Tanzania Journal of Engineering and Technology 2022, **41(4)**:83-94
OPEN ACCESS articles distributed under Creative Commons Attribution Licence [CC BY-ND]
Websites: https://www.ajol.info/index.php/tjet; https://tjet.udsm.ac.tz

ISSN 1821-536X (**print**); ISSN 2619-8789 (**electronic**)
DOI: https://doi.org/10.52339/tjet.v41i4.762

Full Length Research Paper

Occlusion Handler Density Networks for 3D Multimodal Joint Location of Hand Pose Hypothesis

Stanley L. Tito[1†], Jamal F. Banzi [2] and Aloys N. Mvuma[1]

[1]Mbeya University of Science and Technology, P. O. Box 131, Mbeya, TANZANIA
[2]Sokoine University of Agriculture, P. O. Box 3167, Morogoro, TANZANIA
†Corresponding author: stanley.tito@must.ac.tz; ORCID: https://orcid.org/0000-0003-4094-6530

ABSTRACT

Predicting the pose parameters during the hand pose estimation (HPE) process is an ill-posed challenge. This is due to severe self-occluded joints of the hand. The existing approaches for predicting pose parameters of the hand, utilize a single-value mapping of an input image to generate final pose output. This way makes it difficult to handle occlusion especially when it comes from the multimodal pose hypothesis. This paper introduces an effective method of handling multimodal joint occlusion using the negative log-likelihood of a multimodal mixture-of-Gaussians through a hybrid hierarchical mixture density network (HHMDN). The proposed approach generates multiple feasible hypotheses of 3D poses with visibility, unimodal and multimodal distribution units to locate joint visibility. The visible features are extracted and fed into the Convolutional Neural Networks (CNN) layer of the HHMDN for feature learning. Finally, the effectiveness of the proposed method is proved on ICVL, NYU, and BigHand public hand pose datasets. The imperative results show that the proposed method in this paper is effective as it achieves a visibility error of 30.3mm, which is less error compared to many state-of-the-art approaches that use different distributions of visible and occluded joints.

ARTICLE INFO

Submitted: **July 5, 2022**

Revised: **August 9, 2022**

Accepted: **October 18, 2022**

Published: **December 30, 2022**

Keywords: *Deep learning, Convolutional neural networks, Self-occluded joints, Unimodal gaussian distribution, Multiple feasible hypotheses*

INTRODUCTION

Modern computer interactions such as virtual reality, augmented reality, somatosensory games and gesture interaction require a non-invasive interface that achieves a greater user experience (Ge et al., 2018). Hand pose estimation is a vital component of these seamless human-machine interactions. Hand pose estimation has the potential to provide a natural and non-contact solution.

There has been a considerable research effort in this area in the last two decades (Duan et al., 2019). Many of these researches have been developed from data-driven 2D interfaces to 3D joint positions and have achieved accurate results (Farahanipad et al.,2021). However, because of the complexity of hand and self-occlusion among other factors, hand pose estimation is still a challenging task.

In recent years, deep learning has gradually become the mainstream approach for solving problems in computer vision and pattern recognition (Ge et at., 2018; Duan et al.,2019; Banzi et al., 2020).

Previous research results show that CNN-based methods perform better when estimating hand pose. The existing methods for 2D position estimation of hand joints have attained high accuracy(Farahanipad et al.,2021). However, since the depth information in the original image cannot be fully utilized due to occlusion, HPE results are not ideal. Recently, many meaningful models have been proposed to address this issue (Zhang et al., 2021; Xiong et al., 2019). However, self-occlusion still hides some potential information and eventually misleads the accuracy of the pose estimator (Xiongwei and Hoi, 2020). This paper therefore, attempts to address the problem of handling multimodalities of occluded finger joint locations by introducing a hybrid hierarchical mixture density network (HHMDN). The proposed HHMDN provides an end-to-end learning ability through differentiable density functions with a complete description of the hand poses of the given images under occlusions.

The idea behind HHMDN is to model the probability distribution of the joint locations of the fingers in a two-level hierarchy for both single and multivalued mapping under conditioned joint visibility and detection score. The first level hierarchy represents the distribution of a latent variable for the joint visibility. The second level hierarchy represents the distribution of the finger joint locations using a single Gaussian model for visible joints or a Gaussian mixture model (GMM) for occluded joints through detection score. The entire network is trained end-to-end through differentiable density functions.

In the first place, this paper presents a probabilistic framework for detecting hand fingers based on detection scores to determine the visibility correlation of the visible finger. Then, the Gaussian labels were extracted from the detection scores to form multiple labels of a GMM with *i* components. These components are hierarchically regressed in visibility distribution, uni-modal distribution, and multimodal units to determine the joint visibility through CNN and facilitate feature learning. It also models several states in two hierarchical levels to unify a single and multiple-valued mapping in its output. The novelty of the proposed approach over other existing methods includes the integration of the detection score in line with the CNN to estimate visibility probability which is finally used to predict the probability that an input window is the hand pose. Also, using the negative log-likelihood of a multimodal mixture-of-Gaussians through HHMDN enhances the accuracy of HPE.

The remains of this paper are organized as follows: Section 2 describes the related works, and section 3 presents the system description and theoretical concepts. Section 4 presents the datasets and model analysis, while the experimental setup is explained in section 5. Finally, the results and discussion are explained in section 6.

MATERIALS AND METHODS

Related work

HPE approaches are generally classified into generative, discriminative, and hybrid paradigms(Roy et al.,2017). Discriminative and generative are the two complementary approaches, while the hybrid approach combines both generative and discriminative approaches to complement the significant advantages of both methods (Tang et al. 2018). Many of the previous works for HPE were developed based on the generative approach in which multiple hand models called hypotheses are generated and a suitable model that best matches the observed data were found (Ling et al., 2018; Pavllo et al., 2019; Xiongwei and Hoi, 2020). In the generative approach, the predicted pose parameters usually originate from the prior stage. An optimization framework is used to find a

model that minimizes a certain cost function (Banzi et al., 2019).

A self-learning procedure based on Deep Reinforcement Learning (DRL) with a bounding box was proposed by Saha (2018), to localize gesture location. However, the approach did not perform well under self-occlusion, requiring enough memory due to the number of iterations needed during training.

Generative approaches face numerous challenges, including a tremendous overhead in rendering candidate poses(Banzi et al., 2019). Generally, generative methods are computationally expensive and sometimes need to be implemented on GPU to meet a real-time performance (Banzi et al., 2020). Most recent works are based on a discriminative approach, which requires the hand pose to be extracted from a single-depth image through the mapping process using classification and regression techniques (Jariwala and Parmar, 2017). In particular, the tree-structured Region Ensemble Network (REN) was proposed by Guo et al., (2018) to recover a 3D hand pose by dividing the convolution outputs into regions while integrating the results from multiple regressors on each region.

A hierarchical tree-structured CNN discriminative approach was proposed by Madadi et al., (2017) to predict different parts of the kinematic tree and obtain the local poses as a subset of hand joints. Yuan et al., (2018) employ a CNN discriminative approach to minimize a mean square error function when mapping the 3D joints of the hand from a given set of hand input images and their corresponding pose labels. However, the mapping was considered a single-valued with a precise conditional average if all the finger joints from the given image are visible. This conditional average provides a limited description of the joint locations with reduced accuracy. Since occlusion frequently occurs in the egocentric view, the mapping must be multivalued because occluded joints exhibit multiple locations under the same images. Oberweger and Wohlhart (2015) use CNN to predict the joint locations and extract potential features to generate small heat maps for joint locations. The joints were then converted to hand skeletons using an inverse kinematics process. However, only 2D locations of joints were predicted, and the approach was computationally expensive and unable to predict the occluded joints. Generally, discriminative approaches have proven to have some difficulties with occluded joints, or high inaccuracies at fingertips (Oberweger et al., 2018).

The hybrid method was developed by (Chen et al., 2020) to leverage the merits of generative and discriminative learning techniques. The approach first provides the poses of the candidate through discriminative methods and uses them as an initial state of the generative technique to optimize the full hand poses. The early hybrid approach presented by Oberweger and Wohlhart (2015) trains a CNN of real labelled data annotated using a slower generative approach to regress the body pose. The works of (Zhou et al., 2019; Yuan et al., 2018) use hybrid techniques in different scenarios to provide a smooth and robust HPE. Still, their approaches were weak in generalizing to a new unseen object and had less pose estimation accuracy. Tang et al., (2017) demonstrated an adaptive hierarchical classification method to improve the efficiency of forest regression by regressing all the joints in one forest channel per frame. However, the method was prone to error propagation due to self-occlusion leading to wrong poses estimate.

Theoretical descriptions of the proposed system

The hand pose was first described as an articulated object before considering the probabilistic of the two-level hierarch to handle the hand poses under joint occlusion. CNN is then used to learn model parameters discriminatively. The detection scores vector given as $\lambda=[\lambda_1,\lambda_2,\ldots\lambda_k]^T$ are obtained first before estimating the visibility probability *p*

as $p(h|\lambda)$. The visibility probability is finally used to predict the probability that an input window is the hand pose. The present detection window is represented by ω while the detection score λ of the k parts is denoted by $\lambda=[\lambda_1, \lambda_2, \ldots, \lambda_k]^T$. Accordingly, it is assumed that both the deformation scores and appearance scores have been integrated by the part-based model into λ.

A probabilistic framework is designed in this study to effectively estimate the visibility of fingers before learning their visibility relationship based on a deep discriminative model. This considers the finger visibility of the k parts represented by $h=[h_1,h_2,\ldots,h_k]^T \epsilon \{0,1\}^T$ with $h_i=1$ meaning visible and $h_i=0$ meaning invisible. The study considered $\boldsymbol{h}$ as a hidden random vector because it was not given during training or testing phases such that

$$\sum_{i=1}^{k} p(h_i, y|\lambda_i) = p(h_i, y|\lambda_i)\sum_{i=1}^{k} p(h_i|\lambda_i) \qquad (1)$$

A special case framework is considered for pedestrian detection approaches by setting $h_i=1$ or taking the visibility h_i based on λ_i while ignoring the visibility power of the framework. Other approaches construct a deep model based on this power to learn the visibility relationship between finger parts. In this study, the finger visibility relationship $p(h_i|h|h_i,\lambda)$ is integrated with the joint visibility, which is given by the hierarchical mixture density network (HMDN) proposed by (Ye and Kim 2018) to form what is presented in this study as a hybrid hierarchical mixture density network (HHMDN).

The study exploits three publicly available datasets ICVL, NYU, and BigHand which lack visibility information on the finger joints, as shown in Table 1. This helped to investigate those with a higher proportion of occluded joints in the experiment. The learning datasets in our proposed study contain $\{X_n, Y_n, h_n^d | n=1, \ldots, N; d = 1, \ldots, D\}$ where $\{X_n, Y_n,$ and $h_n^d\}$ represents n^{th} hand depth image, the pose labels (represents 3D locations of the d^{th} joint of the n^{th} image) and the visibility variable, respectively. The d^{th} joint is connected with multiple labels $Y_n^d = y_m^d$ where $y_m^d \epsilon R^3$ is the m^{th} label i.e., 3D location. The binary variable was given to the visibility label to signify whether the d^{th} joint of the n^{th} image was visible or not. The D joints are handled independently. The discriminative model was divided into two levels to model occluded hand poses. The top-level which model the visibility constraint and the bottom level send-witched between a uni-model distribution and multimodal distribution based on the joint visibility. The visibility variable h_n^d provides the binary variable, which follows the Bernoulli distribution given as

$$p(h_n^d|v_n^d) = (v_n^d)^{h_n^d}(1-v_n^d)^{(1-h_n^d)} \qquad (2)$$

The term v_n^d is the probability that the joint is visible. Detection scores are used to generate visibility labels from the available pose labels. However, detection scores can only infer the detection of fingers. Nevertheless, the visible finger may be embedded with some occluded joints. Therefore, a sphere model is applied in this study, similar to Qian et al. (2014), to generate visible labels based on the existing pose labels**.**

The finger joints whose spheres have several pixels below a threshold are considered occluded. As described earlier, the visibility is given as a binary value, i.e. $h_n^d = 1$ when the joint is visible in the image and the location can be determined. The noise label y_m^d is generated only from a single Gaussian distribution

$$p(y_m^d|h_n^d = 1) = N(y_m^d \mu_m^d \alpha_m^d) \qquad (3)$$

For the occluded joint, i.e., $h_n^d = 0$, multiple labels are drawn from a GMM with i components such that

$$p(y_m^d|h_n^d = 0) = \sum_{i=1}^{i} \pi_i^N N(y_m^d c_m^d s_m^d) \qquad (4)$$

where c_{ni}^d is the center and s_{ni}^d is the standard deviation of the i^{th} components. Therefore, considering all components in place, the distribution of the joint locations conditioned on the visibility is given as

$$p(y_m^d|h_n^d)=\left[N(y_m^d\mu_n^d\sigma_n^d)^{h_n^d}\left[\sum_{i=1}^{i}\pi_i^d N(y_m^d c_{ni}^d s_{ni}^d)\right]\right]^{-(1-h_n^d)} \quad (5)$$

The distribution of joint through y_m^d and h_n^d is given as

$$p(y_m^d,h_n^d)=\left[v_n^d N(y_m^d\mu_n^d\sigma_n^d)\right]^{v_n^d}\left[(1-v_n^d)\sum_{i=1}^{i}\pi_i^d N(y_m^d c_{ni}^d s_{ni}^d)\right]^{-(1-h_n^d)} \quad (6)$$

This equation is the joint distribution that defines the loss function. Both the multilevel mixture density generation and the joint distribution equations are conditioned by x_n. All model parameters are functional, x_n and the joint distribution equation is differentiable. The CNN then learns the loss function and parameterizes the distribution through the output, as shown in Figure 1. The image x_n is an input to the CNN with the detected scores and the output is HHMDN parameters v_n^d, μ_n^d, σ_n^d, c_n^d, s_{ni}^d, π_{ni}^d for d = 1, …, D and i =1, …, i. The parameters given out have been classified into visibility probability based on equation 2, the uni-modal Gaussian based on equation 3, and the GMM based on equation 4. We utilize different activation functions to confine with the defined range of parameters. For example, exponential functions activate standard deviations and ensure they remain positive, while the SoftMax function ensures that π_{ni}^d it is maintained to [0,1]. The value h_n^d was used to calculate the visibility loss through the visibility label s_{ni}^d, v_n^d. Furthermore, the UG was calculated for the visible joints or GMM for occluded joints based on the visibility label presented above.

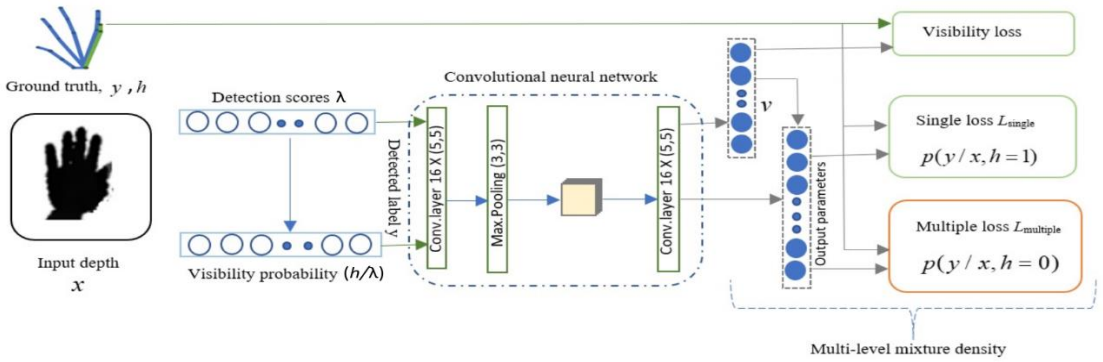

Figure 1: Probabilistic framework showing detection scores on the left and a CNN based learning frame on the right integrated into one pipeline.

Model Training

The likelihood for the complete dataset was first defined as $p=\prod_{n=1}^{N}\prod_{d=1}^{D}\prod_{m} p(y_m^d h_n^d)$. Then the CNN generates the parameters to maximize the likelihood of the dataset. The negative logarithmic likelihood was then applied to equation 6 as the loss function to obtain

$$L=-Log\ p=\sum_{n=1}^{N}\sum_{d=1}^{D}\sum_{m}(L_{visib}+L_{single}+L_{multi}) \quad (7)$$

In Figure 1 above, the L_{visib}, L_{single}, and L_{multi} corresponds to the three branches with their negative log described as

$$L_{visib}=-h_n^d Log(v_n^d)-(1-h_n^d)Log(1-v_n^d) \quad (8)$$

$$L_{single}=-h_n^d Log(N(y_m^d\mu_n^d\sigma_n^d)) \quad (9)$$

$$L_{multi}=-(1-h_n^d)Log(\sum_{i=1}^{i}\pi_{ni}^d N(y_m^d C_{ni}^d s_{ni}^d)) \quad (10)$$

The visibility loss L_{visib} can be calculated by using the estimated value h_n^d. The L_{single} is computed when h_n^d =0 and the L_{multi} is computed when h_n^d=1. At the testing stage, if the image x_n is applied into the network, the joint position estimation is classified into distinct branches through the visibility probability v_n^d. If v_n^d is larger than a specified threshold, the estimation of the location is provided by the uni-modal Gaussian distribution or the GMM. Nevertheless, if the estimation for the visibility is accompanied by errors, then the position of the joint will also be incorrect.

The samples drawn from the estimated distribution were used during the training stage to avoid bias instead of using binary visibility labels h_n^d to calculate the likelihood. When the number of samples is adequately greater, the mean of the samples becomes v_n^d, and the losses equation changes to

$$L_{single} = -v_n^d Log(N(y_m^d \mu_n^d \sigma_n^d)) \quad (11)$$

$$L_{multi} = -(1-v_n^d) Log(\sum_{i=1}^{i} \pi_{ni}^d N(y_m^d C_{ni}^d s_{ni}^d)) \quad (12)$$

Experiments

The experiments conducted to validate the proposed occlusion aware system for the 3D multimodal joint location of hand pose hypothesis are shown in this section. The paper briefly presents the experimental setup and introduces the dataset used. Finally, the results of our experiment were compared with several state-of-art approaches to evaluate the performance of our system.

Experimental setup

In the experiment, the detection scores λ were first obtained before using it to estimate the visibility probability of the positions of the finger to help suggest the occluded part. Multilevel mixture density was then integrated with the CNN layer to help determine the joint visibility and their corresponding visibility losses. The proposed Occlusion Aware Density Networks for 3D Multimodal Joint Location of Hand Pose Hypothesis was then implemented using C++ OpenCV. The number of framerates from different contending approaches is presented are Table1.

Table 1: The total number of frames and rate of the occluded finger joints

SN	Dataset	Train(R/T)	Test(R/T)
1	ICVL(Wang et al. 2018)	0.08/16,000	0.01/1596
2	NYU(Tang et al. 2015)	0.09/72757	0.36/8252
3	BigHand(Asad and Slabaugh 2016)	0.48/969,600	0.24/33,468

RESULTS AND DISCUSSION

Several models were compared with the Single Gaussian Network (SGN) baseline to explore the efficiency of our proposed approach. The SGN is the neural network trained with uni-modal Gaussian dissemination. The tested models are the SGN, HMDN, and HHMDN. The samples were drawn from the dissemination of various approaches. The results are presented in Figure 2 to depict the capacity of the proposed method in modelling mapping differences i.e., one-valued mapping for visible and many-valued mapping for occluded joints. For occluded joints, the samples generated by SGN form a broad range of spheres. The samples from HMDN and HHMDN are distributed as arc-shaped, indicating the range of movement of the fingertips under kinematic constraints. However, the SGN and HMDN yield the samples situated in a dense form around the ground truth position for visible joints.

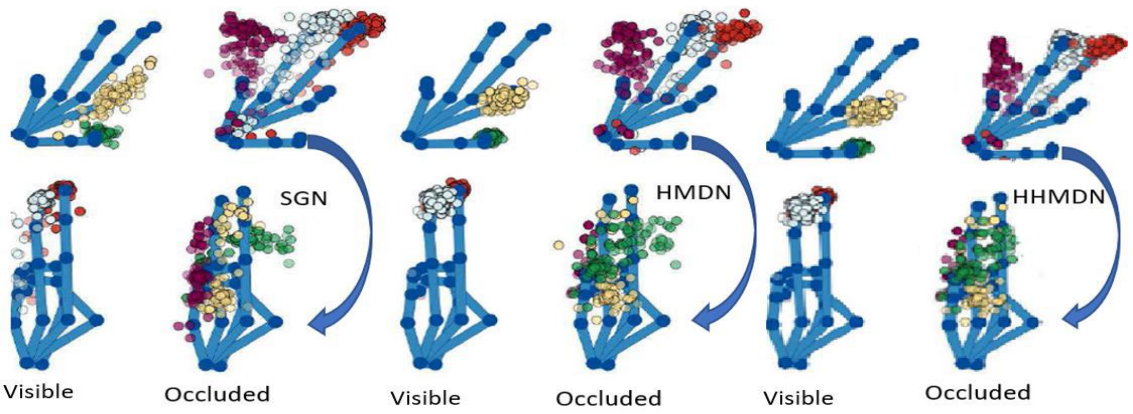

Figure 2: Illustration of samples drawn from different distributions for fingertips in visible and occluded joints.

The study considers the distribution of each method qualitatively. It compares each hypothesis with the ground truth joint locations to measure the displacement errors as reported in Table 2 for visible and occluded joints. The comparisons under normally used metrics, i.e. the proportion of joints within an error threshold were then depicted when the number of joints which is the Gaussian component, was set to J=20. The results are presented in Figure 3. The proposed HHMDN outperformed both SGN and HMDN using different numbers of the Gaussian component. The average errors for both visible and occluded joints are reported in Table 2. The estimation errors of HHMDN and HMDN, as shown in Table 2, do not change much for J=10, 20 and 30. Nevertheless, the parameters of the model were increased linearly with J.

Table 2: Prediction error of different models with different distributions

SN	No. of Gauss(J)	Model	Visibility error(mm)	Occluded error(mm)
1	10	SGN	33.4	37.5
2	20	HMDN	30.7	34.7
3	30	HHMDN	30.3	34.2

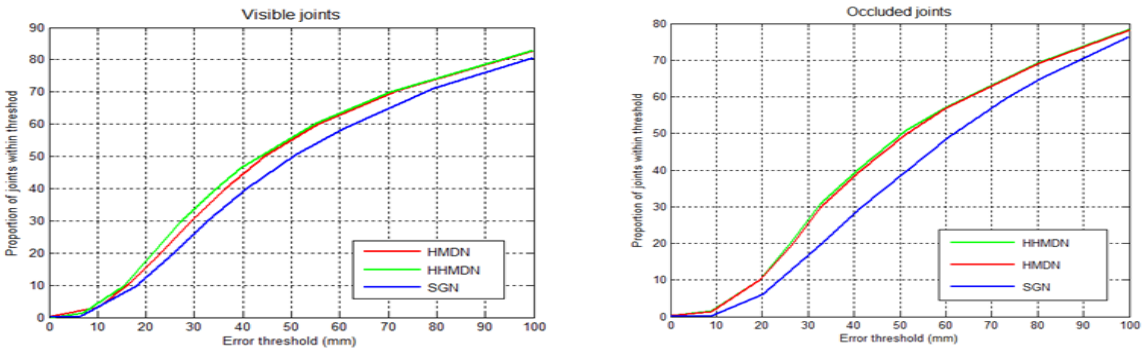

Figure 3: Self-comparison of the three distributions showing the proportion of errors in a given threshold for both visible and occluded joints.

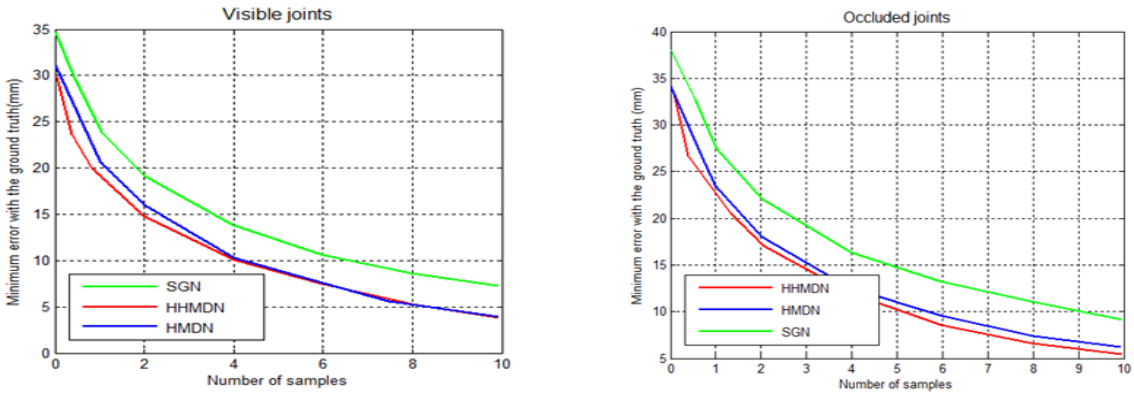

Figure 4: Self-comparison showing the minimum error attained against the ground truth for both visible and occluded joints.

As depicted in Figure 3, HHMDN performs better than SGN and slightly better than HMDN for visible and occluded joints using various components in Gaussian. When an error threshold is 40mm, HHMDN outperformed by 8% for visible joints while improving the SGN by more than 10% when the error threshold is 40mm.

Furthermore, the samples attained from the distributions, as depicted in Figure 4, are varied and measured with the least distance error. Our HHMDN achieved the lowest errors than SGN and HMDN in all samples. However, as the samples increase, the error gap between the two approaches is decreased.

The performance of our proposed method was evaluated on two publicly available datasets, NYU and MSHD, on the recently published works of (Chen et al., 2019; Ye and Kim 2018), with a considerable number of occluded joints. Although different evaluation metrics have been used in the literature, we focus on the segment of sample error distance in a metric of threshold to measure the fraction of success frames with the error distance of each joint less than a certain threshold. A single mistaken joint in this evaluated metric may depreciate the whole hand, making this the most challenging evaluation criterion.

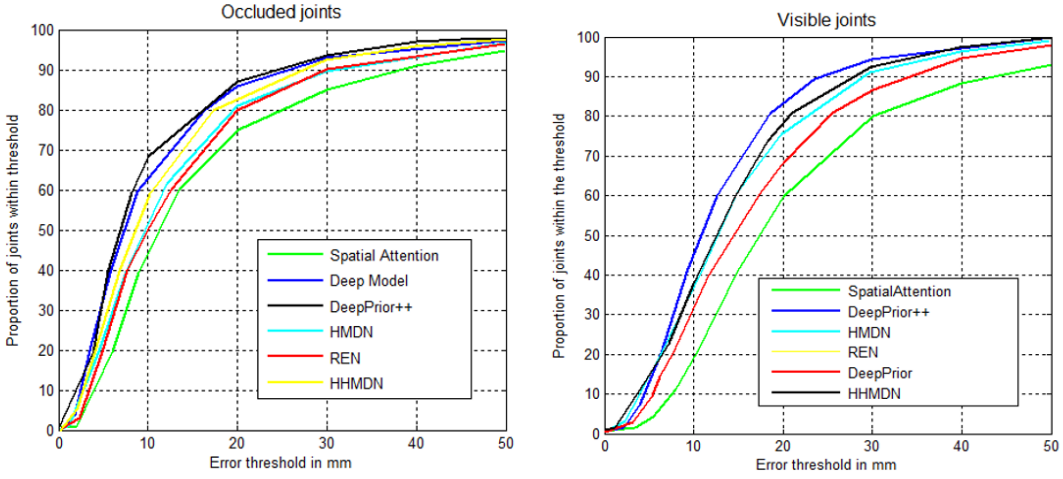

Figure 5: Evaluation with the state-of-the-art approaches on the NYU dataset for visible and occluded joints.

The study was compared with five state-of-art approaches on the NYU dataset based on the fraction of success frame whose distance between all predicted joints is within a certain threshold. These approaches are Spatial attention Ye et al. (2016), DeepModel Oikonomidis et al. (2011),

DeepPrior Oberweger and Wohlhart (2015), DeepPrior++ Oberweger and Lepetit (2017), REN Guo et al. (2018) and HMDN (Ye and Kim,2018). Most of the joints in this dataset are visible, especially in the training set, with up to 36% of occluded joints in the testing set. As depicted in Figure 5, our approach was the third best for occluded joints and the second best for visible joints because it has detection with favorable error in (mm) compared to other approaches with the given proportions of joints.

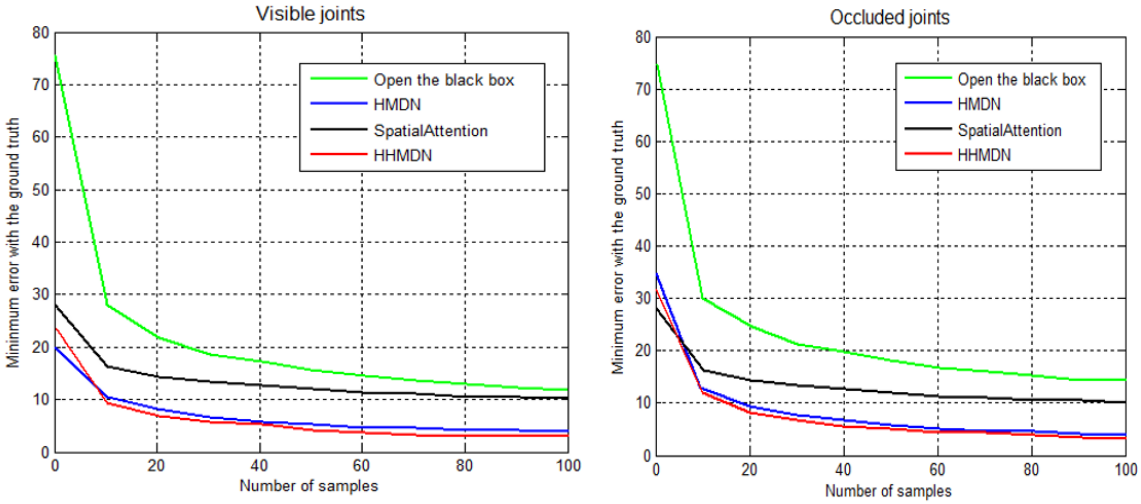

Figure 6: The number of success samples against the ground truth for different visible and occluded joints.

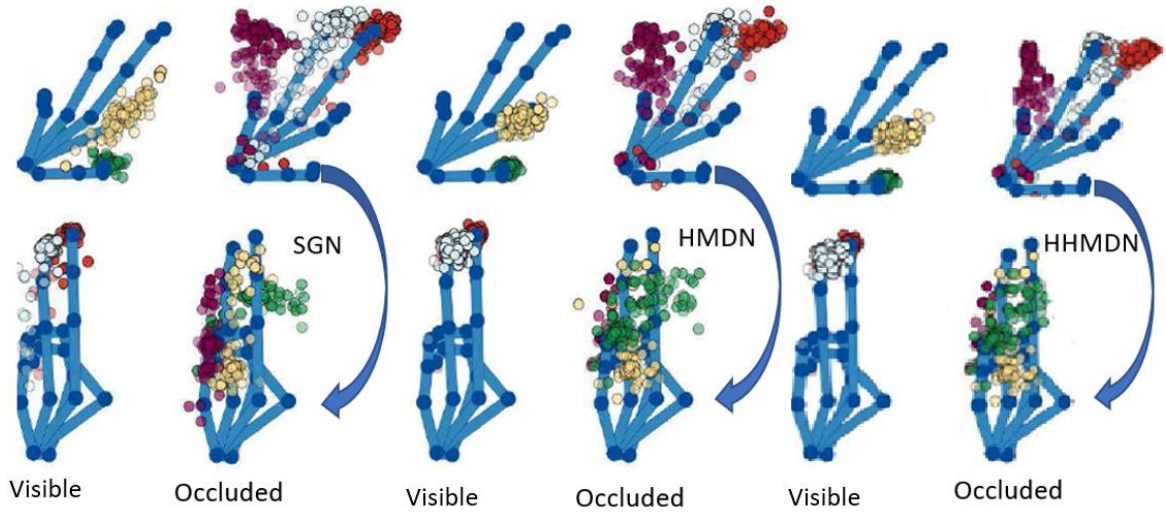

Figure 7: Illustration of samples for one tip joint scattered along the skeletons for visible and occluded joints.

For example, in Figure 5, when the error threshold is between 20 and 30, our method performed 8% better than Spatial attention and more than 5% better than the conventional HMDN and REN. Based on the number of success samples against the ground truth in Figure 6, our method also outperforms all the contending approaches with a slight improvement over HMDN.

We relate our technique with the three state-of-the-art approaches, Spatial attention Ye et al., (2016), open the black box Tang et al., (2015), and HMDN Ye and Kim (2018) by varying the number of distributed samples and measuring the least displacement error. Ye et al., (2016) proposed spatial attention predicts the joint by utilizing the CNN as a uni-modal Gaussian.

On the contrary, the Open Blackbox by Tang et al. (2015) used a decision forest with trees each modeled by GMM.

Our method outperformed all three methods in visible and occluded joints, indicating that a HHMDN improves awareness of the multiple modes of occluded joints more than the traditional HMDN. Finally, the distributions of joints with different samples are presented in Figure 7, from which the samples from Tang et al. (2015) are more diverse for fingertips and span a large region, while that of Ye and Kim (2018) are less diverse and compact but deviate more from the ground truth. Part of Figure 7 for other methods is obtained from the work of (Ye and Kim 2018).

CONCLUSION AND RECOMMENDATIONS

This paper presents an occlusion handling approach for hand pose estimation. Handling occlusion is an essential process for accurate hand pose estimation system. An accurate hand pose estimation system provides accurate interaction platform that achieves greater user experience. Handling occlusion caused by fingers can improve the accuracy of hand pose estimation which ultimately provides the possibility for developing a future seamless multi-touchless interaction.

Posed as a probabilistic based framework, fingers detection can be performed using detection scores. By representing finger parts as detection scores, the probability framework models the visibility of parts as a hidden variable or label.

Additionally, the visible label is generated by a single Gaussian distribution. To model occluded hand poses, the presented discriminative model is divided into two levels.

The top level which models the visibility constraints and the bottom level send-witched between uni-model distribution and multi-model distribution, based on the joint visibility. Evaluation of the approach was performed against two publicly available datasets consisting of several users and a significant percentage of occluded joints. The success joints were compared against the ground truth for both visible and occluded joints and the errors generated are presented in mm. The approach was compared with five state-of-the-art approaches Spatial attention, Deep model, DeepPrior, DeepPrior++, REN, and HMDN. While their approaches utilized large numbers of CNN layers like for example 20 CNN layers by DeepPrior and 50-Resnet layers, by DeepModel, their systems performed similar or slightly better with our proposed method.

Different from the contending approaches, the presented method in this paper is less ambiguous and cost-effective and can be run in a conventional CPU desktop computer. Expanding the proposed method to

accommodate hand flexibility i.e. degree of freedom using machine learning algorithms would be the future expansion of this proposed work.

REFERENCES

Asad, M., & Slabaugh, G. (2016). Learning Marginalization through Regression for Hand Orientation Inference. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 1215–1223. https://doi.org/10.1109/CVPRW.2016.154

Banzi, J., Bulugu, I., & Ye, Z. (2020). Learning a deep predictive coding network for a semi-supervised 3D-hand pose estimation. *IEEE/CAA Journal of Automatica Sinica*, **7**(5): 1371–1379. https://doi.org/10.1109/JAS.2020.1003090

Banzi, J. F., Bulugu, I., & Ye1, Z. (2019). Learning hand latent features for unsupervised 3D hand pose estimation. *Journal of Autonomous Intelligence*, **2**(1): 1. https://doi.org/10.32629/jai.v2i1.36

Chen, X., Wang, G., Guo, H., & Zhang, C. (2020). Pose guided structured region ensemble network for cascaded hand pose estimation. *Neurocomputing*, **395**: 138–149. https://doi.org/10.1016/j.neucom.2018.06.097

Chen, Y., Wang, J., Zhu, B., Tang, M., & Lu, H. (2019). Pixelwise Deep Sequence Learning for Moving Object Detection. *IEEE Transactions on Circuits and Systems for Video Technology*, **29**(9): 2568–2579. https://doi.org/10.1109/TCSVT.2017.2770319

Duan, L., Shen, M., Cui, P. S., & Guo, P. Z. (2019). *Estimat ing 2D Mult i-hand Poses from Single D epth Images*. 257–272.

Farahanipad, F., Rezaei, M., Dillhoff, A., Kamangar, F., & Athitsos, V. (2021). A Pipeline for Hand 2-D Keypoint Localization Using Unpaired Image to Image Translation. *ACM International Conference Proceeding Series*, 226–233. https://doi.org/10.1145/3453892.3453904

Ge, L., Liang, H., Yuan, J., Member, S., & Thalmann, D. (2018). *Robust 3D Hand Pose Estimation from Single Depth Images using Multi-View CNNs*. **7149**(c): 1–15. https://doi.org/10.1109/TIP.2018.2834824

Ge, L., Liang, H., Yuan, J., & Thalmann, D. (2018). Robust 3D Hand Pose Estimation from Single Depth Images Using Multi-View CNNs. *IEEE Transactions on Image Processing*, **27**(9): 4422–4436. https://doi.org/10.1109/TIP.2018.2834824

Guo, H., Wang, G., & Chen, X. (2017). *Towards Good Practices for Deep 3D Hand Pose Estimation*. *arXiv preprint arXiv:1707.07248*.

Guo, H., Wang, G., Chen, X., Zhang, C., Qiao, F., & Yang, H. (2018). *Region ensemble network: improving convolutional network for hand pose estimation*. In *2018 IEEE International Conference on Image Processing (ICIP)* (pp. 4512-4516). IEEE.

Guo, J., Han, K., Wang, Y., Zhang, C., Yang, Z., Wu, H., Chen, X., & Xu, C. (2020). *Hit-Detector : Hierarchical Trinity Architecture Search for Object Detection*. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11405-11414. 2020.

Jariwala, S. S., & Parmar, P. N. (2017). *Survey on Hand Gesture Recognition Using American Sign Language Abstract :* **3**(3): 182–185.

Madadi, M., Escalera, S., Bar, X., & Gonz, J. (2017). *End-to-end Global to Local CNN Learning for Hand Pose Recovery in Depth Data ∗ arXiv : 1705 . 09606v2 [cs . CV] 11 Apr 2018*.

Oberweger, M., & Lepetit, V. (2017). *DeepPrior ++ : Improving Fast and Accurate 3D Hand Pose Estimation*.

Oberweger, M., & Wohlhart, P. (2015). *Hands Deep in Deep Learning for Hand Pose Estimation*. *February*.

Oberweger, M., Wohlhart, P., & Lepetit, V.

(2018). *Generalized Feedback Loop for Joint Hand-Object Pose Estimation*. **14**(8): 1–15.

Oikonomidis, I., Kyriazis, N., & Argyros, A. A. (2011). Full DOF tracking of a hand interacting with an object by modelling occlusions and physical constraints. *Proceedings of the IEEE International Conference on Computer Vision*, 2088–2095. https://doi.org/10.1109/ICCV.2011.6126483

Pavllo, D., Delahaye, M., Porssut, T., Herbelin, B., & Boulic, R. (2019). *Computers & Graphics : X Real-time neural network prediction for handling two-hands mutual occlusions*. *2*. https://doi.org/10.1016/j.cagx.2019.100011

Qian, C., Sun, X., Wei, Y., Tang, X., & Sun, J. (2014). Realtime and robust hand tracking from depth. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1106–1113. https://doi.org/10.1109/CVPR.2014.145

Roy, K., Mohanty, A., & Sahay, R. R. (2017). *Deep Learning Based Hand Detection in Cluttered Environment Using Skin Segmentation*. 640–649.

Saha, S. (2018). *3D Hand Pose Tracking from Depth Images using Deep Reinforcement Learning*.

Q. Y., & Kim, T.-K. (2018). *3D Hand Pose Estimation Using Convolutional Neural Networks*. *In Proceedings of the European conference on computer vision (ECCV) (pp. 801-817).*

Tang, C., Feng, Y., Yang, X., Zheng, C., & Zhou, Y. (2017). *The Object Detection Based on Deep Learning*. https://doi.org/10.1109/ICISCE.2017.156

Tang, C., Ling, Y., Yang, X., Jin, W., & Zheng, C. (2018). *applied sciences Multi-View Object Detection Based on Deep Learning*. https://doi.org/10.3390/app8091423

Tang, D., Taylor, J., Kohli, P., Keskin, C., Kim, T. K., & Shotton, J. (2015). Opening the black box: Hierarchical sampling optimization for estimating human hand pose. *Proceedings of the IEEE International Conference on Computer Vision*, *2015 Inter*, 3325–3333. https://doi.org/10.1109/ICCV.2015.380

Xiong, F., Zhang, B., Xiao, Y., Cao, Z., Yu, T., Zhou, J. T., & Star, A. (2019). A2J : Anchor-to-Joint Regression Network for 3D Articulated Pose Estimation from a Single Depth Image. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, *Iccv*, 793–802. https://doi.org/10.1109/ICCV.2019.00088

Xiongwei, W., & Hoi, S. C. H. (2020). *Recent Advances in Deep Learning for Object Detection*. *January*.

Ye, Q., & Kim, T.(2018) (n). *Occlusion-aware Hand Pose Estimation Using Hierarchical Mixture Density Network*. *In Proceedings of the European conference on computer vision (ECCV) (pp. 801-817).*

Ye, Q., Yuan, S., & Kim, T. K. (2016). Spatial attention deep net with partial PSO for hierarchical hybrid hand pose estimation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9912 LNCS*, 346–361. https://doi.org/10.1007/978-3-319-46484-8_21

Yuan, S., & Bj, G. G. (2018). *Depth-Based 3D Hand Pose Estimation : From Current Achievements to Future Goals Depth-Based 3D Hand Pose Estimation : From Current Achievements to Future Goals*. *December 2018.* In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2636-2645).. https://doi.org/10.1109/CVPR.2018.00279

Zhou, T., Li, Z., & Zhang, C. (2019). *Enhance the recognition ability to occlusions and small objects with Robust Faster R-CNN*. *September*. *International Journal of Machine Learning and Cybernetics*, **10**(11): pp.3155-3166https://doi.org/10.1007/s13042-019-01006-4