Computers & Graphics 85 (2019) 85–91

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Technical Section

Simple and effective deep hand shape and pose regression from a single depth image

Jameel Malik[a,b,c,e,*], Ahmed Elhayek[a,d], Fabrizio Nunnari[a,**], Didier Stricker[a,b,e]

[a] German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany
[b] Department of Informatics, University of Kaiserslautern, Kaiserslautern 67653, Germany
[c] School of Electrical Engineering and Computer Science (SEECS), National University of Sciences and Technology (NUST), Islamabad 44000, Pakistan
[d] University of Prince Mugrin (UPM), Madinah 20012, Saudi Arabia
[e] TUKL (Technical University of Kaiserslautern), Kaiserslautern, Germany

ARTICLE INFO

Article history:
Received 12 July 2019
Revised 7 October 2019
Accepted 8 October 2019
Available online 14 October 2019

Keywords:
Convolutional neural network (CNN)
Depth image
Three-dimensional (3D) hand mesh and pose

ABSTRACT

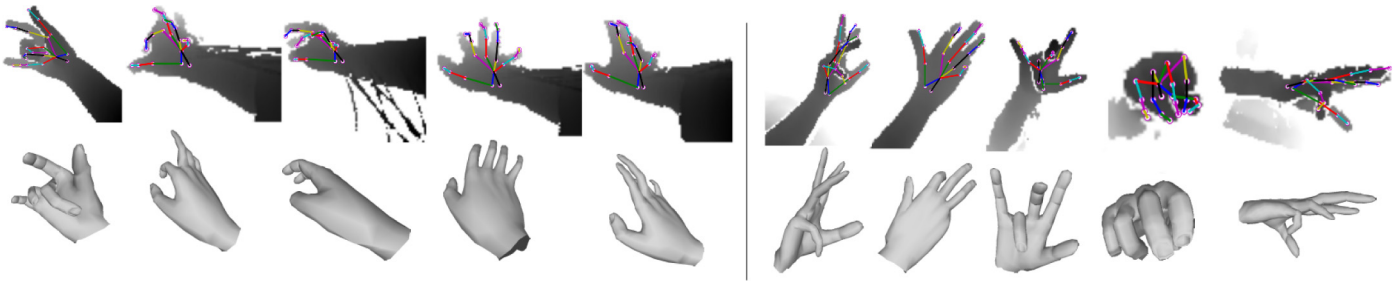
Simultaneously estimating the 3D shape and pose of a hand in real time is a new and challenging computer graphics problem, which is important for animation and interactions with 3D objects in virtual environments with personalized hand shapes. CNN-based direct hand pose estimation methods are the state-of-the-art approaches, but they can only regress a 3D hand pose from a single depth image. In this study, we developed a simple and effective real-time CNN-based direct regression approach for simultaneously estimating the 3D hand shape and pose, as well as structure constraints for both egocentric and third person viewpoints by learning from the synthetic depth. In addition, we produced the first million-scale egocentric synthetic dataset called SynHandEgo, which contains egocentric depth images with accurate shape and pose annotations, as well as color segmentation of the hand parts. Our network is trained based on combined real and synthetic datasets with full supervision of the hand pose and structure constraints, and semi-supervision of the hand mesh. Our approach performed better than the state-of-the-art methods based on the SynHand5M synthetic dataset in terms of both the 3D shape and pose recovery. By learning simultaneously using real and synthetic data, we demonstrated the feasibility of hand mesh recovery from two real hand pose datasets, i.e., BigHand2.2M and NYU. Moreover, our method obtained more accurate estimates of the 3D hand poses based on the NYU dataset compared with the existing methods that output more than joint positions. The SynHandEgo dataset has been made publicly available to promote further research in the emerging domain of hand shape and pose recovery from egocentric viewpoints (https://bit.ly/2WMWM5u).

1. Introduction

In recent years, major advances in deep learning and the availability of low cost depth sensors have significantly facilitated research into the 3D hand pose estimation task. This task has important applications in computer vision and graphics, such as the handling of objects in virtual environments and signing in the air [1]. Deep learning-based direct regression methods (discriminative) [3,4] can achieve state-of-the-art accuracy with depth-based hand pose datasets. However, these methods only regress the sparse 3D hand pose from a single depth image, and thus they do not consider the hand's structure. A hand mesh (i.e., shape) is a richer and more useful representation that is much needed in many computer graphics applications, such as animating a personalized hand shape in virtual reality (VR) and augmented reality (AR). An egocentric viewpoint is important in these applications because the camera is mounted on the head. Inferring only a sparse hand pose is not sufficient for modern gaming environments. Hence, recovering real hand shapes from an egocentric viewpoint is essential, but this is a very challenging task because of finger occlusion and field-of-view limitations. Algorithms for direct regression hand shape estimation from an egocentric viewpoint are not available because none of the existing hand pose datasets provide egocentric ground truth shape information. Annotating real images with shape representations is highly challenging and the results can be sub-optimal. In addition, the SynHand5M synthetic dataset [5] provides information about hand shapes and poses from a third person viewpoint but it lacks realism. The differences between real

* Corresponding author at: German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany.
** Corresponding author.
E-mail addresses: jameel.malik@dfki.de (J. Malik), ahmed.elhayek@dfki.de (A. Elhayek), fabrizio.nunnari@dfki.de (F. Nunnari), didier.stricker@dfki.de (D. Stricker).

https://doi.org/10.1016/j.cag.2019.10.002
0097-8493/



(a) **Egocentric** view (b) **3rd person** view

Fig. 1. *3D hand shape and pose recovery from a real depth image.* We developed a simple and effective real-time convolutional neural network (CNN)-based direct regression approach for simultaneously recovering the 3D hand shape and pose from a single depth image. (a) and (b) show the recovery of the real hand pose and shape based on samples from the BigHand2.2M [2] dataset. Feasible real hand shapes can be recovered when no shape annotations are available in the real dataset.

and synthetic data can be decreased according to recent studies [6,7]. In particular, Malik et al. [5] estimated the hand shape and pose using a model-based deep network, which was trained end-to-end based on a combination of real and synthetic data. However, this method does not support the egocentric viewpoints that are essential for VR/AR applications and it is hindered by low accuracy and the generation of artifacts during shape estimation because of the limited representational capacity of their hand model, difficulty optimizing highly nonlinear kinematic models, and the linear blend skinning function inside the deep network.

In this study, we developed a simple and effective real-time convolutional neural network (CNN)-based approach for regressing the sparse hand joints and hand structure constraints, as well as a 3D hand mesh for both egocentric and third person viewpoints. The structure constraints (i.e., bone lengths, kinematic distances, and inter-finger distances [8]) are optimized simultaneously to maintain the structural relationships between the estimated joints [9]. To the best of our knowledge, directly regressing the 3D hand shape and pose based on a single depth image has not been reported previously. In addition, we produced SynHandEgo as the first egocentric synthetic hand dataset containing accurate ground truth data for 3D meshes, 3D poses, and color segmentations of hand parts. This dataset can facilitate the reconstruction of egocentric hand shapes using learning-based algorithms given that the annotation of real depth images with accurate hand shape information is almost impossible due to severe occlusion issues. Our joint training strategy based on real and synthetic datasets allows shapes to be reconstructed from real depth images without requiring ground truth hand shape data. Our method successfully recovered plausible hand shapes from two real benchmarks, i.e., NYU [10] and BigHand2.2M [2]. Our approach also outperformed the state-of-the-art methods with the SynHand5M [5] dataset. Experiments demonstrated that our method improved the hand pose estimation accuracy based on the NYU dataset compared with the existing methods that produce more than joint positions. We demonstrated that joint regression of the hand pose, shape, and structural constraints improved the accuracy of pose estimation compared with the baseline architecture [11] by 10.6% using the real NYU dataset and by 20.7% using the synthetic SynHand5M dataset. The qualitative shape estimation results were improved compared with the state-of-the-art method [5] using the NYU dataset. The main contributions of the present study are listed as follows.

1. We developed simple and effective real-time CNN-based direct regression approach for simultaneously estimating full hand mesh and 3D pose from a single (egocentric view (Fig. 1(a)) and a third person viewpoint (Fig. 1(b))) depth image. Our approach enhanced the accuracy of 3D hand pose estimation by simultaneously estimating the full 3D hand mesh and the 3D pose (see Sections 5.2 and 5.3).
2. We constructed the first egocentric synthetic hands dataset called SynHandEgo, which provides accurate ground truth data for the 3D shapes, poses, and color segmentation of hand parts from an egocentric viewpoint. This dataset has been made publicly available to promote further research in this emerging domain.

2. Related work

Yuan et al. [2] provided a detailed comparative analysis of several state-of-the-art deep learning-based networks for 3D hand pose estimation from a single depth image. Zhou el al. [12] presented a hand model-based deep learning approach that ensures the geometric validity of the estimated 3D pose. This approach was extended by Malik et al. [5,13] for learning bone lengths and the shape of the hand skeleton from a single depth image. However, these methods are hindered by their low accuracy because the kinematic model is highly nonlinear and difficult to optimize in deep networks [9]. Obreweger et al. [14] proposed a CNN-based feedback network for estimating and refining 3D poses. Ye et al. [15] introduced a hybrid method based on a spatial attention mechanism and hierarchical particle swarm optimization. Recently proposed CNN-based methods ([4,16–20]) obtain higher accuracy compared with the model-based deep learning approaches. In particular, Ge et al. [21] regressed a 3D hand pose using a 3D-CNN. Rad et al. [4] simultaneously learned the mappings between synthetic images and their real counterparts with the 3D pose. Moon et al. [3] introduced a voxel-to-voxel network for accurate 3D pose estimation. However, these methods may yield unstable 3D poses because they lack some physical constraints. A solution to this problem was proposed by Malik et al. [8], which we integrate into our approach.

All of the approaches mentioned above focused only on estimating sparse 3D hand poses using deep neural networks. Malik et al. [5] firstly proposed an end-to-end model-based deep learning network for estimating 3D hand poses and shapes from a single depth image. However, the shape representation capacity of their hand model is limited. Moreover, their method is affected by the generation of artifacts because of the difficulties associated with optimizing complex hand shapes, bone scales, and joint angle parameters inside the deep network. In a recent study, Ge et al. [22] estimated 3D hand shapes and poses from a single RGB image. They also addressed a different and harder problem but their approach depends greatly on pseudo-ground truth data for the real hand shapes. Moreover, the 3D poses were estimated based on reconstructed 3D shapes, and thus the accuracy of the estimated 3D poses was dependent on the recovered 3D shapes, which is

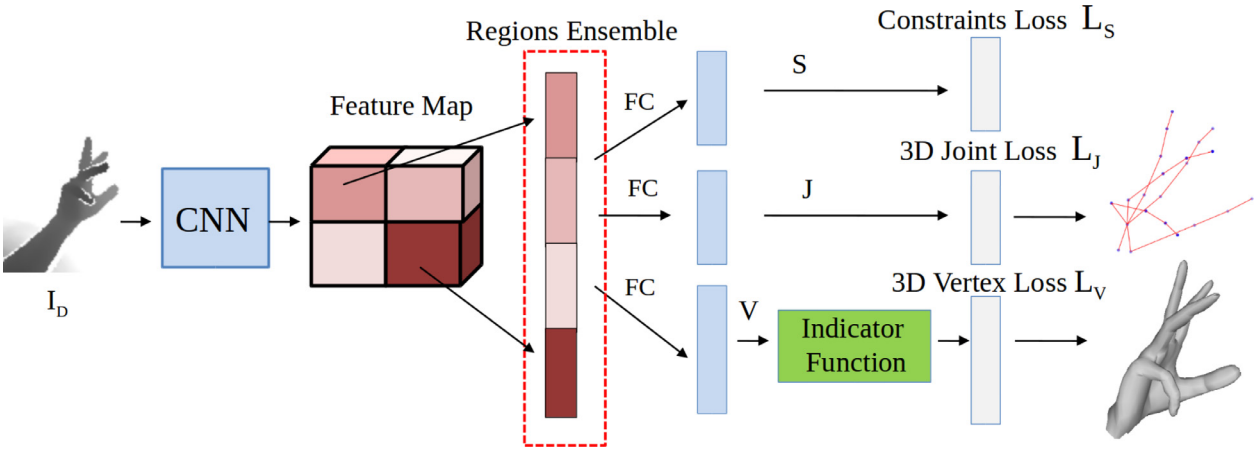


Fig. 2. Overview of our approach for hand shape and pose regression. A depth input (I_D) is given to the convolutional neural network (CNN), which provides a feature map for four distinct regions of I_D. After applying the regions ensemble strategy [11], the structural constraints S, joint positions J, and mesh vertices V are directly regressed. The indicator function specifies whether the ground truth is available for the vertices or not, which allows the network to be trained using a combination of real and synthetic data. FC: fully connected layer.

not feasible in practice because accurate real hand shape ground truth data are not available in the existing benchmarks. To the best of our knowledge, a direct regression approach is not available for simultaneously estimating the 3D hand shape and pose from a single depth image. Thus, we developed a simple and effective real-time CNN-based 3D hand pose and shape regression approach for both egocentric and third person viewpoints by learning from the synthetic depth. We showed that our approach effectively reconstructed hand shapes based on real images and obtained better qualitative results compared with the state-of-the-art DeepHPS method [5] with the NYU hand pose dataset. In order to facilitate the learning of real hand shapes from an egocentric viewpoint, we produced the first egocentric synthetic hands dataset (i.e., SynHandEgo), which provides 1M depth images with accurate 3D shape and pose annotations as well as color segmentations of hand parts. Given that the annotation of real images of hand shapes from an egocentric viewpoint is extremely difficult because of frequent finger occlusion, our dataset will facilitate the development of algorithms for reconstructing hand shapes from egocentric viewpoints.

3. Proposed approach

Fig. 2 presents an overview of our approach. Given a single view and gray scale depth image I_D, the task involves directly regressing the hand joints $J \in \mathcal{R}^{3xP}$, mesh vertices $V \in \mathcal{R}^{3xN}$, and structure constraints S, where P is the number of joints and $N = 1193$ is the number of vertices. As mentioned earlier (Section 2), direct hand pose regression methods (e.g., [4,11]) may lead to unstable pose estimation because they do not explicitly consider the hand structure in the learning process. In order to maintain the structural relationships between the estimated joints, we follow the method proposed by Malik et al. [8] and simultaneously optimize $S \in \mathcal{R}^{(3xP-8)}$, including the bone lengths, kinematic distances, and inter-finger distances. The ground truth for S can easily be obtained from J (see [8] for details). The loss equations are given by the Euclidean distances as:

$$L_J = \frac{1}{2}\|J - J_{GT}\|^2, \quad L_V = \frac{1}{2}\|V - V_{GT}\|^2,$$
$$L_S = \frac{1}{2}\|S - S_{GT}\|^2 \tag{1}$$

where L_J, L_V, and L_S represent the joint, vertex, and constraint losses, respectively, and J_{GT}, V_{GT}, and S_{GT} are the ground truths for the pose, shape, and constraints. The combined loss equation can be written as:

$$L = L_J + L_S + \mathbb{1}L_V, \tag{2}$$

where $\mathbb{1}$ is an indicator function. During the forward pass, $\mathbb{1}$ selects V only for synthetic images using a binary flag value. This value is 1 for synthetic images and 0 for real images. Similarly, back-propagation for V is disabled for real images.

3.1. Network architecture

We employ a state-of-the-art CNN architecture [11] used only for pose regression and modify it to simultaneously regress S, J, and V (see Fig. 2). A depth input I_D with a size of 96×96 is passed through a shared CNN to produce the feature map measuring 12×12×64. The shared CNN comprises six convolutional layers using a filter size of 3×3. Three max pooling layers with a stride of 2 are used after each pair of convolutional layers. Two residual connections are made between the last two pairs of convolutional layers (for more details of the CNN architecture, see [11]). The feature map is divided into four regions where each measures 6x6x64. These regions are flattened to produce fully connected (FC) layers where each has a size of 2048. The FC layers are then ensembled using feature concatenation to create a high-dimensional feature vector with a size of 8192. Finally, three lower dimensional regression FC layers are connected to produce S, J, and V separately (as shown in Fig. 2). V is semi-supervised (see Eq. (2)) so we introduce a layer that implements an indicator function $\mathbb{1}$. This layer forwards only the valid V to the vertex loss layer. Back-propagation is enabled only for synthetic images.

4. Egocentric synthetic dataset

The SynHandEgo dataset was generated with Blender [23] using a humanoid created with the MB-Lab add-on [24]. A virtual character was created and provided with an inverse kinematics controller for moving and rotating its right hand, wrist, elbow, shoulder, and clavicle. Rotation limits were set for the whole arm, including the fingers, according to realistic ergonomic ranges. A virtual camera that simulated a Senz3D depth sensor [25] was mounted between the eyes of the character. Mueller et al. [26] mounted the camera on a shoulder whereas our camera position is optimal for VR/AR applications. The hand was set at the initial position in front of the character (as shown in Fig. 3). Our custom code routine generated 1 million hand configurations by uniformly sampling random

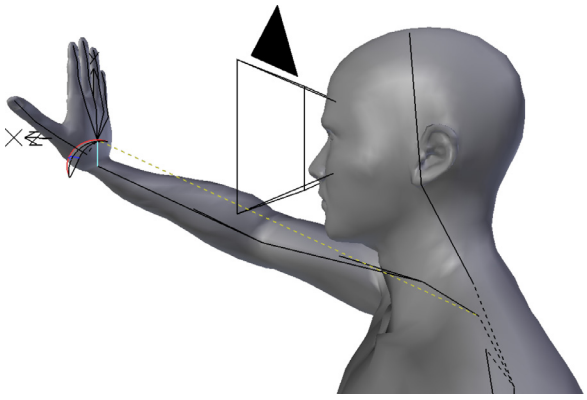


Fig. 3. Setup of the Blender scene used to create the *SynHandEgo* dataset. The virtual camera was placed between the eyes of the character.

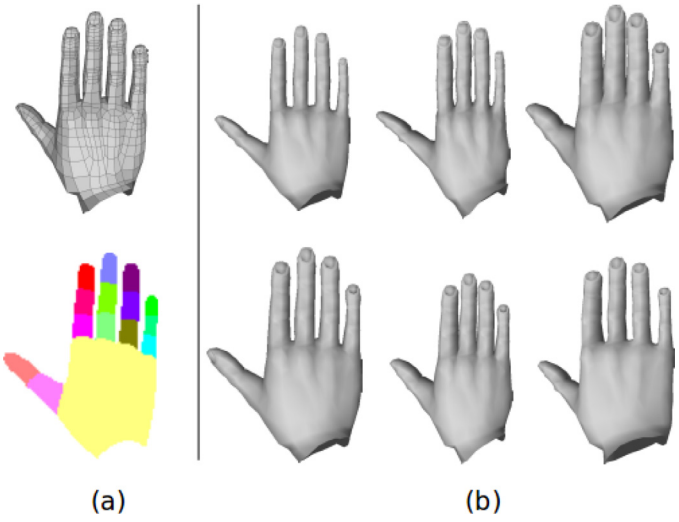


Fig. 4. Proposed egocentric hands dataset. (a) We provide accurate hand segmentation and full 3D hand mesh representations in addition to the 3D pose and depth image. (b) Large hand shape variation. We used realistic hand size measurements provided in the DINED anthropometric database [27].

values within the allowed ranges for three categories of degrees of freedom. The first category was hand rotation where the hand was rotated on three axes in ranges that respected ergonomically realistic positions. The second category was finger rotation where all of the fingers were simultaneously rotated within their rotation limits, which allowed the inclusion of uncomfortable or even unrealistic poses in order to provide samples from border-line conditions in the explored space. A collision detection routine discarded the poses where the fingers penetrate each other. Fig. 5 shows samples of the depth images with overlaid ground truth 3D poses and the respective 3D hand shapes. The third category was hand shapes where the size and proportions of the hand were modulated in the following seven dimensions: *length, mass, size, palm length, inter-finger distance, finger length, and fingertip size*. The realism of the resulting hand proportions was ensured by measuring the hand sizes within the ranges provided in the DINED anthropometry dataset [27] (see Fig. 4(b)). Moreover, we provide accurate color segmentations of the hand parts, as shown in Fig. 4(a), which may be useful for hand part segmentation-based methods such as that proposed by Neverova et al. [28]. Segmentation was conducted at the polygon level by manual assignment of the polygon's color to either a phalanx, the palm, or forearm. The colors were generated by sequentially assigning the value of each RGB component to 0.0, 0.5, or 1.0. These colors might seem similar to the human

Table 1
Quantitative results obtained using the *SynHand5M* [5] test set. The results show that simultaneously learning the pose, shape, and structural constraints improved the accuracy of 3D hand pose estimation by 20.7% compared with the baseline architecture (J) [11]. All of the errors are reported in millimeters (mm).

Method \Error(mm)	3D Joint Loc.	3D Vertex Loc.
DeepModel [12]	11.36	–
HandScales [13]	9.67	–
DeepHPS [5]	6.3	11.8
J [11]	5.83	–
J ∪ V [Ours]	5.14	7.12
J ∪ V ∪ S [Ours]	**4.62**	**6.61**

eye but their values are very different in the RGB color space. We divided the dataset into a training set T_E containing 900K images and a test set of 100K frames. As proposed by Malik et al. [5], P and N are the same in SynHand5M.

5. Experiments and results

In the following, we provide the implementation details and evaluations based on three public datasets (i.e., real NYU [10], real BigHand2.2M [2], and synthetic SynHand5M [5]) and the proposed SynHandEgo dataset. SynHand5M comprises 4.5M training set (T_S) and 500K test set. The NYU provides a training set (T_N) containing 72,757 images from a third person viewpoint and a test set with 8252 frames. BigHand2.2M is the largest real dataset where it contains 956K depth frames with mixed egocentric and third person viewpoints. The test set for pose estimation contains 296K images. However, the annotations for the test set are not publicly available. Therefore, we treated 90% of 956K (i.e., 860K) as the training set (T_B) and the remaining frames (i.e., 96K) as the test set. Three common evaluation metrics are used for public comparisons comprising the average 3D joint location error, average 3D vertex location error, and fraction of images within thresholds.

5.1. Implementation details

We used the method described by Guo et al. [11] for standard preprocessing of the depth frames and the annotations. All of the images were normalized using the hand mass centers and a bounding box with a fixed size of 150. The final values for the preprocessed depth images and annotations were in the range of $[-1, 1]$. To augment the data, we randomly scaled and rotated the training data in the ranges of [0.8, 1.1] and $[-45°, 45°]$, respectively. We trained our network using the Caffe framework [29]. In order to conduct training based on T_S, the learning rate (LR) was set to 0.0005, SGD momentum to 0.9, and the batch size to 512. For combined real and synthetic data training, LR was set to 0.00005. Training was performed on a desktop PC with an Nvidia Geforce GTX 1070 GPU. A single forward pass required only *2.2* ms to generate both the 3D pose and shape. The networks were trained until they reached convergence.

5.2. Synthetic hand shape and pose recovery

SynHand5M dataset: We trained three different implementations of our network to determine the effectiveness of simultaneously learning the hand shape, pose, and structure constraints. In the first implementation, which is similar to that given by Guo et al. [11], 22 3D joint key-points from the SynHand5M dataset (i.e., $J \in \mathcal{R}^{66}$) were directly regressed. The network converged after 1000K iterations using LR = 0.05. The quantitative results are shown in Table 1. In the second implementation, the shape was optimized together with the pose ($J \cup V$) by adding two additional layers to the first implementation: an FC layer with a size of 3579

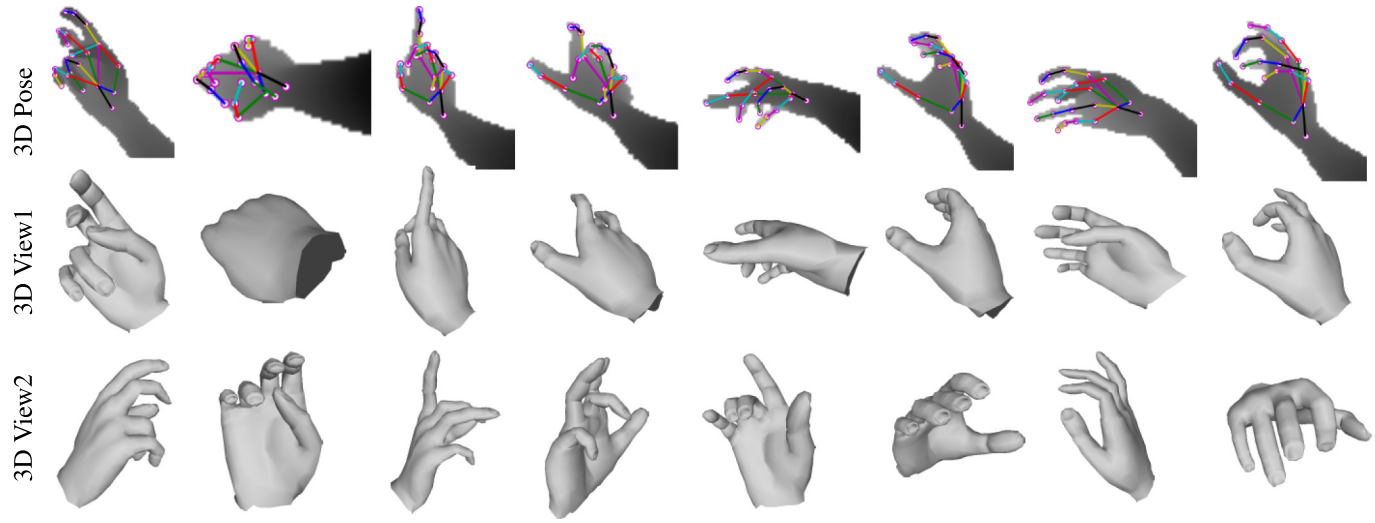


Fig. 5. Sample images from the *SynHandEgo* egocentric dataset. Preprocessed depth images with overlaid ground truth 3D hand poses and the respective ground truth 3D hand meshes from two different viewpoints. Our dataset includes a wide range of hand poses and shapes.

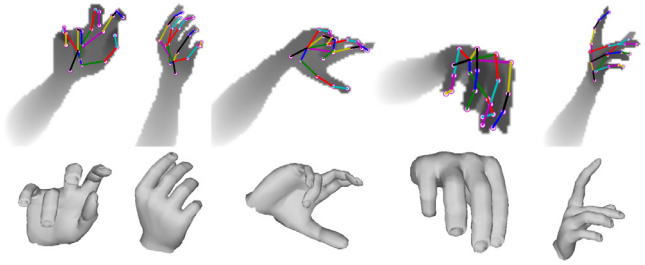


Fig. 6. Qualitative 3D hand pose and shape inference results based on the *SynHand5M* [5] synthetic dataset.

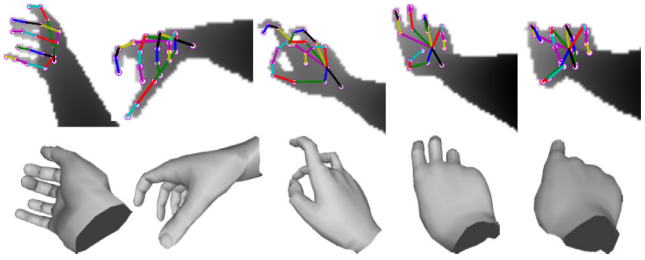


Fig. 7. Qualitative results obtained with the *SynHandEgo* dataset: 3D hand shape and pose recovery from sample egocentric images with high occlusion.

for regressing the mesh vertices and a non-parametric indicator function layer (as shown in Fig. 2). This network implementation required 2000K iterations to converge with LR = 0.0005. The estimated pose improved by 11.8% compared with the first implementation. In addition, the shape estimation accuracy improved by 39.6% compared with DeepHPS [5]. In the third implementation, the structural constraints were learned simultaneously with the pose and shape (J∪V∪S), where convergence occurred in 2500 K iterations with LR = 0.0005 and the performance was better than the other approaches (Table 1). This network implementation contained an additional FC layer with a size of 58 to regress S (see Section 3). Fig. 6 shows the qualitative 3D pose and shape estimation results for some challenging hand poses.

SynHandEgo dataset: We trained our network (J∪V∪S) on T_E with full supervision based on the joint positions, mesh vertices, and structural constraints. The network required 500K iterations to converge with LR = 0.0005. Fig. 7 shows the qualitative 3D pose and shape recovery results. Quantitatively, the joint and vertex location errors with the test set were 5.5 mm and 7 mm, respectively.

5.3. Real hand shape and pose recovery

The synthetic data provided weak supervision of the mesh vertices for real hand shape recovery. However, training using both the synthetic and real data allowed our network to learn the shapes and poses of real hands despite the lack of ground truth shape information for real images. We aimed to simultaneously recover both the real hand shape and the pose. However, for the sake of completeness, we conducted comparisons with the state-of-the-art hand pose estimation methods using the NYU dataset.

NYU dataset: We trained our network (J∪V∪S) based on four datasets, which were combined to form a single training set: $T_{NBSE} = T_N \cup T_B \cup T_S \cup T_E$. Not all of the joints in the NYU dataset were consistent with the other datasets, so we followed the method proposed by Malik et al. [5] for selecting the 16 closely matching joints present in all of the datasets. After training based on T_{NBSE} with full supervision for J and S, and semi-supervision for V (see Eq. (2)), we recovered the plausible 3D hand shapes from the NYU dataset. Fig. 9 shows reconstructions of the 3D shapes obtained from the sample test depth images in the NYU dataset. The network required 5000K iterations to reach convergence using LR = 0.00005. We qualitatively compared our reconstructed 3D hand shapes with those obtained using the state-of-the-art DeepHPS method [5]. DeepHPS is hindered by the generation of artifacts during shape reconstruction because of the limited representational capacity of the hand model as well as difficulties optimizing complex hand shapes, bone scales, and joint angle parameters inside the deep network. Moreover, all of these parameters were implicitly learned. To compare the performance of our method in the 3D pose estimation task, we trained our network using a subset of 14 joints from the NYU dataset and the corresponding closely matching joints in other datasets. Fig. 8 shows quantitative comparisons with several state-of-the-art methods. Fig. 8 (left) shows the errors based on individual joints from the NYU dataset and the mean error. Fig. 8 (right) shows the fraction of successful frames within various thresholds (in *mm*). The joint location errors over all of the test frames are presented in Table 2. As mentioned earlier, DeepHPS [5] is the only existing method that estimates both 3D the hand shape and pose with the NYU dataset. Our approach

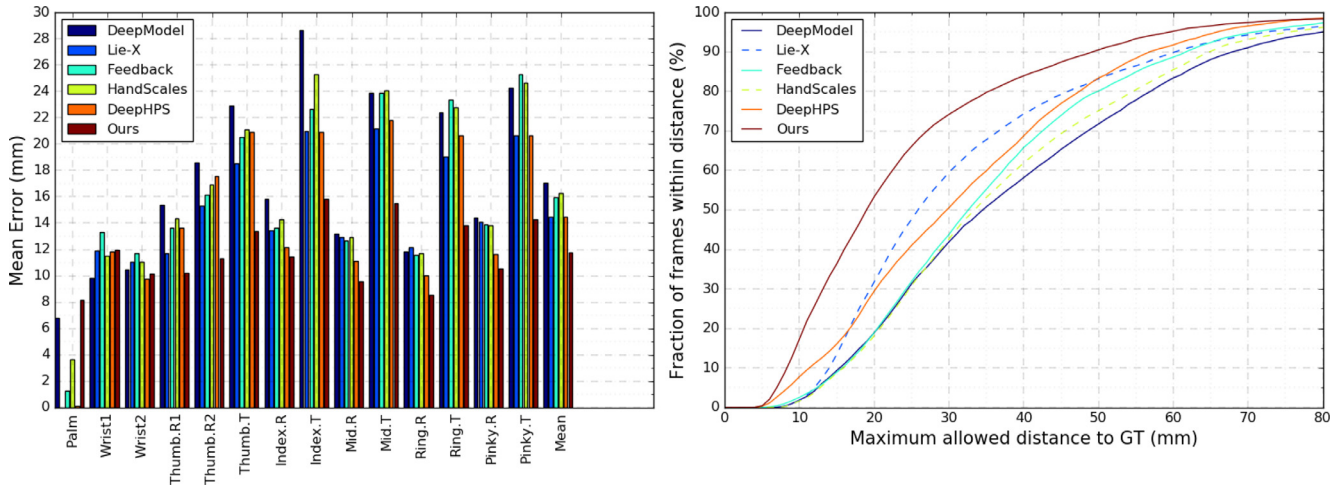


Fig. 8. Comparisons with state-of-the-art methods based on the NYU [10] dataset: mean error (left) and fraction of successful frames (right). Our method improved the accuracy in the 3D pose estimation task compared with the state-of-the-art methods *that output more than the joint positions*, i.e., DeepModel [12], HandScales [13], lie-X [30], Feedback [14], and DeepHPS [5].

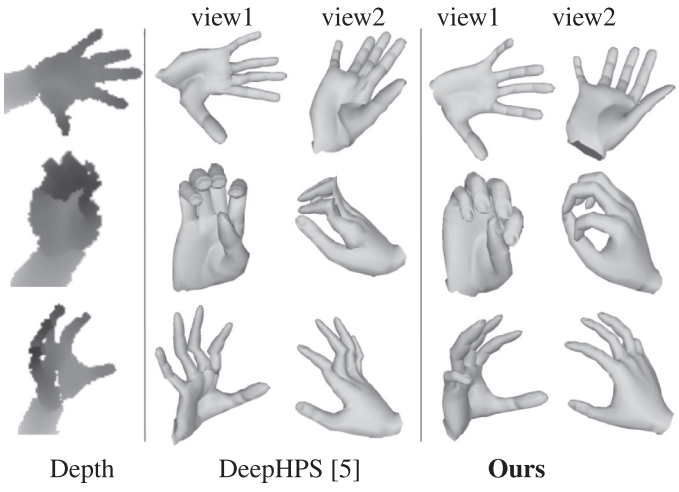


Fig. 9. Qualitative comparison of the 3D hand shapes recovered with the state-of-the-art DeepHPS method [5] using the *NYU* [10] dataset. The DeepHPS method produced artifacts due to the limited representational capacity of their hand model as well as difficulties optimizing complex hand shapes, bone scales, and joint angle parameters. By contrast, our algorithm recovered more accurate real hand shapes.

Table 2
Quantitative comparison based on the **NYU** [10] test set using several state-of-the-art methods. Our method improved the accuracy compared with the methods that output more than the joint positions. In addition, our method improved the accuracy by 10.6% compared with the baseline method [11]. All of the errors are reported in mm.

Methods	3D Joint Location Error
Crossing Nets [31]	15.5
Feedback [14]	15.9
DeepHPS [5]	14.2
REN-4x6x6 [11]	13.2
Ours	*11.8*

performed significantly better than the DeepHPS method in the pose estimation task, thereby demonstrating the benefit of directly regressing the dense mesh together with the sparse joints. In addition, our approach improved the accuracy by 10.6% compared with the baseline REN architecture [11], which only regresses the joint positions. Thus, our method significantly improved the accuracy of hand pose estimation compared with the state-of-the-art methods

VR/AR applications.

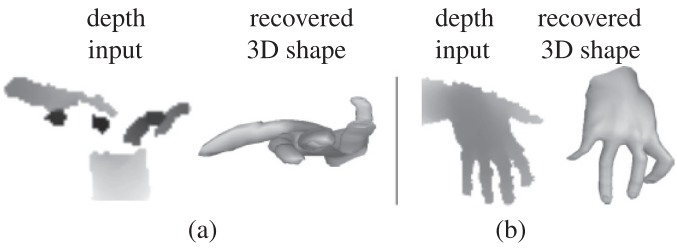


Fig. 10. Failure case: Our method failed to recover correct hand shapes from real depth images when large amounts of depth information were missing (a), or when the hand pose in the real depth image was not covered by the pose space in the synthetic dataset, which was a rare occurrence (b).

that produce more than joint positions, i.e., DeepModel [12], HandScales [13], Feedback [14], and DeepHPS [5]. In addition to 3D pose estimation, DeepModel [12] estimates the joint angle parameters, HandScales [13] predicts the joint angles and bone lengths in the hand skeleton, Feedback [14] synthesizes 2D depth images, and the DeepHPS [5] method estimates the joint angles, bone lengths, complex hand shape parameters, and 3D hand mesh vertices. We did not conduct comparisons with direct regression methods [3,4] that do not incorporate hand structure in their pipelines and that only produce 3D hand poses.

BigHand2.2M dataset: We combined the BigHand2.2M, SynHand5M, and SynHandEgo datasets into one training set: $T_{BSE} = T_B \cup T_S \cup T_E$. The 21 joints in the BigHand2.2M dataset were consistent with the joints in both of the synthetic datasets. However, the palm center positions were missing from the BigHand5M dataset. Thus, we calculated the palm centers by taking the mean of the centers of the metacarpal joints and the wrist joint. Hence, all 22 joints were used in the combined training process. After simultaneously training with the real and synthetic data, we recovered 3D hand shapes for the challenging poses in BigHand2.2M from both egocentric and third person viewpoints (as shown in Fig. 1). The network converged within 5000K iterations using LR = 0.00005. Quantitatively, the 3D joint location error with the test set was 13.5 mm.

Failure cases: When significant depth information was missing from the real images, our algorithm failed to recover the plausible hand shapes (as shown in Fig. 10(a)). In addition, if the hand

pose in the real depth image differed significantly from the pose space covered by the synthetic dataset, which occurred rarely, our network could not recover the correct hand shape (see Fig. 10(b)).

6. Conclusion and future work

In this study, we developed a simple and effective real-time CNN-based approach for directly regressing the 3D hand shape and pose for both egocentric and third person viewpoints by learning from the synthetic depth. We also produced the first egocentric synthetic hand pose dataset, which provides accurate annotations for 3D hand shapes and poses. In addition, we provided color segmentations of the hand parts. This dataset will facilitate future research into full hand shape and pose estimation from egocentric viewpoints, provided that obtaining the real hand shape ground truth is a hard and sub-optimal problem. Our network is trained simultaneously using real and synthetic data, which allows the successful recovery of plausible real hand shapes. Learning the pose and structural constraints is fully supervised, whereas shape learning is semi-supervised. Experiments showed that our approach performed better than the state-of-the-art methods with the synthetic SynHand5M dataset in terms of both hand shape and pose estimation tasks, and it also improved the pose estimation accuracy based on the real NYU dataset compared with the existing methods that output more than the joint positions. Inspired by the deep learning-based approaches proposed in previous studies [6,7], we plan to generate a realistic egocentric synthetic hands dataset and to extend our approach for improved hand shape and pose estimation to allow its direct use in real-time VR/AR applications.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was partially funded by the Federal Ministry of Education and Research of the Federal Republic of Germany as part of the research project VIDETE (Grant number 01IW18002).

References

[1] Malik J, Elhayek A, Ahmed S, Shafait F, Malik M, Stricker D. 3DAirSig: a framework for enabling in-air signatures using a multi-modal depth sensor. Sensors 2018;18(11):3872.
[2] Yuan S, Ye Q, Stenger B, Jain S, Kim T-K. BigHand2. 2M benchmark: hand pose dataset and state of the art analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2017. p. 2605–13.
[3] Moon G, Chang JY, Lee KM. V2V-posenet: voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. arXiv:171107399 2017.
[4] Rad M, Oberweger M, Lepetit V. Feature mapping for learning fast and accurate 3D pose inference from synthetic images. arXiv:171203904 2017.
[5] Malik J, Elhayek A, Nunnari F, Varanasi K, Tamaddon K, Heloir A, et al. Deephps: End-to-end estimation of 3d hand pose and shape by learning from synthetic depth. In: Proceedings of the International Conference on 3D Vision (3DV). IEEE; 2018. p. 110–19.
[6] Mueller F, Bernard F, Sotnychenko O, Mehta D, Sridhar S, Casas D, et al. Ganerated hands for real-time 3D hand tracking from monocular RGB. arXiv:171201057 2017a.
[7] Shrivastava A, Pfister T, Tuzel O, Susskind J, Wang W, Webb R. Learning from simulated and unsupervised images through adversarial training. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3; 2017. p. 6.
[8] Malik J, Elhayek A, Stricker D. Structure-aware 3D hand pose regression from a single depth image. In: Proceedings of the International Conference on Virtual Reality and Augmented Reality. Springer; 2018. p. 3–17.
[9] Sun X, Shang J, Liang S, Wei Y. Compositional human pose regression. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2; 2017. p. 7.
[10] Tompson J, Stein M, Lecun Y, Perlin K. Real-time continuous pose recovery of human hands using convolutional networks. ACM Trans. Gr. (ToG) 2014;33(5):169.
[11] Guo H, Wang G, Chen X, Zhang C, Qiao F, Yang H. Region ensemble network: improving convolutional network for hand pose estimation. In: Proceedings of the ICIP; 2017.
[12] Zhou X, Wan Q, Zhang W, Xue X, Wei Y. Model-based deep hand pose estimation. In: Proceedings of the IJCAI; 2016.
[13] Malik J, Elhayek A, Stricker D. Simultaneous hand pose and skeleton bone-lengths estimation from a single depth image. Proceedings of the 3DV2017.
[14] Oberweger M, Wohlhart P, Lepetit V. Training a feedback loop for hand pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision; 2015. p. 3316–24.
[15] Ye Q, Yuan S, Kim T-K. Spatial attention deep net with partial PSO for hierarchical hybrid hand pose estimation. In: Proceedings of the European Conference on Computer Vision. Springer; 2016. p. 346–61.
[16] Wan C, Probst T, Van Gool L, Yao A. Dense 3D regression for hand pose estimation. arXiv:171108996 2017a.
[17] Chen X, Wang G, Zhang C, Kim T-K, Ji X. SHPR-Net: deep semantic hand pose regression from point clouds. IEEE Access 2018;6:43425–39.
[18] Ge L, Ren Z, Yuan J. Point-to-point regression pointnet for 3D hand pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018. p. 475–91.
[19] Wang G, Chen X, Guo H, Zhang C. Region ensemble network: towards good practices for deep 3D hand pose estimation. J. Vis. Commun. Image Represent. 2018;55:404–14.
[20] Oberweger M, Lepetit V. Deepprior++: Improving fast and accurate 3D hand pose estimation. In: Proceedings of the ICCV workshop, 840; 2017. p. 2.
[21] Ge L, Liang H, Yuan J, Thalmann D. 3d convolutional neural networks for efficient and robust hand pose estimation from single depth images. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017.
[22] Ge L, Ren Z, Li Y, Xue Z, Wang Y, Cai J, et al. 3D hand shape and pose estimation from a single RGB image. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2019. p. 10833–42.
[23] Blender. Blender v2.79b. https://www.blender.org; 2019.
[24] MB-Lab MB-Lab v1.5.0.https://github.com/animate1978/MB-Lab;2019.
[25] Creative. Senz3D interactive gesture camera. https://us.creative.com/p/web-cameras/blasterx-senz3d; 2019.
[26] Mueller F, Mehta D, Sotnychenko O, Sridhar S, Casas D, Theobalt C. Real-time hand tracking under occlusion from an egocentric RGB-D sensor. In: Proceedings of the IEEE international conference on computer vision (ICCV); 2017. p. 1163–72.
[27] Molenbroek J. Dined, anthropometric database. https://dined.io.tudelft.nl/ – Accessed: 6 Feb2019.
[28] Neverova N, Wolf C, Nebout F, Taylor GW. Hand pose estimation through semi-supervised and weakly-supervised learning. Comput. Vis. Image Underst. 2017;164:56–67.
[29] Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, et al. Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the twenty-second ACM international conference on Multimedia. ACM; 2014. p. 675–8.
[30] Xu C, Govindarajan LN, Zhang Y, Cheng L. Lie-x: depth image based articulated object pose estimation, tracking, and action recognition on lie groups. Int J Comput Vis 2017;123:454–78.
[31] Wan C, Probst T, Van Gool L, Yao A. Crossing nets: Combining GANs and VAEs with a shared latent space for hand pose estimation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2017.