A Review Paper on Real-Time Hand Motion Capture

Niji K Raj[1]**, Prof. Parvathi V. S.**[2]

[1]PG Student, Dept. of Electronics & Communication Engineering, LBSITW, Kerala, India
[2]Assistant Professor, Dept. of Electronics & Communication Engineering, LBSITW, Kerala, India

---***---

Abstract - *Hand Image Understanding (HIU) can be used for a variety of human-computer interaction applications such as physical or gesture-based controls for virtual reality and augmented reality devices. It is a novel framework that extracts comprehensive information about hand objects from a single RGB image. This paper gives a brief study of the design of a hand motion capture system for real-time extraction of hand shape and poses, of all data modalities, including synthetic and real-image datasets with either 2D or 3D annotations.*

Key Words**: Hand Image Understanding, Human-Computer Interaction, Virtual Reality (VR), Augmented Reality (AR), Motion Capture, RGB Image.**

1. INTRODUCTION

The hand is our most useful tool for manipulating physical objects and communicating with the outside world. As a result, capturing vision-based 3D hand motion has a wide range of applications, including gaming, biomechanical analysis, robotics, human-computer interaction such as augmented reality and virtual reality (AR/VR), and many others. Despite years of research, it remains an unsolved problem due to the high dimensionality of hand shape, pose and shape variations, self-occultations, and so on. Previous methods concentrated on sparse 3D hand joint location from monocular RGB images. However, discriminative methods based on convolutional neural networks (CNNs) have been used to estimate dense hand poses including 3D shapes from RGB images or depth maps (Fig.1). [1]

Convolutional neural networks (CNNs)-based discriminative methods have demonstrated very promising performance in estimating 3D hand poses from RGB images or depth maps. However, the predictions are often based on coarse skeletal representations with no explicit kinematics or geometric mesh constraints. Establishing a personalized hand model, on the other hand, necessitates a generative approach that optimizes the hand model to fit 2D images. Aside from their complexity, optimization-based methods are prone to local minima, and personalized hand model calibration contradicts the ability to generalize for hand shape variations.

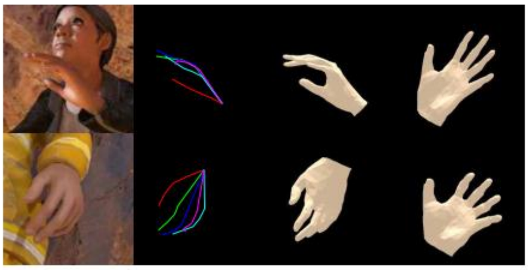

Fig -1: Hand pose examples

2. REVIEW ON RELATED WORK

Moeslund *et al.* [1] proposed a novel approach to recovering and tracking a human hand's 3D position, orientation, and full articulation from markerless visual observations obtained by a Kinect sensor. It provides a thorough review covering the general problem of articulated objects for tracking 3D, visual human motion capture, and analysis. The 3D tracking of human hands has several applications, towards developing an effective solution, one has to struggle with the number of interacting factors such as the problem of high dimensionality, and self-occultations that occur while the hand is in action. Therefore, approached an optimization problem, looking for hand model parameters that minimize the difference in the appearance and 3D structure of hypothesized instances of a hand model and actual hand observations. This optimization problem is effectively solved using a Particle Swarm Optimization variant (PSO). The proposed method does not necessitate the use of special markers or a complex image acquisition setup. It provides continuous solutions to the problem of tracking hand articulations because it is model-based. Extensive experiments with the proposed method's prototype GPU-based implementation show that accurate and robust 3D tracking of hand articulations can be achieved in near-real-time (15Hz).

Erol *et al.* [2] proposed a method based on the output, that differentiates between partial pose estimation and full pose estimation, different existing approaches are divided into model and appearance-based. Model-based approaches, provide visual information through a multicamera system but are computationally high in cost. Appearance-based methods, associated with less hardware complexity and computational cost are low.

The proposed method takes as input an image captured with the Kinect sensor and its accompanying depth map. To isolate the hand in 2D and 3D, skin color detection followed by depth segmentation is used. The chosen 3D hand model is made up of a collection of appropriately assembled geometric primitives. A vector with 27 parameters represents each hand position. Hand articulation tracking is formulated as an estimation problem. The 27 hand model parameters minimize the difference between hand hypotheses and hand the actual observations. Thus, use graphics rendering techniques to generate comparable skin and depth maps for a given hand pose hypothesis in order to quantify this disparity. An appropriate objective function is thus defined, and a variant of PSO is used to solve it.

Romero *et al*. [3] pre-deep learning and deep learning were introduced, prior to deep learning, there were attempts to estimate 3D hand pose from the monocular RGB input image, which tells about discriminative and generative approaches. The majority of the proposed cutting-edge 3D hand pose estimation methods, however, are based on convolutional neural networks (CNNs). A non-parametric method for estimating the 3D sequential pose of hands in interaction with objects was presented. The development of a method that not only handles severe occlusion from objects in the hand but also takes object shape into account in 3D hand reconstruction is one of the contributions of this paper. Furthermore, the method is nonparametric and provides 3D hand reconstruction in real-time while accounting for time continuity constraints. Experiments revealed that the method estimates hand pose in real-time while being robust against segmentation errors, and that taking multiple hypotheses of previous hand pose into account improves the method's robustness to temporary estimation errors.

J. Tompson *et al*. [4] proposed a non-parametric method for estimating the 3D sequential pose of hands in interaction with objects was presented. The development of a method that not only handles severe occlusion from objects in the hand but also takes object shape into account in 3D hand reconstruction is one of the contributions of this paper. Furthermore, the method is nonparametric and provides 3D hand reconstruction in real-time while accounting for time continuity constraints. Experiments revealed that the method estimates hand pose in real-time while being robust against segmentation errors and that taking multiple hypotheses of previous hand pose into account improves the method's robustness to temporary estimation errors. The accuracy of offline model-based dataset generation routines is used to support a robust real-time convolutional network architecture for feature extraction in this pipeline.

Melax *et al*. [5] Tracking the full skeletal pose of the hands and fingers is a difficult problem with numerous applications for user interaction. Existing techniques either necessitate wearable hardware, restrict user pose, or necessitate significant computational resources. This study investigates a novel method for tracking hands, or any articulated model, using an augmented rigid body simulation. As a result, we can define 3D object tracking as a linear complementarity. a problem with a clear solution Based on the data from a depth sensor samples, the system generates constraints that limit motion orthogonal to the surface of the rigid body model. These constraints, together with a projected Gauss-Seidel solver is used to resolve problems involving prior motion, collision/contact constraints, and joint mechanics. The numerous surface constraints are impulse capped to avoid overpowering mechanical constraints due to camera noise properties and attachment errors. Multiple simulations are spawned at each frame and fed a variety of heuristics, constraints, and poses to improve tracking accuracy. A 3D error metric selects the best-fit simulation, assisting the system in dealing with difficult hand motions. This method allows for real-time, robust, and accurate 3D skeletal tracking of a user's hand on a variety of depth cameras while only requiring a single x86 CPU core for processing. Tracking fidelity improves as resolution, model accuracy, and camera frame rate increase. The system runs comfortably on a single CPU core, but it could be extended to search and simulate more possible poses using additional computing resources to improve robustness even further.

Zimmermann *et al.* [6] proposed a method for estimating 3D hand pose from regular RGB images in this paper. Because of the missing depth information, this task has far more ambiguities. We propose a deep network that learns a network-implicit 3D articulation prior to accomplish this. This network produces good estimates of the 3D pose when combined with detected keypoints in the images. For training the involved networks, we present a large-scale 3D hand pose dataset based on synthetic hand models. Experiments on a variety of test sets, including one for sign language recognition, show that 3D hand pose estimation on single color images is feasible.

Shanxin *et al* [7] Shanxin et al [7] proposed an approach taxonomy Learning-based approaches have been found to be effective for solving single-frame pose estimation, with hand model fitting as an option for greater precision. compared methods on a new dataset and came to the conclusion that deep models are ideal for pose estimation It also emphasized the requirement for large-scale training sets in order to train generalizable models. By comparing deep learning methods to previous analyses. Performing fine-grained analysis on a large-scale dataset with a variety of error sources and design options. The HIM2017 challenge methods provide insights into the current state of 3D hand pose estimation. (1) 3D volumetric representations used with a 3D CNN perform well, possibly because they capture the spatial structure of the input depth data better. (2) While detection-based methods outperform regression-based methods, regression-based methods can achieve good performance when explicit spatial constraints are used. Using richer spatial models, such as bone structure, can help even more. In extreme viewpoint cases where severe occlusion occurs, regression-based methods perform better.

Yidan *et al.* [8] introduced a technique for both accuracy and real-time performance, that is Hand Branch Ensemble network (HBE), a novel three-branch Convolutional Neural Network with three branches representing the three parts of a hand: the thumb, index finger, and other fingers. The HBE network's structural design is inspired by an understanding of the differences in the functional importance of different fingers. Furthermore, a feature ensemble layer, in conjunction with a low-dimensional embedding layer, ensures that the overall hand shape constraints are met. The experimental results on three public datasets show that the approach outperforms state-of-the-art methods with less training data and shorter training times.

Yujun *et al*. [9] turned to render low-cost synthetic hands with 3D models, from which 3D joint ground truth can be easily obtained. Although this method performs well on the synthetic dataset, it does not generalize well to real image datasets due to domain shift between image features. A discriminative approach was used to localize the 2D keypoints, and a model fitting method was used to calculate the 3D pose. CycleGANs to generate a "real" dataset from a synthetic dataset. However, poor performance demonstrates that there is still a gap between generated "real" images and real-world images.

Seungryul *et al.* [10] Estimates 2D-3D mapping ambiguities with limited training data, then estimating 3D hand meshes from single RGB images is difficult. It uses a 3D parametric hand model that is compact and represents deformable and articulated hand meshes. Thus, investigate and contribute in three ways to achieve model fitting to RGB images: 1) Neural rendering: Inspired by recent work on the human body, our hand mesh estimator (HME) is implemented using a neural network and a differentiable renderer, which is supervised by 2D segmentation masks and 3D skeletons. HME outperforms other methods for estimating hand shapes and improves pose estimation accuracy. 2) Refinement through iterative testing: Our fitting function is differentiable. In the spirit of iterative model fitting methods such as ICP, we use gradients to iteratively refine the initial estimate. 3) Self-data supplementation: obtaining sized RGB-mesh (or segmentation mask)-skeleton triplets for training is a significant challenge. After successfully fitting the model to the input RGB images, it meshes, i.e., shapes and articulations, are realistic, and we augment view-points on top of estimated dense hand poses. Experiments with three RGB-based benchmarks show that our framework outperforms the state-of-the-art in 3D pose estimation and recovers dense 3D hand shapes. Each of the technical components listed above significantly improves the accuracy of the ablation study.

Adnane *et al* [11] proposed a CNN (Convolutional Neural Network), that forecasts a set of hand and view parameters. The decoder is made up of two parts: a pre-computed articulated mesh deformation hand model that generates a 3D mesh from hand parameters and a re-projection algorithm. View parameters-controlled module that projects the generated hand into the domain of images We demonstrate this by utilizing the shape and pose prior knowledge encoded in the A hand model embedded in a deep learning framework achieves cutting-edge performance in 3D pose prediction from images. on standard benchmarks, and yields geometrically valid results as well as plausible 3D reconstructions. Furthermore, it also demonstrates that training with weak supervision in the form of 2D joint annotations on datasets of wild images, combined with full supervision in the form of 3D joint annotations on limited available datasets, allows for good generalization to 3D shape and pose predictions on wild images.

3. CONCLUSION

3D hand motion has a wide range of applications, including gaming, biomechanical analysis, robotics, human-computer interaction such as augmented reality and virtual reality (AR/VR), and many others. Therefore proposed the first learning-based approach for estimating monocular hand pose and shape using data from two completely different modalities: image data and MoCap data. A trained inverse kinematics network that directly regresses joint rotations is included in our new neural network architecture. In terms of accuracy, robustness, and runtime, these two factors result in a significant improvement over the state of the art. Another avenue is the simultaneous capture of two interacting hands from a single RGB image, which is currently only possible with depth sensors. This work is a summary of different techniques for real-time hand motion capture with their advantages and limitations.

ACKNOWLEDGEMENT

We would like to thank the Director of LBSITW and the Principal of the institution for providing the facilities and support for our work.

REFERENCES

[1] Thomas B Moeslund, Adrian Hilton, and Volker Kru. A Survey of Advances in Visionbased Human Motion Capture and Analysis. Computer Vision and Image Understanding, 104:90–126, 2006

[2] Ali Erol, George Bebis, Mircea Nicolescu, Richard D. Boyle, and Xander Twombly. Vision-based Hand Pose Estimation: A review. Computer Vision and Image Understanding, 108(1-2):52–73, 2007.

[3] J. Romero, H. Kjellstrm, and D. Kragic. Hands in action: real-time 3d reconstruction of hands in interaction with objects. In ICRA, 2010

[4] Stan Melax, Leonid Keselman, and Sterling Orsten. Dynamics based 3d skeletal hand tracking. In Proceedings of Graphics Interface 2013, pages 63–70. Canadian Information Processing Society, 2013.

[5] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. Real-time continuous pose recovery of human hands using convolutional networks. ACM Transactions on Graphics (ToG), 33(5):169, 2014.

[6] C. Zimmermann and T. Brox. Learning to estimate 3d hand pose from single rgb images. In ICCV, 2017

[7] Shanxin Yuan, Qi Ye, Bjorn Stenger, Siddhant Jain, and TaeKyun Kim. Bighand2.2m benchmark: Hand pose dataset and state of the art analysis. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[8] Yidan Zhou, Jian Lu, Kuo Du, Xiangbo Lin, Yi Sun, and Xiaohong Ma. Hbe: Hand branch ensemble network for realtime 3d hand pose estimation. In The European Conference on Computer Vision (ECCV), September 2018

[9] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In The European Conference on Computer Vision (ECCV), 2018.

[10] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[11]Adnane Boukhayma, Rodrigo de Bem, and Philip H.S. Torr. 3d hand shape and pose from images in the wild. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.