Finger Bendedness Classification from 3D Pose Regression

Jihyeon Kim[1], Changhwa Lee[1], Seongyeong Lee[1], Junuk Cha[2], Hansoo Park[2], Donguk Kim[2] and Seungryul Baek[2]

[1] UNIST, Computer Science and Engineering, Ulsan, South Korea, {jihyeon, changhwalee, skwithu}@ unist.ac.kr
[2] UNIST, AI Graduate School, Ulsan, South Korea, {jucha, hansupark, dukim, srbaek}@unist.ac.kr

Abstract

Understanding hand poses has been popular in the field of computer vision thanks to its potential to be used in the real-world applications such as augmented reality (AR), virtual reality (VR), human computer interaction (HCI) and etc. In this paper, we propose the fraemwork for the finger bendedness classification which classifies 5 finger's bendedness into 'bending', 'half-bending' or 'unbending'. Since directly estimating such bendedness of each finger from RGB images is non-trivial, we constituted our framework having 3 distinct stages: we first achieve the 3D skeletal regression that estimates 3D coordinate values of 21 keypoints lying on the human's hands. Then, we calculate the angles between finger joints using the inverse kinematics from the regressed 3D keypoint coordinates. Finally, we map the angles for each finger towards 3 types of finger bendedness label ('bending', 'half-bending' or 'unbending'). During the mapping, we propose to use 2 types of finger bendedness classification models for thumb finger and others, as the angle distribution for the thumb finger is different from those for other fingers; while angle distributions for other fingers are similar each other. Experiments demonstrated the effectiveness of our method showing the superior performance compared to the finger bendedness classification baseline based on the ResNet-50 architecture.

Keywords— *Hand pose classification, Hand pose estimation, Inverse kinetic, Regression*

I. Introduction

Mobile devices such as smart phones, laptops, are making our lives convenient everyday and becoming indispensable in our daily lives. On top of that, the real-world applications such as virtual reality (VR), augmented reality (AR), convenient human-computer interaction (HCI) and human-machines interface (HMI) applying AI technologies are increasingly introduced and commercialized recently. For these applications, understanding hand poses is crucial, as hand poses are the primary interacting user interface for humans to interact with the outer environments.

Hand pose estimation is usually framed as the problem to regress 3D coordinates of the keypoints lying on the human's hands or to classify the discrete hand gesture labels. Continuous coordinate values for 21 hand joints contain rich information, however post-processing is further required to disentangle them towards the human intention; while the discrete gesture label is already disentangled, however it is non-trivial to extend it to encode the diverse human intention. We try to find the compact representation that better represents the diverse human intention. In this paper, we propose to encode it as the bendedness of each finger as the representation is simple and general enough to encode diverse human intentions and gestures.

Directly estimating the finger bendedness from single RGB images is non-trivial and we constituted our framework having 3 distinct stages: We first estimate the 3D coordinates of the 21 hand keypoints using the Mediapipe [5] algorithm. By applying the inverse kinematics (IK) to the estimated 3D coordinate values, we reveal the 9 angles (3 angles between 4 joints in x, y and z direction) from 4 joints constituting each finger. Finally, the finger bendedness classification is performed for each finger by training the classification model. Due to the lack of finger bendedness classification datasets, we created datasets by collecting the estimated 3D skeletons and their ground truths for finger bendedness. Since 5 fingers are independent, it is inefficient to map angles of 5 fingers altogether into the finger bendedness label. Instead, we map 3 angles obtained from each finger into the finger bendedness labels and therefore we obtained 5 finger bendedness labels from a hand image input.

The contribution of our paper is summarized as follows:

- We tackled the relatively new task, finger bendedness classification, for understanding human hand poses in the form of discrete finger bendedness information.
- Due to the lack of relevant datasets, we further developed the user interface for achieving the efficient data

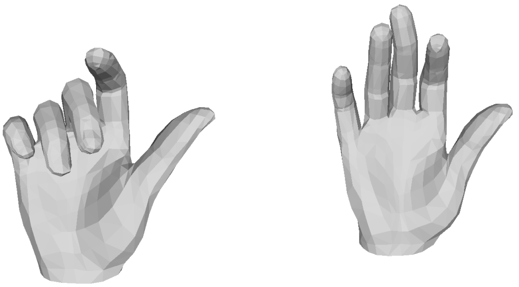


Fig. 1. The visualization of angles re-constructed by the inverse kinematic (IK) process. The angles obtained from 3D skeletons estimated from the Mediapipe could be used for the MANO [17] mesh model's pose parameter .

collection.

- Given collected datasets, we proposed the novel finger bendedness classification pipeline using inverse kinematics and classification models. Also, we have demonstrated its effectiveness by comparing it with ResNet-based architecture that is learned to directly output the finger bendedness from RGB images.

II. RELATED WORK

Hand pose estimation. For many years, hand pose estimation studies use convolutional neural network (CNNs) and most of researches proposed for single 3D hand pose estimation [4, 14, 23]. [14] proposed real-time hand tracking system that tracks global 3D hand pose from RGB-only images and also suggested deep generative model to learn latent for hand. [23] proposed a self-supervised method for 3D hand pose estimation from depth maps. Also, they introduced a approach to couple unsupervised model-based fitting with supervised discriminative approaches for hand pose estimation. And there are studies of 3D single and interacting hand pose estimating papers [15, 18, 13]. [13] introduced InterNet (using ResNet) for 3D single and interacting hand pose estimation. And InterNet estimates handedness, hand pose, and right hand-relative left hand depth from a single RGB image. [19, 16, 27, 8] perform regression for pose estimation. [8] introduced new 2.5D representation of hand pose and then provide method to reconstruct the 3D pose from 2.5D using regression. And in our paper, we use regression for hand classification.

Model-based Hand classification and recognition. Most studies on hand classification and recognition are often based on various models. Based on the most representative CNN, [1, 25] uses CNN to recognize egocentric hand gesture and detect fingertip or to solve the problem of large model size and slow execution speed. In addition to CNN, there are many studies based on other models. [11] suggested Pose-TGCN that model spatial and temporal dependencies in human pose trajectories simultaneously. And there are also two model-based classification. For instances, [10] recognize dynamic hand gesture of video stream in real-time using lightweight CNN architecture and then classify the hand using ResNeXt-101 model. And [7] designed end-to-end learnable model for joint 3D reconstruction of hands and objects using MANO [17] model. As such, in the field of hands using mesh, it often depends on MANO model. However, in this paper, the mesh can be created using the IK process without using MANO parameter, and can be applied to classification.

Inverse kinematic process. The inverse kinematic (IK) process is process of changing the position of the fingertip to the value of joint space. In recent years, IK process is one of the extensively studied fields. And there have been various suggestions to implement IK process. The simple way is computational methods. For example, [2] suggest computational ways for high nonlinearities using adaption control laws. [6] described PODA system that utilizes pseudo inverse control in order to solve redundant limbs problem. And [24, 22, 3] also proposed new numerical solution to the general version of inverse kinematic. But these methods has problem of optimization. The optimization was quite repetitive time-consuming task. So, various researches exist to solve this problem. Among them there are heuristic methods such as CDC, FABRIK. They pay a low computational cost for each heuristic iteration. In addition to heuristic method, there are also analytical solutions. And there is a study that proposed using combination algorithm of numeric and analytical method [20]. And most recently, there are studies using neural network to solve IK problems [9, 21, 12]. [9] designed HOPS-Net using CNN for Hand-held Object Pose and Shape estimation and [21] proposed recurrent neural network architecture for unsupervised motion retargeting. [12] proposed differentiable HybrIK, a hybrid analytical-neural IK solution that converts the accurate 3D joint locations to full 3D human mesh. And we used IK process introduced by [12] for 3D joint estimation.

III. METHOD

A. 3D hand pose estimation

We used the Mediapipe framework [5] for achieving the 3D hand pose estimation, which is developed based on the CNN architecture and provides the 3D coordinate values from 21 hand keypoints in the real-time manner. The input to the Mediapipe network ($\mathbf{H}$) is RGB images (I), and the output is 3D joint location ($l = \{u, v, d\}$), where d denotes the relative depth. The network $\mathbf{H}$ is trained to map the image input I towards the hand joint location l.

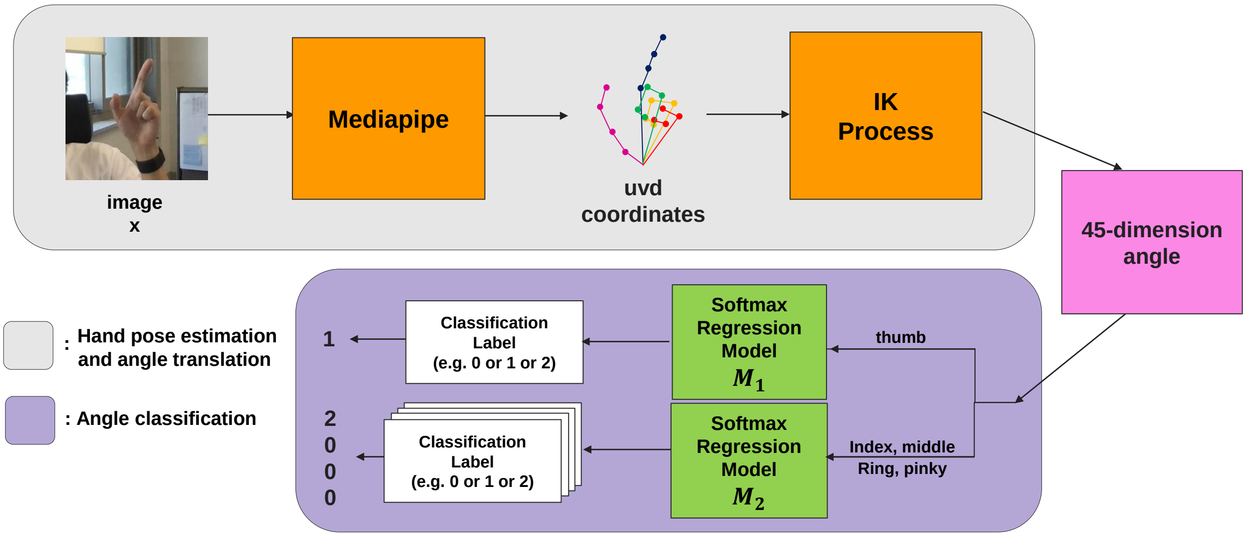


Fig. 2. Overview of our finger bendedness classification framework. First, Mediapipe [5] is exploited to achieve the 3D hand pose estimation. It outputs 3D uvd coordinates from the input RGB hand images. Then, the uvd coordinate is translated to 45-dimension angles using the inverse kinetic (IK) process. Two classification models are used to efficiently achieve the angle classification: The first model M_1 is used for classifying bendedness of the thumb finger; and the second model M_2 is used for classifying bendedness of other fingers (index, middle, ring, pinky).

B. *Obtaining angles from 3D coordinates*

We proposed to use the inverse kinematics to obtain the finger angle from the estimated 3D hand keypoints. While the forward kinematics problem is well-posed, the inverse kinematics problem is ill-posed, since there is either no or many solutions for the inverse kinematic problem. We defined the fundamental equation of inverse kinematic process using the Jacobian matrix ($\mathbf{J}$) considering single joint with its joint at the origin. The world coordinates of the end point with L_1 distance away from origin are obtained as follows:

$$x(q) = \begin{bmatrix} L_1 cos(q_1) \\ L_1 sin(q_1) \end{bmatrix} \tag{1}$$

The Jacobian matrix (J) of this point obtained by taking derivative with respect to each of the joint coordinates (q_1) is obtained as follow:

$$J(q) \overset{\text{def}}{=} \frac{\partial}{\partial q} x(q) = \begin{bmatrix} \frac{\partial(L_1 cos(q_1))}{\partial q} \\ \frac{\partial(L_1 sin(q_1))}{\partial q} \end{bmatrix} = \begin{bmatrix} -L_1 sin(q_1) \\ L_1 cos(q_1) \end{bmatrix} \tag{2}$$

Since J is neither square nor invertible matrix and it suffers from the singularity problem. Instead of solving it accurately, we solve it using the approximation. To reduce the inference time and to embed differentiable model, we adopt HybridIK method [12], which composes the entire rotation recursively along the kinematics tree. A keypoint of this method is that there is no need for the additional optimization procedure and it is differentiable, which allow us 3D hand mesh in an end-to-end manner. This process can be denoted as:

$$\mathbf{R} = \mathrm{IK}(\mathbf{P}, \mathbf{T}), \quad \mathbf{T} = \mathbf{H}(I) \tag{3}$$

where $\mathbf{R} = \{R_{\mathrm{pa(k)},k}\}_{k=1}^{K}$, with desired locations of input hand joints $\mathbf{P} = \{p_k\}_{k=1}^{K}$ and rest pose template $T = \{t_k\}_{k=1}^{K}$. K is the number of hand joints, $t_k \in \mathbb{R}^3$ is k-th joint location of the rest pose template. $\mathrm{pa}(k)$ return the parent's index of the k-th joint, and $R_{\mathrm{pa}(k),k}$ is the relative rotation of k-th joint with respect to its parent joint. Ideally, generated rotation matrix should satisfy the following condition:

$$p_k - p_{\mathrm{pa}(k)} = R_k(t_k - t_{\mathrm{pa}(k)}) \quad \forall 1 \leq k \leq K. \tag{4}$$

For our method, we could get 16 joint angle (there is no angle of finger tips) described in the rotation matrix. These joint angles could be input to the classification model. To apply the estimated pose to the MANO [17] model as in Fig. 1, we convert the rotation matrix to the axis-angle representation.

C. *Angle classification*

For angle classification, we trained two classification models (M_1, M_2). To classify 5 finger's angles into their bendedness, it may require to train 5 independent classification models. However we trained only 2 classification models as the distribution since the thumb finger is different from that for remaining 4 fingers; while the distribution is similar for 4 fingers. The first model M_1 is the classification model which is responsible for the thumb finger classification, and the second model M_2 is reserved for classifying the remaining 4 fingers. Furthermore, if we use only

2 classification models, the data collection becomes easier. We collect the data for only thumb and index fingers: As the middle finger, ring finger, and pinky finger are limited in their rotation angles compared to the index finger, data obtained by the index finger could span the coverage of these fingers. Therefore, the M_2 classification model which is trained by index finger data is used for classifying the bendedness of the 4 fingers.

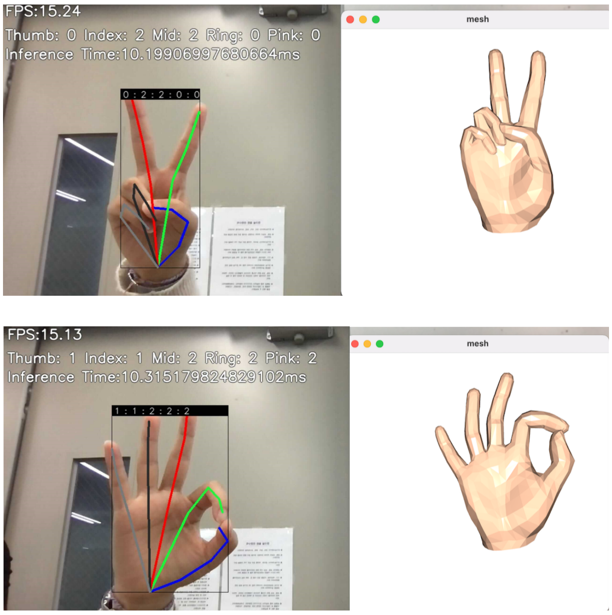
After applying the inverse kinematic (IK) process, we obtain the 45-dimensional angle vector which represents 9 angles for 5 fingers. So, the input to the classification models M_1 and M_2 are the 9-dimensional vector, which encodes the rotations about the z (3-dim), ϕ (3-dim), and θ (3-dim) axes from 4 bone connections, respectively. Via the classification models, we map the 45-dimensional angle vector into three types of bending labels: ‘bending(0)’, ‘half-bending(1)’ and ‘unbending(2)’.

D. *Independent finger-level data collection*

We collect a dataset using the webcam and our annotation tool. As mentioned in the previous section, there are only two classification models: M_1, M_2 which are trained by thumb and index finger data. Compared to the data collection using whole 5 finger configuration, the data collection becomes much more efficient when using only 2 fingers: If we collect data in the holistic way for 5 finger configuration as in [26], the overall configuration of the hand pose becomes $3^5 = 243$ (having 0, 1 and 2 for 5 fingers). However, if we collect data for 2 finger configuration and regard each finger independent, the overall configuration becomes only $3 \times 2 = 6$ (collect 0, 1 and 2 for 2 independent fingers). We could rule out the viewpoint variation as the angle inputs are invariant to the viewpoints. Our annotation tool records hand images, the Mediapipe and IK process are applied subsequently to obtain the angles from the images. Then, we manually annotate bendedness of the thumb and index fingers. During the data collection, we fold and unfold thumb and index fingers sequentially. Overall, we have collected 400 number of angle and bendedness label pairs to train M_1 and M_2 classification models, respectively.

IV. Experiments and Results

Tables 1 is the quantitative results of cross-subject angle classification for thumb, index finger, middle finger, ring finger, and pinky finger, respectively. The experimental metrics are accuracy, precision, recall and f1-score. Overall, the classification results are good. We could also see the validity of the second classification model M_2 that is trained by the index finger data for middle, ring and pinky fingers. Also, we make the real-time demo using our framework. The Fig 3 is the visualization obtained from our classification models.



FPS:15.24
Thumb: 0 Index: 2 Mid: 2 Ring: 0 Pink: 0
Inference Time:10.19906997680664ms

FPS:15.13
Thumb: 1 Index: 1 Mid: 2 Ring: 2 Pink: 2
Inference Time:10.315179824829102ms

Fig. 3. Examples of real time demos using our framework. Upper image shows the annotation of the finger gesture as [0, 2, 2, 0, 0] written in black box. And the lower image shows the annotation of the finger gesture as [1, 1, 2, 2, 2] written in black box.

Finger	Label	Precision	Recall	F1-Score	Acc
Thumb	0	0.93	0.84	0.89	0.80
	1	0.19	0.38	0.25	
	2	0.77	0.85	0.81	
Index	0	0.99	0.89	0.94	0.89
	1	0.46	0.86	0.60	
	2	0.72	0.90	0.80	
Middle	0	1.00	0.95	0.97	0.92
	1	0.50	0.52	0.51	
	2	0.69	0.95	0.80	
Ring	0	1.00	0.87	0.93	0.87
	1	0.52	0.94	0.67	
	2	0.79	0.75	0.77	
Pinky	0	0.99	0.66	0.79	0.74
	1	0.40	0.81	0.54	
	2	0.69	0.98	0.81	

Table 1. The test result of cross-subject using our method. It shows each classification result of finger labels. It is better when it closed to 1.

Finger	Label	Precision	Recall	F1-Score	Acc
Thumb	0	0.34	0.75	0.47	0.26
	1	0.00	0.00	0.00	
	2	0.00	0.00	0.00	
Index	0	0.76	0.60	0.67	0.47
	1	0.00	0.00	0.00	
	2	0.08	0.10	0.09	

Table 2. The test result of cross-subject fingers using ResNet-50 baseline. It shows each classification result of finger labels. The 0 result indicates that the model mis-predict the corresponding finger label.

We also trained and tested the ResNet-50 model by our training datasets which consist of thumb and index fingers. Since the middle, ring and pinky finger datasets are limited, we are not able to experiment on these finger with the same environment of our method. So, the result of ResNet-50 model have only thumb and index finger as table 2. As shown in table 2 the accuracy of the ResNet-50 model is far less than that of our method.

V. Conclusion

In this work, we propose the finger bendedness classification framework using the Mediapipe, inverse kinetic(IK) and two classification models. The Mediapipe is involved to estimate 3D hand pose in real-time and via IK process, we are able to obtain 45-dimension angles. After training two classification models by our collected datasets, models become able to classify the finger bendedness. By regarding each finger independent, we could secure our data collection process efficiently. Experimental results demonstrate that the finger bendedness is performed well compared to the state-of-the-art image classification baseline using ResNet-50 architecture.

VI. Acknowledgement

This work was supported by the IITP grants (No. 2021–0–01778, Development of human image synthesis and discrimination technology below the perceptual threshold) funded by the Korea government (MSIT).

References

[1] Mohammad Mahmudul Alam, Mohammad Tariqul Islam, and SM Mahbubur Rahman. Unified learning approach for egocentric hand gesture recognition and fingertip detection. *Pattern Recognition*, 2022.

[2] Aldo Balestrino, Giuseppe De Maria, and Lorenzo Sciavicco. Robust control of robotic manipulators. *IFAC Proceedings Volumes*, 1984.

[3] Samuel R Buss and Jin-Su Kim. Selectively damped least squares for inverse kinematics. *Journal of Graphics tools*, 2005.

[4] Martin de La Gorce, David J Fleet, and Nikos Paragios. Model-based 3d hand pose estimation from monocular video. *IEEE transactions on pattern analysis and machine intelligence*, 2011.

[5] Andrey Vakunov Andrei Tkachenka George Sung Chuo-Ling Chan Fan Zhang, Valentin Bazarevsky and Matthias Grundmann. Mediapipe hands: On-device real-time hand tracking. *CoRR*, 2020.

[6] Michael Girard and Anthony A Maciejewski. Computational modeling for the computer animation of legged figures. *ACM SIGGRAPH Computer Graphics*, 1985.

[7] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[8] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[9] Mia Kokic, Danica Kragic, and Jeannette Bohg. Learning to estimate pose and shape of hand-held objects from rgb images. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019.

[10] Okan Köpüklü, Ahmet Gunduz, Neslihan Kose, and Gerhard Rigoll. Real-time hand gesture detection and classification using convolutional neural networks. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 2019.

[11] Dongxu Li, Cristian Rodriguez, Xin Yu, and Hongdong Li. Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2020.

[12] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. 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*, 2021.

[13] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, 2020.

[14] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[15] Franziska Mueller, Micah Davis, Florian Bernard, Oleksandr Sotnychenko, Mickeal Verschoor, Miguel A Otaduy, Dan Casas, and Christian Theobalt. Real-time pose and shape reconstruction of two interacting hands with a single depth camera. *ACM Transactions on Graphics (TOG)*, 2019.

[16] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017.

[17] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (ToG)*, 2017.

[18] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017.

[19] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[20] Deepak Tolani, Ambarish Goswami, and Norman I Badler. Real-time inverse kinematics techniques for anthropomorphic limbs. *Graphical models*, 2000.

[21] Ruben Villegas, Jimei Yang, Duygu Ceylan, and Honglak Lee. Neural kinematic networks for unsupervised motion retargetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[22] Charles W Wampler. Manipulator inverse kinematic solutions based on vector formulations and damped least-squares methods. *IEEE Transactions on Systems, Man, and Cybernetics*, 1986.

[23] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. Self-supervised 3d hand pose estimation through training by fitting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[24] William A Wolovich and H Elliott. A computational technique for inverse kinematics. In *The 23rd IEEE Conference on Decision and Control*, 1984.

[25] Fan Yang, Yang Wu, Sakriani Sakti, and Satoshi Nakamura. Make skeleton-based action recognition model smaller, faster and better. In *Proceedings of the ACM multimedia asia*. 2019.

[26] Shanxin Yuan, Qi Ye, Bjorn Stenger, Siddhant Jain, and Tae-Kyun Kim. Bighand2.2m benchmark: hand pose dataset and state of the art analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017.

[27] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Weakly-supervised transfer for 3d human pose estimation in the wild. In *IEEE international conference on computer vision, ICCV*, 2017.

SUMMARY OF THIS PAPER

A. Problem Setup

Previous hand pose estimation or classification and gesture recognition studies, there were many model-based methods. But, since these model-based methods are parameter-based, they must be pre-defined, and the process of learning parameters is a nonlinear process. Also, the image-model misalignment be occured. Because of these problem, the performance will be further reduced. On the other way, the 3D keypoint estimation method can obtain pixel-level localization accuracy by combining deep CNN network and volumetric representation, but it can predict unrealistic body structure. These problems result in poor task performance.

B. Novelty

In this paper, we used the inverse kinetic(IK) process, which can reduce the difference between mesh and 3D keypoint, to solve unrealistic body structure prediction problems without using model-based methods. Also, by using Mediapipe, it is possible to create a demo that can infer hand pose classification in real time. Finally, we easily collected datasets for training the regression model.

C. Algorithms

From the image input, we can get the uvd coordinates through the hand pose estimation. This uvd coordinate is translated into 45-dimension angle using inverse kinetic process. And use two classification models to classify from a 45-dimension angle. First classification model is only for classification of thumb. And the second model is for the other fingers (index, middle, ring, pinky). Note that the second model trained by only index finger dataset. Therefore we can get the classification label for each finger.

D. Experiments

The experiment result metrics of angle classification are accuracy, precision, recall, f1-score. Overall results of classification for each label and each finger show good performances. So, we can check that our proposed algorithms results in good inference score. Here, we also see the validity of second model which trained by only index finger dataset. Also, we make the real time demo using our framework. It works well in real time inference and classification.