCV, 2020.

[3] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Augmented skeleton space transfer for depth-based hand pose estimation. In CVPR, June 2018.

[4] William H Beluch Bcai, Andreas Nürenberger, and Jan M Köhler Bcai. The power of ensembles for active learning in image classification. In CVPR, 2018.

[5] R. Qi Charles, Hao Su, Mo Kaichun, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In CVPR, IEEE, 2017.

[6] Ali Erol, George Bebis, Mirczea Nicolescu, Richard D. Boyle, and Xander Twombly. Vision-based hand pose estimation: A review, 2007.

[7] Yarin Gal and Zoubin Ghahramani. Bayesian convolutional neural networks with bernoulli approximate variational inference, 2015.

[8] Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In ICML, 2016.

[9] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian Active Learning with Image Data. In ICML, 2017.

[10] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian Active Learning with Image Data. In ICML, 2017.

[11] Dun-Yu Hsiao, Min Sun, Christy Ballweber, Seth Cooper, and Zoran Popović. Proactive sensing for improving hand pose estimation. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 2016.

[12] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding.

[13] Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In NeurIPS, 2017.

[14] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In CVPR, 2018.

[15] Diederik P Kingma and Jimmy Lei Ba. ADAM: A Method for Stochastic Optimization. In ICLR, 2015.

[16] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010.

[17] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In CVPR, June 2018.

[18] Markus Oberweger and Vincent Lepetit. Deepoprior++: Improving fast and accurate 3d hand pose estimation. In ICCV, 2017.

[19] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. Hands deep in deep learning for hand pose estimation. In CVWW, 2015.

[20] Ian Osband. Risk versus uncertainty in deep learning : Bayes , bootstrap and the dangers of dropout. 2016.

[21] Robert Pinsler, Jonathan Gordon, Eric Nalisnick, and José Miguel Hernandez-Lobato. Bayesian Batch Active Learning as Sparse Subset Approximation. In NeurIPS, 2019.

[22] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, 2015.

[23] Ozan Sener and Silvio Savarese. Active Learning for Convolutional Neural Networks: A Core-set approach. In ICLR, 2018.

[24] Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[25] yawar Siddiqui, Julien Valentin, and Matthias Nießner. Viewal: Active learning with viewpoint entropy for semantic segmentation. In CVPR, 2020.

[26] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational Adversarial Active Learning. In ICCV, 2019.

[27] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. J. Mach. Learn. Res., 2014.

[28] Danhang Tang, Hyung Jin Chang, Alykhan Tejani, and Tae-Kyun Kim. Latent regression forest: Structured estimation of 3d articulated hand posture. In CVPR, 2014.

[29] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. Real-time continuous pose recovery of human hands using convolutional networks. ACM Transactions on Graphics, 33, August 2014.

[30] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. Dense 3d regression for hand pose estimation. In CVPR, June 2018.

[31] Gert Wolf. Facility location: concepts, models, algorithms and case studies. In Contributions to Management Science, pages 331–333, 2011.

[32] Fu Xiong, Boshen Zhang, Yang Xiao, Zhiguo Cao, Taidong Yu, Joey Tianyi Zhou, and Junsong Yuan. A2j: Anchor-to-joint regression network for 3d articulated pose estimation from a single depth image. In ICCV, 2019.

[33] Qi Ye, Shanxin Yuan, and Tae-kyun Kim. Spatial Attention Deep Net with Partial PSO for Hierarchical Hybrid Hand Pose Estimation. In ECCV, 2016.

[34] Donggeun Yoo and In So Kweon. Learning Loss for Active Learning. In CVPR, 2019.

[35] Shanxin Yuan, Guillermo Garcia-Hernando, Björn Stenger, Gyeongsik Moon, Ju Yong Chang, Kyoung Mu Lee, Pavlo Molchanov, Jan Kautz, Sina Honari, Liuhao Ge, Junsong Yuan, Xinghao Chen, Guijin Wang, Fan Yang, Kai Akiyama, Yang Wu, Qingfu Wan, Meysam Madadi, Sergio Escalera,
Shile Li, Dongheui Lee, Iason Oikonomidis, Antonis Argyros, and Tae-Kyun Kim. Depth-based 3d hand pose estimation: From current achievements to future goals. In CVPR, 2018.

[36] Shanxin Yuan, Qi Ye, Guillermo Garcia-Hernando, and Tae-Kyun Kim. The 2017 hands in the million challenge on 3d hand pose estimation, 2017.

[37] Shanxin Yuan, Qi Ye, Bjorn Stenger, Siddhant Jain, and Tae-Kyun Kim. Bighand2.2m benchmark: Hand pose dataset and state of the art analysis. In CVPR, pages 2605–2613, 07 2017.

[38] Jiaming Zeng, Adam Lesnikowski, and Jose M. Alvarez. The relevance of bayesian layer positioning to model uncertainty in deep bayesian active learning, 2018.