Applied Intelligence (2023) 53:8735–8745
https://doi.org/10.1007/s10489-022-03969-4

3D real-time human reconstruction with a single RGBD camera

Yang Lu[1] · Han Yu[1] · Wei Ni[2] · Liang Song[1]

Accepted: 5 July 2022 / Published online: 2 August 2022

Abstract

3D human reconstruction is an important technology connecting the real world and the virtual world, but most of previous work needs expensive computing resources, making it difficult in real-time scenarios. We propose a lightweight human body reconstruction system based on parametric model, which employs only one RGBD camera as input. To generate a human model end to end, we build a fast and lightweight deep-learning network named Fast Body Net (FBN). The network pays more attention on the face and hands to enrich the local details. Additionally, we train a denoising auto-encoder to reduce unreasonable states of human model. Due to the lack of human dataset based on RGBD images, we propose an Indoor-Human dataset to train the network, which contains a total of 2500 frames of action data of five actors collected by Azure Kinect camera. Depth images avoid using RGB to extract depth features, which makes FBN lightweight and high-speed in reconstructing parametric human model. Qualitative and quantitative analysis on experimental results show that our method can improve at least 57% in efficiency with similar accuracy, as compared to state-of-the-art methods. Through our study, it is also demonstrated that consumer-grade RGBD cameras can provide great applications in real-time display and interaction for virtual reality.

Keywords FBN · Real-time · Parametric model · RGBD · Indoor-Human

1 Introduction

Our interaction with the world is no longer limited to text, voice or video. As 3D reconstruction technology continues advancing, the representation method of people is developing towards 3D as well. Taking social network as an example, Facebook launched a virtual reality(VR) social platform Horizon [24], in which people are represented as different 3D cartoon models and are brought to a 3D world full of exploration. During the 2021 COVID-19, Horizon extends an application called Horizon Workrooms to enable people to work from home. Technically, it realizes 3D communication between multiple people.

The works [9, 27, 29, 35] of reconstructing static human body have achieved excellent results. These methods obtain depth data through active measurement or depth estimation, and then fuse 3D models using point cloud registration or deep learning. Unfortunately, the speed of these methods is always pretty slow.

Many scenarios require not only the ability to reconstruct a static human body, but the process in real time, such as remote conference [19], VR fitting [26] and online VR education [33]. The above applications commonly need to represent human in real time. To meet the speed requirement, most of the existing methods adopt 3D cartoon models and focus on the deformation of some certain parts, such as face [30], hand [28] or pose [2]. However, these methods are not straightforward or realistic enough. To get the same shape and texture with the real human body, there have been categories of solutions. One is the non-parametric reconstruction methods [6, 7, 11, 31] based on multi-camera calibration and point cloud fusion, while the other is the parametric reconstruction methods [3, 8, 14, 16] based on the deformation of 3D human model template. Parametric reconstruction methods have high performance without relying on expensive computing resources. Recent works show that it is becoming the mainstream method. Existing results have either high reconstruction accuracy or high reconstruction speed, but usually not both.

Depth cameras can have absolute advantages in 3D static human reconstruction tasks. The multimodal inputs enrich

✉ Liang Song
songl@fudan.edu.cn

[1] Academy of Engineering and Technology, Fudan University, Shanghai, China

[2] Shanghai Key Research Laboratory of NSAI, Shanghai, China

the dimension of data and reduce the influence of illumination and other factors. Additionally, the depth data obtained by active measurement can help to reduce the difficulty of extracting depth features. Therefore, using depth data is expected to improve the speed performance in 3D real-time human reconstruction system. The increased speed can be then exchanged to improve the representation of local details.

In this paper, we investigate a real-time 3D human body reconstruction scheme by a single RGBD camera, based on parametric methodology. The scheme can reconstruct rich local details in more reasonable states with real-time speed. The result shows that the additional depth information is proven to accelerate the reconstruction computation. The contributions of our work include the following aspects:

- We build a deep neural network called FBN to predict parameters from RGBD data end-to-end. FBN pays more attention to details and reduces the unreasonable states of human body. To train the network, we use Azure Kinect camera to collect the RGBD data of five actors and build an Indoor-Human dataset.
- We propose a lightweight 3D real-time human body reconstruction system based on parametric methodology with a single RGBD camera.

We further show that FBN can achieve huge improvement in efficiency and similar accuracy, by better employing depth information, as compared to state-of-the-art parametric methods.

2 Related work

2.1 Non-parametric reconstruction methods

In non-parametric reconstruction methods, Holoportion [22] system proposed by Microsoft leads the upsurge. Holoportion used Conditional Random Field for foreground segmentation and Fusion4D [7] for voxel fusion on time series. Based on embedded deformation (ED), Fusion4D proposed an energy equation to estimate non-rigid deformation field and proposed a fusion error correction mechanism to search corresponding point cloud. These optimizations adapted Fusion4D to fast human movements and topology changes. Motion2Fusion [6] and VolumeDeform [11] proposed after Holoportation were focused on optimizing detail, texture and topology. Xu et al.[32] proposed UnstructuredFusion, whcih employed three depth cameras to reconstruct the whole body by online multi-camera calibration and skeleton warping based non-rigid tracking. UnstructuredFusion almost perfectly removed the dependence on tiresome pre-calibration, reducing the difficulty of system construction. The above methods are based on spacial model fusion, and the key point is to find an efficient calibration method or registration method. Another approach is temporal fusion. Yu et al.[34] proposed Function4d, which used dynamic sliding fusion to fuse neighboring depth observations together with topology consistency. Similar works include RobustFusion [25], POSEFusion [15] and so on.

2.2 Parametric reconstruction methods

For the specific object such as human body, with its prior knowledge, parametric method can greatly remove the dependency on computing requirements and reduce the complexity of the model representation. The process can be divided into two steps: one is to create a 3D template model of the human body, and the other is to control the deformation of the template model with several key parameters.

In recent years, SMPL(A Skinned Multi-Person Linear Model) [17]has been widely used in the field of human reconstruction. The method learned mixed shapes from massive data and described the posture as a linear combination of rotation matrices. As compared to LBS(Linear Blending Skinning), SMPL is more standard, simple, realistic and has better generalization ability. However, it lacks the representation of expression and gesture. The face and hands only occupy a small part of the body, but they transmit the most interactive information. Therefore, Pavlakos et al. [23] proposed SMPL-X model which paid more attention to local details. At the same time, SMPL-X is 8 times faster than SMPL. SMPL and SMPL-X both used low dimensional parameters as inputs to generate a high-dimensional human model. In this paper, we use SMPL-X as the template human model as we expect to improve local details.

After the human model can be controlled by parameters, the method of generating parameters needs to be designed. SMPLify and SMPLify-X are basic methods based on optimization given by the authors [17, 23]. Compared with SMPLify, SMPLify-X trained a gender detector to optimize the performance on different genders. Both methods needed to obtain the key points of human body from OpenPose [2], and then predicted the shape and pose parameters through numerical fitting. However, it took more than 40 seconds for both SMPLify and SMPLify-X to process a frame. Additionally, the two methods depend on 2D attitude and optimization-based method, which could suffer from the issues of local optimum. HMR [12] trained a network that can generate 3D human body model only by 2D annotation. The method took the back-projection of key points as the loss, to get rid of the limitation of 2D-3D matching data. Omran et al. [21] proposed NBF method, inferring SMPL parameters based on bottom-up human semantic segmentation and top-down model constraints. ExPose proposed by Choutas et al. [23] focused more on

the hand and face through multi-stage and back projection, making the details more abundant.

All current methods based on parametric methods get 3D shape and pose from 2D images. The mapping from 2D to 3D is such a complex process that it requires a great deal of computation. In this paper, we use the depth information directly to accelerate the 3D reconstruction process accurately.

3 System overview

Our reconstruction system is shown in Fig. 1. The entire system consists of three main parts: data acquisition and pre-processing, real-time 3D reconstruction and texture mapping.

3.1 Data acquisition and pre-processing

In terms of sensing equipment, we adopt a depth camera that can output depth data together with RGB data. There are two advantages as follows.

First, depth camera can obtain depth information directly. Indirect methods using RGB cameras require complex algorithms or deep neural networks, which usually consume plenty of computing resources. If the depth data can be obtained directly, the prediction process will naturally be greatly simplified.

Second, the depth data is more robust. Due to uneven illumination, the quality of RGB images is easy to decline. Subsequently, the image contrast will bias the recognition of the human contour. Depth data can help to reduce these effects. ToF method used by Kinect and structured light method used by RealSense are not insensitive to illumination, for they both adopt infrared source to actively measure distance. Advantageously, the reconstruction algorithm using the depth camera can be more robust.

Combining the above two advantages, we chose the depth camera as our sensing equipment. Specifically, we select the Azure Kinect DK camera which uses the ToF method. Compared with the RealSense D435 camera using structured light, it has longer effective distance and lower sensitivity to the lighting environment.

The depth image output by the camera has non-uniform response noise, random noise and fixed noise, which is disadvantageous to subsequent steps. To reduce the noise, raw outputs are buffered and pre-processed before entering to the network. In the buffering step, depth images are smoothed temporally. Raw data is sequentially entered into the buffer queue and the average value is output in turn. In data pre-processing step, we use Gaussian filter to filter out the background noise. After that, we use Poisson filter to smooth the local concave convex of the depth image.

3.2 Real-time 3D reconstruction

For the representation of human model, we use the parametric representation method. Specifically, we choose SMPL-X.

To predict parameters from RGBD data, we build a deep-learning network, Fast Body Net (FBN), achieving real-time results (Fig. 2). However, for such a high-dimensional regression problem, it's difficult to control all the parameters in a balanced way, especially for the hand and face. Certain body parts occupy only a small part of the pixels but are decisive for the expression of human model. FBN adopts a parallel multi-branch structure, which makes the model pay more attention to faces and hands while maintaining the original resolution.

3.3 Texture mapping

Textures can be mapped to 3D model in two ways: per-vertex and per-face. Per-vertex method obtains the face

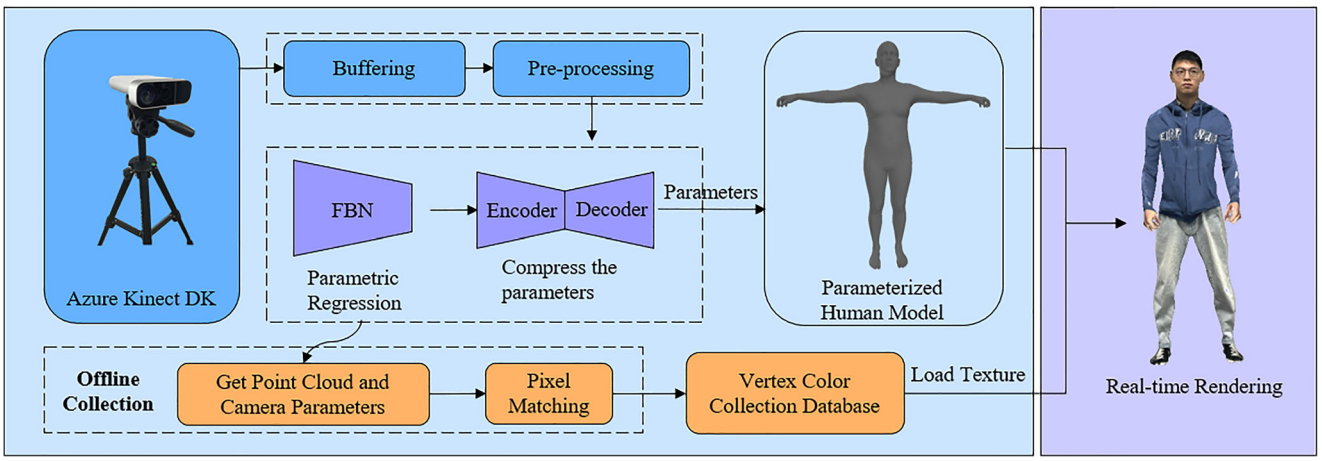

Fig. 1 Real-time 3D human body reconstruction system

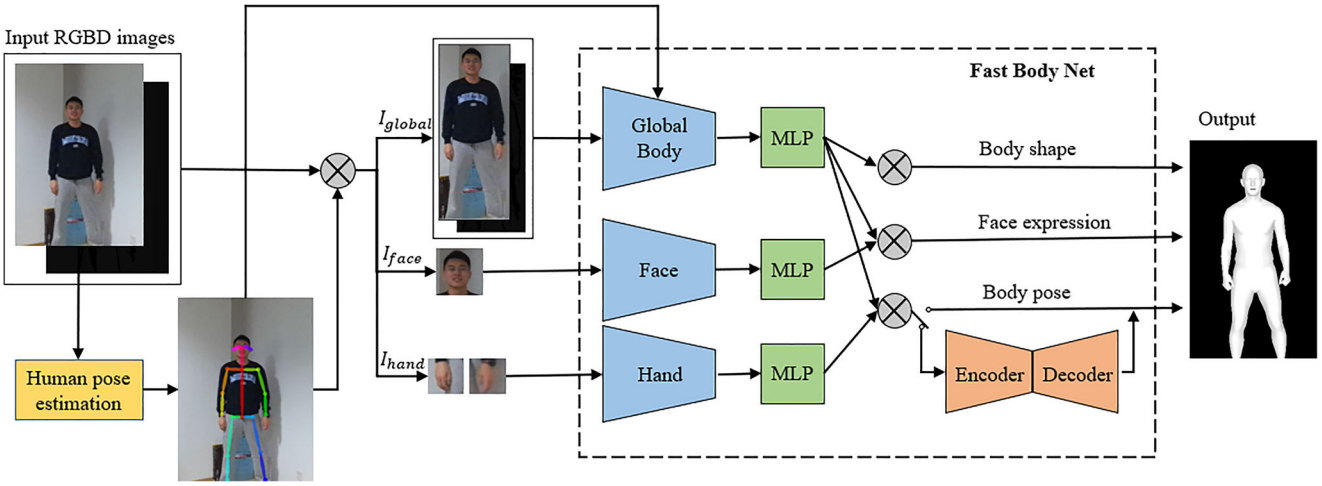

Fig. 2 Given a single RGBD image, we get the pose of joint points and use the result to extract the face and hand separately. Then we send them to the network parallel to the global body network. Each parallel network is weighted and fed into SMPL-X model. For the pose parameter, we can choose to use an automatic encoder to reduce the unreasonable output

color by interpolating and the mesh color resolution is equal to the vertex resolution. Per-face method obtains the face color directly from the texture and the mesh color resolution is equal to the texture resolution. According to SMPL-X [23], the vertex index remains unchanged during model deformation while the face index does not. Assuming that colors can be stored according to the vertex index, the real-time texture can be realized only by loading a color sequence. Therefore, we adopt per-vertex texture mapping method and correspondingly pre-establish color sequences according to the vertex index and texture image.

Color sequences depend on texture images which need to be collected offline. The input RGBD data can be taken from multiple angles. However, there will be discontinuous textures in the overlapping areas. In order to make the texture of the model smoother, we only collect input data from the front and back. The complete steps are as follows:

1. Collect RGBD data of the target human from the front and back. The resolution is recommended to be greater than 1920x1080.
2. Predict the 3D point cloud and camera parameters by FBN and SMPL-X.
3. Calculate the projection transformation and generate texture image by pixel matching.

4 Fast body net

4.1 Network architecture

We use SMPL-X model for human body representation, which includes shape $\beta \in \mathbb{R}^{10}$, expression $\psi \in \mathbb{R}^{10}$, and pose $\theta \in \mathbb{R}^{99}$. The pose parameter θ includes hand pose $\theta_h \in \mathbb{R}^{24}$, body pose $\theta_b \in \mathbb{R}^{21\times3}$, jaw pose $\theta_j \in \mathbb{R}^{1\times3}$, eye pose $\theta_e \in \mathbb{R}^{2\times3}$ and global orient $\theta_g \in \mathbb{R}^{1\times3}$. The hand pose parameter θ_h is reduced by PCA, otherwise it will be $\theta_h \in \mathbb{R}^{30\times3}$. SMPL-X generates vertices v and faces f as (1), and then a 3D human mesh m can be obtained by (2).

$$[v,\ f] = M(\beta, \theta, \psi) \tag{1}$$

$$m = F(v, f),\ v \in \mathbb{R}^{10475} \tag{2}$$

In order to enrich the details, we focus the attention of the network on the face and hand. ExPose [4] first predicted a rough human model, and then back projected the model onto the original image to get the local pixels of the face and hand. We take a different approach that FBN directly uses OpenPose [1] to get the local parts of human body. According to the output of OpenPose $O \in \mathbb{R}^{3\times25}$, we calculate the boundary points $x_{max}, x_{min}, y_{max}, y_{min}$ in the set of human joint points. The center c and size s of the bounding box can be computed as:

$$c = \left(\frac{x_{min} + x_{max}}{2}, \frac{y_{min} + y_{max}}{2}\right),\ (x, y) \in O \tag{3}$$

$$s = \gamma\,(x_{max} - x_{min}, y_{max} - y_{min}),\ (x, y) \in O \tag{4}$$

where γ represents a magnification factor. We use the calculated bounding box to make an affine transformation $T_p(c, s)$, $p \in [global, face, hand]$ to clip the original RGB image I and depth image D. The same operation is applied to the face node and hand node. Finally, we get three RGB images I_p, $p \in [global, face, hand]$ and three depth images D_p, $p \in [global, face, hand]$ containing human body, face and hand respectively.

$$I_p = ST\left[I; T_p(c_p, s_p)\right], D_p = ST\left[D; T_p(c_p, s_p)\right],\quad p \in [global, face, hand] \tag{5}$$

where ST represents spatial transformers. Compared with ExPose, we focus on the face and hands faster without back projection. Moreover, the network describing details is rarely coupled to the overall network. The quality of local details does not depend on the output of the first step, which is more robust. FBN adopts parallel multi branch structure that global sub-network uses Resnet50 and the other two use Resnet18. We extract features $\varphi_{global} \in \mathbb{R}^{2048}$ from I_{global}, D_{global} and O. At the same time, we extract features $\varphi_{face} \in \mathbb{R}^{512}$ from I_{face} and $\varphi_{hand} \in \mathbb{R}^{512}$ from I_{hand}. After that, FBN uses a MLP containing 3 layers to predict all parameters.

The output of the three sub networks is weighted and fused to get the final output. The loss can be calculated by a combination of parameters loss L_{params}, joint loss L_{joint} and re-projection loss $L_{re-project}$. The equations are as follows:

$$L = L_{params} + L_{joint} + L_{re-project} \tag{6}$$

$$\begin{aligned} L_{params} &= L_{shape} + L_{pose} + L_{expression} \\ &= \|\{\hat{\beta}, \hat{\theta}, \hat{\psi}\} - \{\beta, \theta, \psi\}\|_2^2 \end{aligned} \tag{7}$$

$$L_{\text{joint}} = \sum_{j=1}^{J} \left\| \hat{X}_j - X_j \right\| \tag{8}$$

where X_j represents the 2D joint position. All capped variables represent ground-truth quantities. After the three sub networks are trained, we stop the gradient propagation and output the weighted results. The final output is:

$$[\hat{\beta},\ \hat{\theta},\ \hat{\psi}] = [\hat{\beta},\ \lambda_g\hat{\theta}_g + \lambda_h\theta_h,\ \zeta_g\hat{\psi}_g + \zeta_f\psi_f] \tag{9}$$

where λ denotes the pose weight between global sub-network and hand sub-network, ζ denotes the expression weight between global sub-network and face sub-network.

The whole pipeline is implemented through Pytorch.

4.2 Implementation details

4.2.1 Datasets

The existing 3D human body model datasets do not contain depth image. Therefore, we establish Indoor-Human dataset with 2500 frames of RGBD images. Indoor-Human uses Azure Kinect DK camera as sensor equipment to collect raw data. The depth image obtained by Azure Kinect DK is more complete and smoother than Kinect2. In order to verify the robustness of our model in indoor scenes, we also adjust the indoor light intensity and collect a few samples. Instead of providing raw data and camera parameters, we provide aligned depth images and RGB images. After pre-processing, the dataset can be used more widely and conveniently. Our labels are obtained by running SMPLify-X , and some unreasonable data trapped in local optima are removed manually.

In addition, we use depth data to expand the dataset. We first project the depth map into a point cloud in 3D space, and then rotate it by plus or minus 45 degrees around the three axes of X, Y, Z to obtain six side views from the point cloud. In fact, we can rotate randomly from 0 to 90 degrees, but the small angle keeps almost the same information as the original data, while the large angle is completely out of the view range.

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

4.2.2 Compression of the parameters

Based on the network, FBN can generate SMPL-X parameters from RGBD data. However, due to occlusion, the whole body will be in some unreasonable states, such as palm valgus and elbow joint inward. As shown in the Fig. 3, when hands appear almost flat, the algorithm cannot recover the true posture. We find this problem in most of the datasets like 3DPW [18] and MPI-INF-3DHP [20]. To make the states of the body more reasonable, we train a denoising auto-encoder and put it at the end of FBN. In this paper, the samples with unreasonable states are regarded as noise samples and the clean dataset is obtained through manual intervention. Although the encoder will inevitably lose some accuracy, it filters out unreasonable states and makes the human reconstruction more realistic. Besides, reducing the number of parameters is also beneficial in scenarios where parameter transmission is required.

5 Results

5.1 Qualitative and quantitative analysis of FBN

5.1.1 Evaluation dataset and evaluation metrics

The existing datasets do not contain depth data, so we cannot apply FBN to them. Therefore, we only use Indoor-Human dataset for comparison.

We have adopted two indicators to measure accuracy: Mean Per-Joint Position Error (MPJPE) and Vertex-to-Vertex (V2V). To compare the computing speed, we calculate the average running time under the same configuration and environment. It is worth noting that running time here refers to all the time from inputting the original data to rendering the final human body model, including pre-processing, pose estimation and other necessary processes.

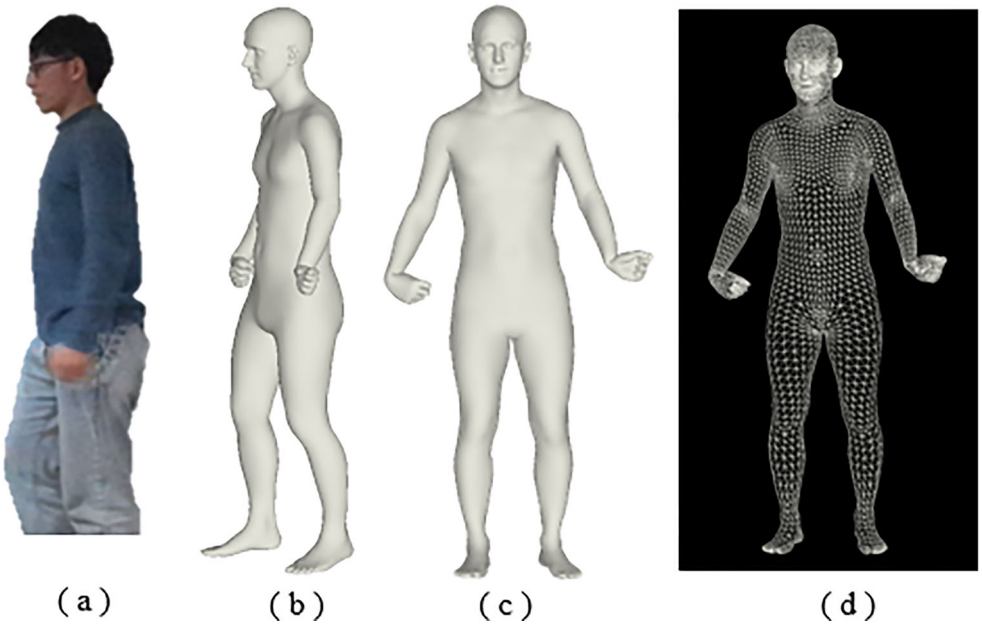

Fig. 3 Unreasonable states of reconstruction results. (a) is the input image. (b), (c), (d) are different perspectives of the model reconstructed by ExPose

Additionally, to compare with methods using SMPL model like HMR and SPIN, we retrain a FBN version based on SMPL. Secondly, the output parameters of ExPose are not standard. It used the buildlayer function to build the model which has no dimensionality reduction. We transform the output to make a fair comparison.

All our experimental environments are: i79700k CPU, RTX2080ti GPU, 32G RAM.

5.1.2 Experimental results and analysis

The reconstruction results of FBN are shown in the Fig. 4. We compare HMR, SPIN, ExPose, and FBN on Indoor-Human dataset.

Quantitative result are given in Table 1. SPIN, ExPose and FBN have similar performance in MPJPE and V2V. However, HMR and SPIN do not concern about face or hand parts. So, when considering face and hand details, ExPose and FBN have higher reconstruction accuracy. In terms of reconstruction time, FBN and SPIN take the least time. They are 57% faster than ExPose if Pyrender is used as the render and 63% faster than ExPose if Open3D is used as the render. When considering face and head details, FBN has the fastest reconstruction speed.

From the Fig. 5, we can also make a qualitative analysis. Compared with SPIN, ExPose fits better in body proportion and body size. But neither of them can express the details of the hand. FBN performs better for the local details than the previous two methods, while maintaining the advantages of ExPose.

After proving that FBN can be effectively applied to 3D human reconstruction, we further conduct ablation experiments.

First, we evaluate whether sub-networks are effective for reconstructing the details of human body. As shown in Table 2, a single global network has the worst performance while parallel sub-networks with or without clipping have better performance. Additionally, if we use the result of OpenPose as a mask to clip the hand and face parts, the details are slightly improved. Resnet18 as feature extractor in the sub-network is not so deep that clipping will help it more intensively focus on local parts.

Another ablation experiment is about the denoising auto-encode added at the end of the network. As shown in Table 3, the accuracy is reduced a little if the automatic encoder is used. This is due to the correction of the unreasonable body states. On the other hand, the results show that the total size of the parameters has decreased by 35.7%. The reduction is beneficial in scenarios where parameter transmission is required.

5.2 Texture acquisition and texture mapping

As shown in Fig. 6, the texture we generate is of great fidelity despite that some small parts outside the camera's view are lost. This can be attributed to the method that we get the depth data directly from the depth camera rather than indirectly calculated from RGB images. This method is consistent with the whole reconstruction system.

In Fig. 7, we show the reconstruction results of different actors' textures with different shapes and poses.

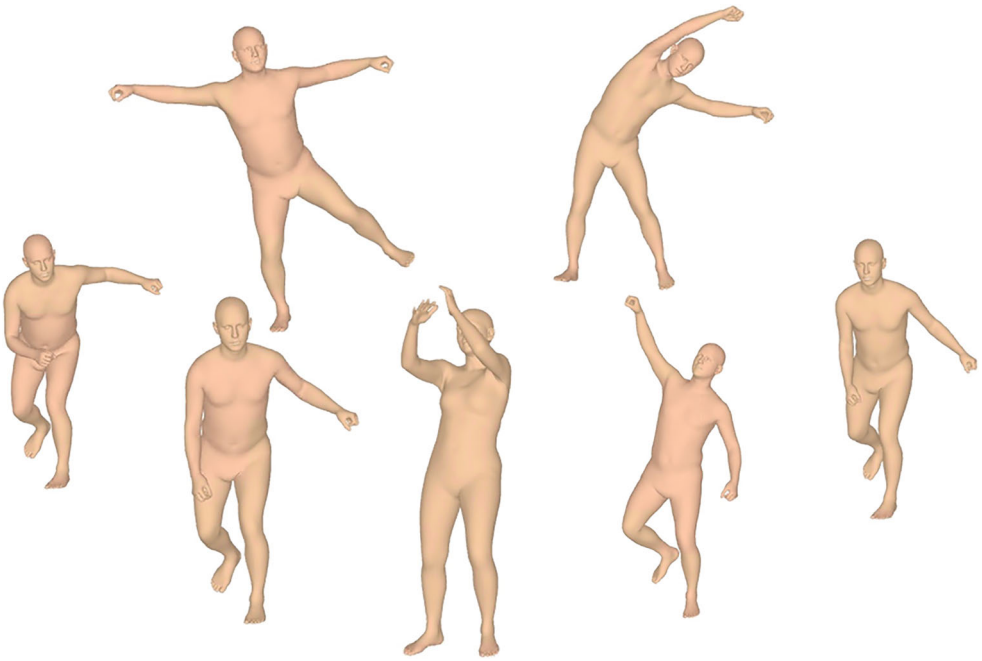

Fig. 4 Parameter prediction of human model and real-time reconstruction results

Table 1 Comparison of algorithms in Indoor-Human dataset

	MPJPE (mm)	V2V (mm)	Time with Pyrender (s)	Time with Open3D (s)
HMR[12]	80.1	119.2	0.28	0.14
SPIN[13]	59.2	110.8	0.21	0.07
ExPose[4]	61.5	96.5	0.33	0.19
FBN (Ours)	60.3	97.8	0.21	0.07

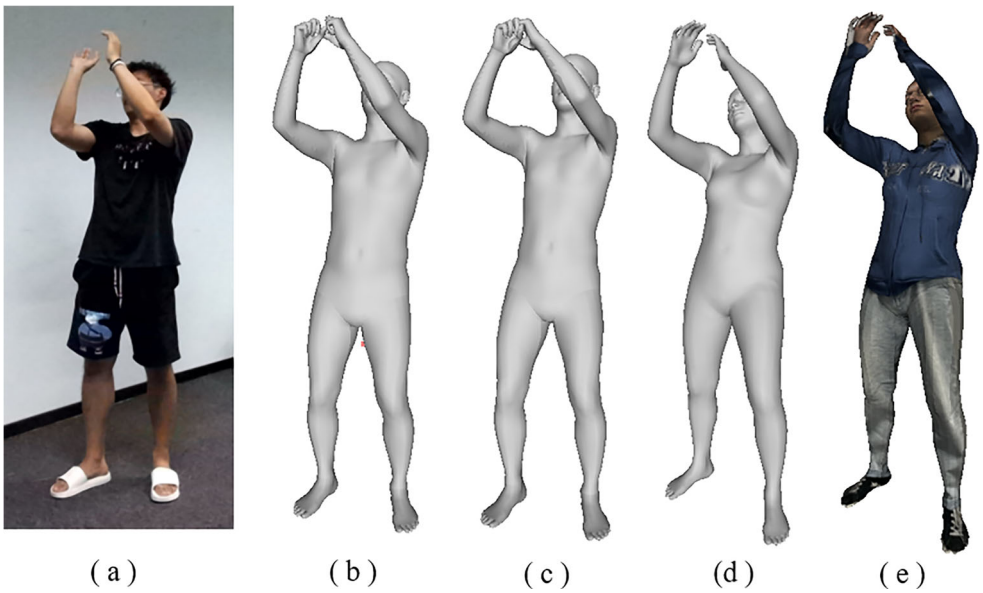

Fig. 5 Performance comparison of different methods. (a) is the input. (b), (c), (d) and (e) are Results of SPIN, ExPose, FBN and textured FBN

Table 2 Comparison made with parallel sub-networks and cropped face and hand

	MPJPE (mm)	V2V (mm)	Time with Pyrender (s)	Time with Open3D (s)
Not parallel	89.1	130.2	0.19	0.05
Parallel only	62.5	103.8	0.21	0.07
Parallel & clip	60.3	97.8	0.21	0.07

Table 3 Comparison made with compressed parameters

	MPJPE (mm)	V2V (mm)	Size of parameters (KB/frame)
With compression	62.0	106.1	2.92
Without compression	60.3	97.8	4.54

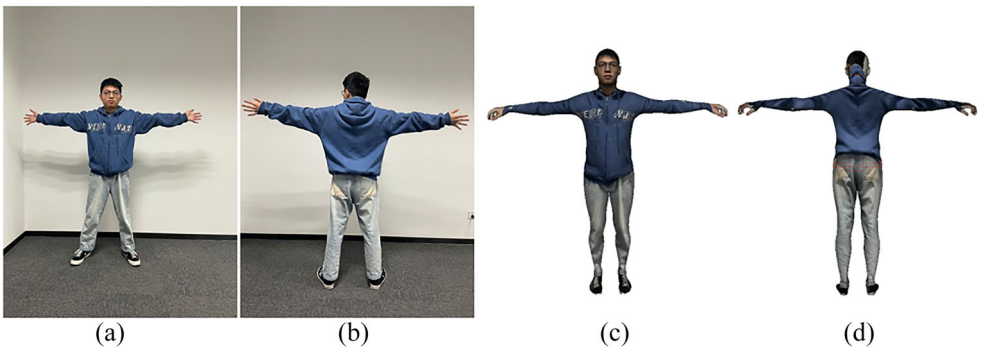

Fig. 6 Texture acquisition and texture mapping. (a) and (b) are inputs. (c) and (d) map textures to the human model

Fig. 7 Reconstruction results of different actions and textures

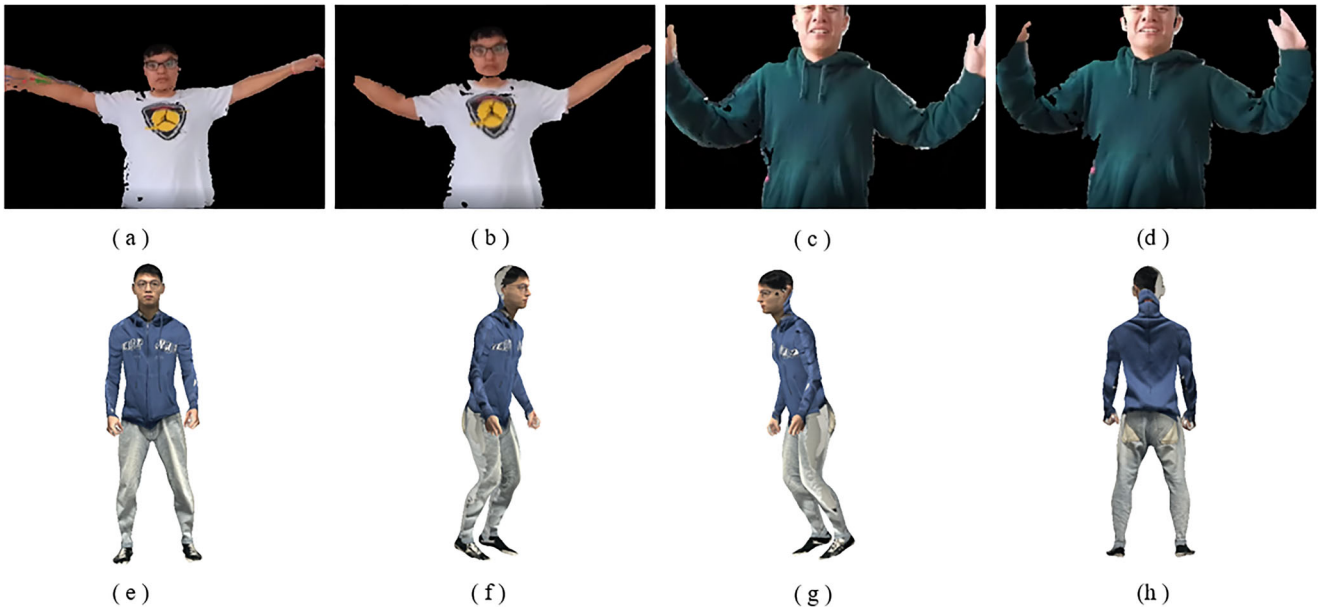

Fig. 8 The reconstruction results of parametric and non-parametric methods are compared. (a), (b), (c) and (d) are the reconstruction results of SurfelWarp [10]. (e), (f), (g) and (h) are the reconstruction results of our method

5.3 Comparison with non-parametric methods using RGBD images

We compare our method with SurfelWarp [10], a non-parametric human reconstruction method using a single RGBD camera as the sensor equipment. The comparison result is shown in Fig. 8. SurfelWarp has high fidelity, but the point cloud is rough. Besides, it can only reconstruct the parts in the camera's view. In contrast, our system can robustly reconstruct an unbroken human body with smooth skin even if some parts of the body are out of camera's view. In terms of efficiency, our system takes only 0.21 seconds to run a frame, while SurfelWarp takes more than 1 second. In addition, our system can deploy FBN on the server side and only put the rendering part on the client side. This CS architecture makes client only need very few computing resources. However, SurfelWarp is difficult to separate the architecture due to direct process of point cloud.

5.4 Integration with static scenes

The above works realize the real-time reconstruction with a single RGBD camera. To be more realistic, the 3D human models can be integrated with static scene models like laboratory or bedroom. We keep the mesh of the static scene unchanged, and dynamically update human mesh nodes by identification. The renderer we use in this work is Pyrender, which is a pure Python library for physically-based rendering and visualization.

We use BundleFusion [5] to reconstruct the house. Results of the fusion is shown in Fig. 9.

Fig. 9 Dynamically changing human model placed in a static scene

6 Conclusion

In this paper, we investigate a real-time 3D human body reconstruction scheme by a single RGBD camera, based on parametric methodology. Under the condition of reconstructing hand and face details, our method achieves the fastest reconstruction speed. The use of the depth camera allows us to take 3D information directly and the additional depth information is proven to accelerate the reconstruction computation. In the future, we hope to improve accuracy and efficiency by using higher dimensional semantic information between frames.

The real-time 3D representation of non-rigid objects such as humans and animals is considered to be one of the most difficult tasks. Such a lightweight real-time reconstruction system of human body will be of great social and commercial value. As such, in the future, people can shuttle freely between the real world and the virtual world, e.g., study, work, make friends, shop and travel in the metaverse.

Acknowledgements This work is sponsored by Shanghai Key Research Laboratory of NSAI.

References

1. Cao Z, Hidalgo G, Simon T, Wei SE, Sheikh Y (2019) Openpose: realtime multi-person 2d pose estimation using part affinity fields. IEEE Trans Pattern Anal Mach Intell 43(1):172–186
2. Cao Z, Simon T, Wei SE, Sheikh Y (2017) Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 1302–1310 Honolulu. https://doi.org/10.1109/CVPR.2017.143. http://ieeexplore.ieee.org/document/8099626/
3. Choi H, Moon G, Chang JY, Lee KM (2021) Beyond static features for temporally consistent 3d human pose and shape from a video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1964–1973
4. Choutas V, Pavlakos G, Bolkart T, Tzionas D, Black MJ (2020) Monocular expressive body regression through body-driven attention. In: European conference on computer vision. Springer, pp 20–40
5. Dai A, Nießner M, Zollhö fer M, Izadi S, Theobalt C (2017) Bundlefusion: real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. ACM Trans Graph (ToG) 36(4):1
6. Dou M, Davidson P, Fanello SR, Khamis S, Kowdle A, Rhemann C, Tankovich V, Izadi S (2017) Motion2fusion: real-time volumetric performance capture. ACM Trans Graph 36(6):1–16. https://doi.org/10.1145/3130800.3130801
7. Dou M, Khamis S, Degtyarev Y, Davidson P, Fanello SR, Kowdle A, Escolano SO, Rhemann C, Kim D, Taylor J, Kohli P, Tankovich V, Izadi S (2016) Fusion4D: real-time performance capture of challenging scenes. ACM Trans Grap 35(4):1–13. https://doi.org/10.1145/2897824.2925969
8. Fang Q, Shuai Q, Dong J, Bao H, Zhou X (2021) Reconstructing 3d human pose by watching humans in the mirror. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12814–12823
9. Gabeur V, Franco JS, Martin X, Schmid C, Rogez G (2019) Moulding humans: non-parametric 3d human shape estimation from single images. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 2232–2241
10. Gao W, Tedrake R (2019) Surfelwarp: Efficient non-volumetric single view dynamic reconstruction. arXiv:1904.13073
11. Innmann M, Zollhöfer M, Nießner M, Theobalt C, Stamminger M (2016). In: Leibe B., Matas J., Sebe N., Welling M. (eds) VolumeDeform: Real-Time Volumetric Non-rigid Reconstruction, vol 9912. Springer International Publishing, Cham, pp 362–379. DOI10.1007/978-3-319-46484-8_22 Series Title: Lecture Notes in Computer Science
12. Kanazawa A, Black MJ, Jacobs DW, Malik J (2018) End-to-end recovery of human shape and pose. In: 2018 IEEE/CVF conference on computer vision and pattern recognition, pp 7122–7131 Salt Lake City. https://doi.org/10.1109/CVPR.2018.00744. https://ieeexplore.ieee.org/document/8578842/
13. Kolotouros N, Pavlakos G, Black MJ, Daniilidis K (2019) Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: Proceedings of the IEEE/CVF International conference on computer vision, pp 2252–2261
14. Li J, Xu C, Chen Z, Bian S, Yang L, Lu C (2021) Hybrik: a hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3383–3393
15. Li Z, Yu T, Zheng Z, Guo K, Liu Y (2021) Posefusion: Pose-guided selective fusion for single-view human volumetric capture. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 14162–14172
16. Lin K, Wang L, Liu Z (2021) End-to-end human pose and mesh reconstruction with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1954–1963
17. Loper M, Mahmood N, Romero J, Pons-Moll G, Black MJ (2015) SMPL: a skinned multi-person linear model. ACM Trans Graph 34(6):1–16. https://doi.org/10.1145/2816795.2818013
18. von Marcard T, Henschel R, Black MJ, Rosenhahn B, Pons-Moll G (2018) Recovering accurate 3d human pose in the wild using imus and a moving camera. In: Proceedings of the European conference on computer vision (ECCV), pp 601–617
19. May D, Auer M (2021) Cross reality and data science in engineering - proceedings of the 17th international conference on remote engineering and virtual instrumentation. https://doi.org/10.1007/978-3-030-52575-0
20. Mehta D, Rhodin H, Casas D, Fua P, Sotnychenko O, Xu W, Theobalt C (2017) Monocular 3d human pose estimation in the wild using improved cnn supervision. In: 2017 International conference on 3d vision (3DV). IEEE, pp‘506–516
21. Omran M, Lassner C, Pons-Moll G, Gehler P, Schiele B (2018) Neural body fitting: unifying deep learning and model based human pose and shape estimation. In: 2018 international conference on 3D vision (3DV). IEEE, pp 484–494, Verona. https://doi.org/10.1109/3DV.2018.00062. https://ieeexplore.ieee.org/document/8491000/
22. Orts-Escolano S, Rhemann C, Fanello S, Chang W, Kowdle A, Degtyarev Y, Kim D, Davidson PL, Khamis S, Dou M, Tankovich V, Loop C, Cai Q, Chou PA, Mennicken S, Valentin J, Pradeep V, Wang S, Kang SB, Kohli P, Lutchyn Y, Keskin C, Izadi S (2016) Holoportation: virtual 3D teleportation in real-time. In: Proceedings of the 29th annual symposium on user interface software and technology. ACM, pp 741–754, Tokyo. https://doi.org/10.1145/2984511.2984517
23. Pavlakos G, Choutas V, Ghorbani N, Bolkart T, Osman AA, Tzionas D, Black MJ (2019) Expressive body capture: 3d hands, face, and body from a single image. In: 2019 IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp 10967–10977, USA. https://doi.org/10.1109/CVPR.2019.01123. ttps://ieeexplore.ieee.org/document/8953319/
24. Stiegler C (2020) No line on the horizon: virtual reality in digital ecosystems and the politics of immersive storytelling. In: Handbook of research on recent developments in internet activism and political participation. IGI Global, pp 53–68
25. Su Z, Xu L, Zheng Z, Yu T, Liu Y, Fang L (2020) Robustfusion: human volumetric capture with data-driven visual cues using a rgbd camera. In: European conference on computer vision. Springer, pp 246–264
26. Vasylevska K, Kaufmann H (2017) Compressing VR: fitting large virtual environments within limited physical space. IEEE Comput Graph Appl 37(5):85–91. https://doi.org/10.1109/MCG.2017.3621226. http://ieeexplore.ieee.org/document/8047456/
27. Venkat A, Jinka SS, Sharma A (2018) Deep textured 3D reconstruction of human bodies. arXiv:1809.06547
28. Wan C, Probst T, Van Gool L, Yao A (2019) Self-supervised 3D hand pose estimation through training by fitting, p 10854. https://doi.org/10.1109/CVPR.2019.01111
29. Wang L, Zhao X, Yu T, Wang S, Liu Y (2020) Normalgan: learning detailed 3D human from a single rgb-d image. In: European conference on computer vision. Springer, pp 430–446
30. Wu F, Bao L, Chen Y, Ling Y, Song Y, Li S, Ngan K, Liu W (2019) MVF-Net: multi-view 3D face morphable model regression, p 968. https://doi.org/10.1109/CVPR.2019.00105
31. Xu H, Alldieck T, Sminchisescu C (2021) H-nerf: neural radiance fields for rendering and temporal reconstruction of humans in motion. Adv Neural Inf Process Syst, vol 34
32. Xu L, Su Z, Han L, Yu T, Liu Y, Fang L (2019) Unstructured-fusion: realtime 4d geometry and texture reconstruction using commercial rgbd cameras. IEEE Trans Pattern Anal Mach Intell 42(10):2508–2522
33. Ying L, Jiong Z, Wei S, Jingchun W, Xiaopeng G (2017) VREX: virtual reality education expansion could help to improve the class experience (VREX platform and community for VR based education), p 5. https://doi.org/10.1109/FIE.2017.8190660
34. Yu T, Zheng Z, Guo K, Liu P, Dai Q, Liu Y (2021) Function4d: real-time human volumetric capture from very sparse consumer rgbd sensors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5746–5756
35. Zhao T, Li S, Ngan KN (2018) Wu, f.: 3-d reconstruction of human body shape from a single commodity depth camera. IEEE Trans Multimedia 21(1):114–123