HAND-3D-STUDIO: A NEW MULTI-VIEW SYSTEM FOR 3D HAND RECONSTRUCTION

Zhengyi Zhao[⋆] *Tianyao Wang*[⋆] *Siyu Xia*[⋆] *Yangang Wang*[⋆†]

[⋆] School of Automation, Southeast University, Nanjing, China
[†] Shenzhen Research Institute of Southeast University, Shenzhen, China

ABSTRACT

This paper proposes a new system named as Hand-3D-Studio to capture the 3D hand pose and shape information. Our system includes 15 synchronized DSLR cameras, which can acquire high quality multi-view 4K resolution color images in a circular manner. We then introduce a 2D hand keypoints guided iterative pixel growth matching strategy for 3D reconstruction, where the 2D keypoints are obtained via convolution neural network. We find that the pre-detected 2D hand keypoints can greatly remove the matching noise, and thus improve the performance of reconstruction. After that, a non-rigid iterative closest points algorithm is performed to drive a template hand to fit the point clouds and register all the hand meshes. As a consequence, we captured more than 20K high quality hand color images, annotated 2D hand keypoints, 3D point cloud as well as the registered hand meshes (>200). All the data are public on the website `http://www.yangangwang.com` for future research.

***Index Terms*—** Multi-view, 3D Reconstruction, 3D Hand Pose Estimation, Dataset

1. INTRODUCTION

Capturing a high fidelity 3D hand model is a important problem, which has seen a significant amount of research interests due to its applications in computer graphics, animation, human computer interaction, rehabilitation and *etc*. Typically, two main strategies, that are depth sensors [1] and RGB sensors [2], can be utilized to perform this task. In recent years, researches have demonstrated that multi-mode data types (*e.g*., point cloud, color image or even multi-view color images) could improve the performance of 3D hand pose estimation. However, due to the limitations of acquisition equipment, many of the existing public hand datasets do not meet such requirements.

This work aims to build a system to collect the multi-mode data types of hands. We propose a high-precision multi-camera acquisition system, which consists of 15 high-quality DSLR cameras. A miniaturized collecting "yurt" with cameras and a led light set was specially designed for hand capture as shown in Fig. 1. The height and tilt angle of the proposed system can be adjusted up to 5.9 feet and 45 degrees respectively, which fits all the hands capture. The customized led light set minimizes specularity and makes the hand details be clear, where the monotonous texture and specularity on the hand may lead to noisy reconstruction results. It is noted that most people are unable to prevent their fingers from trembling, and even they cannot hold their hands still for seconds. In order to solve this challenge, we propose several strategies to make the synchronization delay of our system be less than 10 milliseconds.

Our system combines the advantages of the single depth sensor systems and RGB sensor systems. With the proposed system, we build a new hand dataset with high fidelity 3D hands, which consists 4K multi-view RGB images, point cloud data from stereo matching, 2D & 3D hand joints and fitted 3D hand mesh models. Our system and data would greatly improve the accuracy and fidelity of existing works to estimate the 3D hand pose. The large-scale texture data also provides more possibilities for obtaining high-precision pore-scale hand surfaces. All the multi-mode hand data are public on the website `http://www.yangangwang.com` for future research.

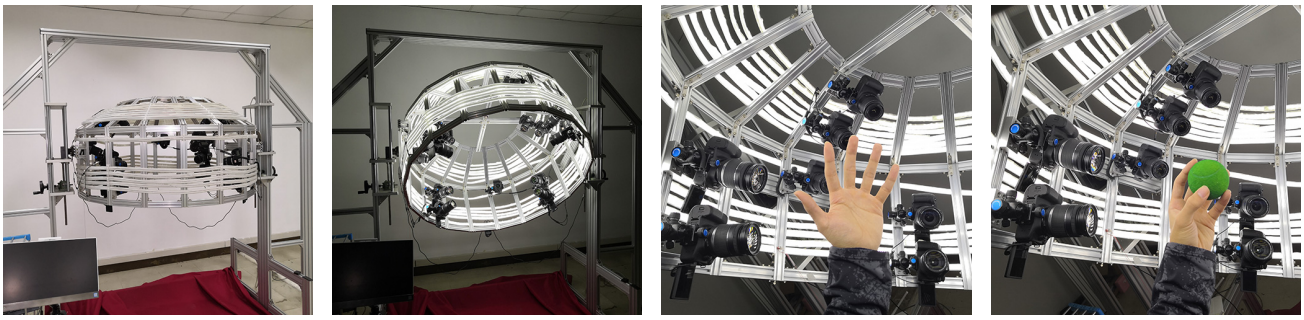
2. RELATED WORK

In order to estimate the 2D and 3D hand poses, many efforts have been made to acquire the hand datasets. Currently, public datasets can be mainly divided into three categories by the sensors: RGB, depth and mixtures. The monocular RGB data is often used for the 2D joints estimation [3, 4, 5] while multi-view systems[6, 7] perform well in the 3D pose estimation. The amount of the depth data[8, 9] is increasing with the recent progress in consumer RGB-D sensors[10]. However, these datasets lack unbiased 3D surfaces of hands. The recent work [11] tries to deploy multiple cameras to annotate 3D hand keypoints and shapes, but is still lack of accurate surface constraint in hand 3D space.

Some commercial solutions for a 3D body scan can be also used for 3D hand modeling. The 3dMDhand[12] system

This work was supported in part by the National Natural Science Foundation of China (No. 61806054), in part by the Natural Science Foundation of Jiangsu Province (No. BK20180355), in part by the Shenzhen Science and Technology Innovation Committee (STIC) (JCYJ20180306174459972) and "Zhishan Young Scholar" Program of Southeast University.

Corresponding author: Yangang Wang. E-mail: ygwangthu@gmail.com



(a) (b) (c) (d)

Fig. 1. **Acquisition setup.** This acquisition setup consists of 15 cameras, a led light set and the customized "yurt" cage. (a) Horizontal off state; (b) Tilted working state, the system is adjusted according to the size of the subjects; (c) Capturing hands only; (d) Capturing hands with an object.

synchronizes five modular units of machine vision cameras and an industrial-grade flash system in a single capture. But it requires the special equipment and only generates a model without multi-view images. TEN24[13] creates a capture rig consisting of hundreds of DSLR cameras and generates fine hand models by the multi-view stereo. But the raw scan must be cleaned up manually to fix the bulges and holes on the surface, which limits the quantity of outputs. TechMed 3D[14] uses a handheld 3D scanner to scan hands. A single scan usually lasts dozens of minutes. Therefore, it can only generate a processed smooth hand with very few details.

For synchronization, a typical multi camera control solution is Smarter Shooter[15]. The system is able to control up to 100 cameras at the same time through a USB connection. Unfortunately, the synchronous time is more than 1 second, which is too long to capture a highly-transient hand.

3. METHOD

The pipeline of the proposed method for 3D hand reconstruction is illustrated in Fig. 2. Our system has four main parts, which includes: hardware system, 2D hand pose estimation, multi-view point cloud reconstruction and non-rigid icp fitting. In the following sections, we will introduce each individual part in details.

3.1. Hardware System

Customized Holder. The system uses a customized holder for the hand capture. The core of the holder is a "yurt" cage, the height and angle of which can be adjusted depending on the height and shape human. The height range is 4.26 feet to 5.90 feet, and the angle range is 0 to 45 degrees. We use the industrial aluminum to build the holder. This material is lighter and can hold more weight. In order to adjust the position and angle of each camera, we attach the cameras on the holder by camera mounts. Soft-light baffles can be optionally mounted around on the pillars of the cage to restrain specularity and provide an enclosed environment that avoids much of the ambient light interference.

Camera Arrangement. The capture setup consists of 8 Canon EOS-80D DSLR cameras and 7 Canon EOS-1200D DSLR cameras, where eight 80Ds are divided into 4 groups and fixed about every 90 degrees inside the customized holder. The 1200Ds are fixed on the top of the holder, as illustrated in Fig. 1. The cameras at the top effectively supplement the missing parts on surfaces obtained from the 80Ds. Other camera arrangements, such as all the cameras are evenly placed in a ring, would lead to an incomplete gestures capture because of the lack of images from the top side. Another arrangement is a jagged ring with a larger baseline, which leads to plenty failed matches and numerous holes.

Synchronization. Different from the face capture, all cameras must take the pictures of the hand at the same time, because it is really difficult for people to keep their hands still for a while. We use high speed wireless remote controls to generate the shutter signal and the common $2.5mm$ audio cable to split the signal to control all the cameras. All the cameras are wired to a wireless receiver. This wired connection allows the cameras to shoot at almost the same time. The synchronization delay is less than 10 milliseconds, far less than the USB software control scheme.

Lighting. The light set also plays an important role in generating a high quality hand model. A bright illumination is necessary to avoid noises and increase the shutter speed. A total length of 50 meters LED strips are mounted on the outside of the "yurt" to provide a shadowless illumination around the hand. Light passes through the diffuser and forms diffuse light to minimize specularity.

3.2. 2D Hand Keypoints Estimation

We use the 2D hand keypoints to help the 3D hand reconstruction, especially for guiding the pixel matching among different cameras. The 2D hand keypoints are obtained through an encoder-decoder hand pose estimation network [3] other than manual annotations. The main idea of the method is to simul-

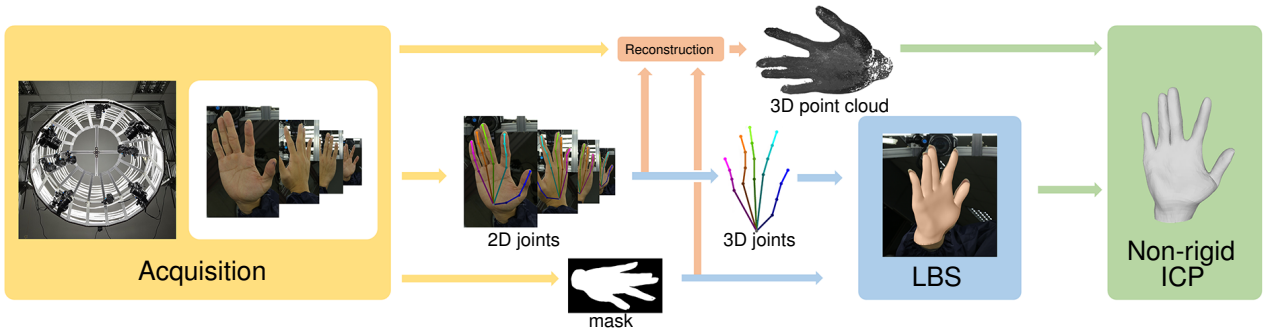


Fig. 2. **Pipeline.** Multi-view images are first acquired from our HAND-3D-STUDIO hardware system. We use an encoder-decoder network to obtain the 2D hand keypoints and masks. The 3D hand joints are then computed from 2D keypoints. We perform the multi-view reconstruction with the help of 3D hand joints. Following that, the point clouds, 3D hand joints and masks are all used to drive a template hand mesh with linear blend skinning (LBS) and non-rigid iterative closest points (ICP).

taneously get the hand region (ROI) and the joints of the hand. To improve the detection accuracy, we use the regional information as feedback to guide the neural network to re-detect the hand region and the position of the joints. We find that this strategy can sufficiently achieve good results for our the 3D hand reconstruction as shown in Fig. 4.

3.3. Multi-view Reconstruction

Because the hand skin has almost the same color, and the hand postures are variable, traditional reconstruction methods for face will output much noise. To de-noise, apart from the image pyramid and constraint methods, we propose a depth initialization method based on the estimated 2D keypoints and an iterative pixel growth method based on Visual Hull.

Depth Initialization. Inspired by the image deformation [16], we find that the two-dimensional skeleton keypoints of the hand can be viewed as the control points. The deviation of the pixels after deformation is very close to the real disparity. Based on this, we initialize the depth in the pairwise matching combined with the 2D joints location information. For a pixel p in the image $\mathcal{I}$, we compute the normalized cross-correlation value and the scaled location value along the epipolar line.

Taking joints in image $\mathcal{I}$ as control points, we deform every pixel in image $\mathcal{I}$, targeting joints in image $\mathcal{J}$ as,

$$\begin{aligned} E(q) =& \frac{1 - NCC(p,q)}{2} + K\left(1 - e^{-\|q - q_{ref}\|^2}\right) \\ K =& K_0 \frac{\min\left(\|p - joints\|^2\right)}{\min'\left(\|p - \text{joints}\|^2\right)}. \end{aligned} \tag{1}$$

where the deformed coordinates are stored in map Q_{ref}. q_{ref} is the reference coordinate of each pixel p. K makes sure the location constraints will not be too strong. The $joints$ represents all the 2D joints in image $\mathcal{I}$, $\min'$ represents the second minimum, and K_0 is a user defined value. For a 1000 pixel wide image, K_0 is 0.05. The noise greatly is suppressed in initialization step in this way.

Iterative Pixel Growth Matching. We also use smoothness constraint, uniqueness constraint and ordering constraint as [17] to further exclude noise. Continuous points on the image may be discontinuous in the 3D space. For example, the multiple fingers overlap in image but separate in space. In these cases, noises will not be eliminated. A Visual Hull[18] constraint is added to counter this problem. Partial matching results must be projected to all the masks. The values out of the volume intersection will be directly deleted. The deleted pixels are re-matched around their neighbors to grow new points. This process are implemented iteratively. Iteration times is dependent on the image resolution, *e.g.*, 20 iterations for 200 $\times$ 300.

3.4. Mesh Model Fitting

We perform the Linear Blend Skinning (LBS) algorithm and non-rigid iterative closest points (ICP) algorithm by fitting a template hand mesh to the reconstructed point cloud, as illustrated in Fig. 2. More details can be referred in [19]. Finally, unbiased registered hand meshes are obtained for the output of our system.

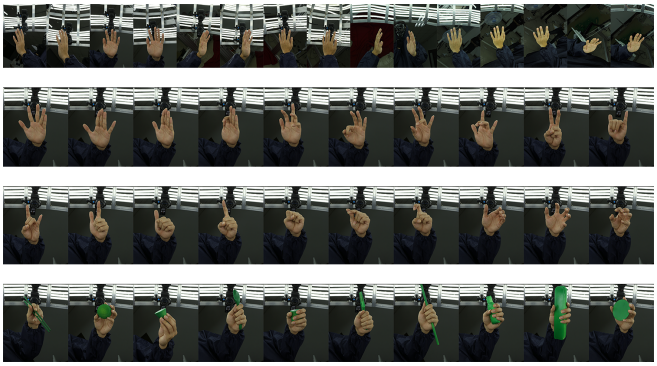


Fig. 3. **Dataset**. Top Row: 15 views of color images; Middle Two Rows: selected hand gestures; Bottom Row: selected hand-object interactions.

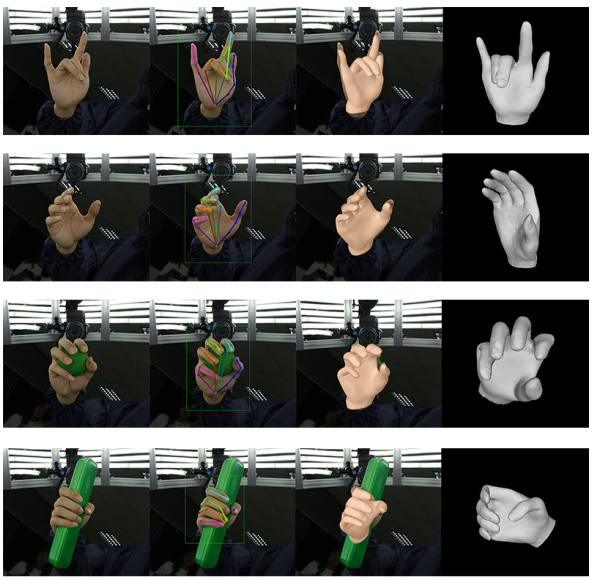
4. EXPERIMENT

Dataset. With the proposed system, we have collected multi-view hand color images with 10 persons of different genders and skin colors as shown in Fig. 3 and the summary of the dataset is shown in Tab. 1. The gestures we collected are very representative and common in daily life. For each person, we collected 50 one-handed gestures and 27 hand-object interaction gestures, both for the left and right hand. The gestures can be divided into 3 categories: finger-movements[10], common gestures and hand-object interaction gestures.

Table 1. Multi-view 4K dataset

attributes	value
num.subjects	10
num.objects	27
num.cameras	15
num.frames	22K/-
resolution	**4K**
3D joints annotation	✓
3D shape annotation	✓

Furthermore, we computed the annotations automatically for the captured multi-view hand color images, which includes: 2D & 3D hand joints, 3D point cloud and registered hand mesh model. Particularly, we detected the multi-view 2D joints with a convolution neural network and computed the 3D hand joints with camera parameters. Combined with a template hand mesh, we fitted a LBS model and deformed it to register 3D point clouds by nonrigid ICP. Fig. 4 shows a small portion of our multi-mode results. We hope the multi-view hand color images dataset as well as the computed annotations will promote the research for high-fidelity hand reconstruction and hand pose estimation in the future.



(a) (b) (c) (d)

Fig. 4. (a): Selected color images; (b): 2D hand keypoints; (c): Fitted hand mesh models overlaid on color images; (d): Fitted hand mesh models.

(a) (b) (c)

Fig. 5. Comparison with different calibration patterns. (a) Random pattern with color calibration; (b) Multi-camera calibration with a ball; (c) A cylinder with random patterns.

Beyond that, we also conducted several experiments to demonstrate the performance of our hardware system. Typically, we found that camera parameters are crucial for the accurate hand reconstruction.

Calibration. Since it is difficult to extract consistent features on hand surface, all the cameras must be calibrated before capturing instead of the camera self-calibration. We calibrated the cameras using intensive grayscale features and color features. Specifically, we made a wooden geometry full of the features using a random pattern from Li and Heng [20]. The object's size should be limited to the scan volume. Blurred images of a too big object, shown in Fig. 5(c), leads to mistakes in calibration. We also tried another multi-camera calibration in [17] using a ball, and found that its robustness and accuracy were not as good as the adopted method. Tab. 2 shows the comparison results of different calibration patterns.

Table 2. Comparison with different calibration patterns

Calibration Patterns	reprojection error	pairwise matches	robustness
(a)	<**0.3 px**	>300	✓
(b)	<1.0 px	<40	✗
(c)	<3.0 px	>**400**	✗

5. CONCLUSION

In this paper, we present a new multi-view 3D hand reconstruction system, named as HAND-3D-STUDIO. The main contribution of this paper is that we public a novel multi-mode hand dataset including multi-view color images, 2D hand keypoints, 3D hand skeletal joints and registered hand mesh models. The proposed dataset would be beneficial for high fidelity hand reconstruction and hand pose estimation from single RGB images in the future.

6. REFERENCES

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

[2] Christian Zimmermann and Thomas Brox, "Learning to estimate 3d hand pose from single rgb images," in *ICCV*, 2017.

[3] Yangang Wang, Cong Peng, and Yebin Liu, "Mask-pose cascaded cnn for 2d hand pose estimation from single color image," *IEEE Trans. CSVT*, vol. 29, no. 11, pp. 3258–3268, 2019.

[4] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt, "Ganerated hands for real-time 3d hand tracking from monocular rgb," in *CVPR*, 2018.

[5] Yangang Wang, Baowen Zhang, and Cong Peng, "Srhandnet: Real-time 2d hand pose estimation with simultaneous region localization," *IEEE Trans. IP*, vol. 29, no. 1, pp. 2977–2986, 2020.

[6] Luca Ballan, Aparna Taneja, Jürgen Gall, Luc Van Gool, and Marc Pollefeys, "Motion capture of hands in action using discriminative salient points," in *ECCV*. Springer, 2012.

[7] Francisco Gomez-Donoso, Sergio Orts-Escolano, and Miguel Cazorla, "Large-scale multiview 3d hand pose dataset," *Image and Vision Computing*, vol. 81, no. 1, pp. 25–33, 2019.

[8] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt, "Real-time hand tracking under occlusion from an egocentric rgb-d sensor," in *ICCV*, 2017.

[9] Toby Sharp, Cem Keskin, Duncan Robertson, Jonathan Taylor, Jamie Shotton, David Kim, Christoph Rhemann, Ido Leichter, Alon Vinnikov, Yichen Wei, et al., "Accurate, robust, and flexible real-time hand tracking," in *HFCS*. ACM, 2015.

[10] Javier Romero, Dimitrios Tzionas, and Michael J Black, "Embodied hands: Modeling and capturing hands and bodies together," *ACM TOG*, vol. 36, no. 6, pp. 245, 2017.

[11] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox, "Freihand: A dataset for markerless capture of hand pose and shape from single rgb images," 2019.

[12] 3dMD, "Static 3dmdhand system," `http://www.3dmd.com/static-3dmd_systems/3dmdhandfoot-system/`, [Accessed August 5, 2019].

[13] TEN24, "T170 capture stage," `https://ten24.info/3d-scanning/`, [Accessed August 5, 2019].

[14] TechMed3D, "Bodyscan scanner," `https://techmed3d.com/products/bodyscan-scanner/`, [Accessed August 5, 2019].

[15] KUVACODE, "Capturegrid," `https://kuvacode.com/`, [Accessed August 5, 2019].

[16] Scott Schaefer, Travis McPhail, and Joe Warren, "Image deformation using moving least squares," *ACM TOG*, vol. 25, no. 3, pp. 533–540, 2006.

[17] Thabo Beeler, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross, "High-quality single-shot capture of facial geometry," *ACM TOG*, vol. 29, no. 4, pp. 40, 2010.

[18] Aldo Laurentini, "The visual hull concept for silhouette-based image understanding," *IEEE Trans. PAMI*, vol. 16, no. 2, pp. 150–162, 1994.

[19] Hao Li, Robert W Sumner, and Mark Pauly, "Global correspondence optimization for non-rigid registration of depth scans," *Computer graphics forum*, vol. 27, no. 5, pp. 1421–1430, 2008.

[20] Bo Li, Lionel Heng, Kevin Koser, and Marc Pollefeys, "A multiple-camera system calibration toolbox using a feature descriptor-based calibration pattern," in *IROS*. IEEE, 2013.