Deep Learning-based Hand Pose Estimation from 2D Image

Jungpil Shin*, Md Abdur Rahim, Okuyama Yuichi, Yoichi Tomioka

School of Computer Science and Engineering, University of Aizu, Aizuwakamatsu, Fukushima, Japan.
*Corresponding Author: Email: jpshin@u-aizu.ac.jp

2020 3rd IEEE International Conference on Knowledge Innovation and Invention (ICKII) | 978-1-7281-9333-5/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICKII50300.2020.9318917

Abstract

Human hands, the most significant part of a human body, are used as a healthy and secure medium in modern technology for the development of human-computer interaction. There are various technical applications like virtual reality, touch-free writing, aerial handwriting, sign language which are performed based on the user's hand gestures. The ability to recognize hand shape and motion can be the basis for understanding the hand gesture control. In the above context, we propose a deep learning technique for hand gesture estimation that identifies hand pose from input images that reveal depth information. This system uses the convolutional neural network (CNN) to detect 3D hand gestures with the help of palm joints and fingertips. We analyzed the positions of 2D joints and fingertips and determined the depth information. From the experimental results, the proposed system is able to present an accurate estimate of the depth information from 2D hand images.

Keywords: hand pose, convolutional neural network (CNN), depth information.

Introduction

The human hand pose is a useful means of communication that allows people to interact with the computer as naturally as possible and perform various tasks intensively. It has various technological applications such as human-computer interaction [1], virtual reality [2], touch-free communication, non-touch writing [3], sign language recognition [4]. Currently, the estimation of hand poses has attracted the attention of researchers. However, performing efficient and effective work is still challenging due to the complexity of diversity, multiple perspectives, self-identifying, different shapes, and sizes. Many researchers have used a deep camera to detect hand gestures from time-series data. Microsoft Kinect, a short-range depth sensor, Primer sense Carmine, Leap Motion are used to estimate body pose [5-6]. The single channel of the hand pose is inferred from the image which contains the value of the depth of each pixel. Depth images contain the information about the image and depth values that are related to the data needed to estimate 3D hand poses. However, due to more degrees of freedom (DoF), hand pose estimation has a unique challenge than body postures. Moreover, the depth camera can capture RGB information as well as depth information, being effective as a non-touch interface [7]. However, these types of devices are special and used for a limited purpose. In addition, deep learning-based models are appearing in 3D hand tracking and pose estimation. Reference [8] presented some constraints of hand joints and simulated from the real images. The 3D shape of the hand model was proposed for moving hand posture identification using silhouette features and motion prediction [9]. However, hand pose estimates are not yet sufficient to perform flexible and reliable work for potential applications.

In this study, we improved 3D hand pose estimation from 2D images of RGB cameras that are cheap and easily accessible and available on regular laptops and mobiles. We trained convolutional neural networks that showed enhanced performance and vector representations for 3D poses with a direct heatmap method for 2D pose inference. We focused on hand localization, left and right classification, 3D hand pose retrieval, and tracking joints.

Proposed Approach

The following steps are considered for 3D hand pose estimation: hand localization, hand image frames, and tracking of joints and fingers. The first step is to find the area of the hand from an image to crop. We, then proceed to create a heatmap of the hand keypoints to identify the 2D joints. Finally, 3D joints are estimated from 2D joints. However, this system uses tracking methods for continuous image frames to create smooth hand pose estimation. Figure 1 depicts the process of the 3D hand pose estimation system.

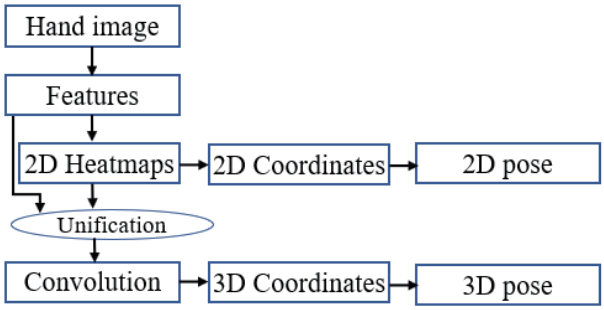

Fig. 1 The estimation process of the 3D hand pose.

A. Hand Localization

The pose of the hand is estimated from a 2D image where we give a hint for understanding the location of the hand. For this, we train the CNN model to output a bounding box that takes a cropped image that is fed as the input for a network of hand pose estimation. To identify the right and left hands, we used data augmentation and train the network to work for both hands. The image mask M was made for each hand for background augmentation by using Eq. (1).

$$I_a = I_o * \mathrm{M} + I_b * (1\text{-M}) \quad (1)$$

where I_a, I_o, I_b represents the augmented, original, and background images, respectively.

B. Keypoints

We used 21 keypoints in this study. (Fig. 2) We followed a sequence of keypoints similar to Synthhand's [10]. The

ISBN: 978-1-7281-9333-5

locations of the 2D keypoints are calculated using a heatmap that is displayed in coordinates (x, y) and converted to a heatmap set. To create a heatmap, a zero array of hand size is created and placed at the (x, y) coordinate position of the value 1 where the hand belongs. Otherwise, the value is set to be 0. A Gaussian blur technique overfits the model and then normalize the heatmap (HT) by using Eq. (2).

$$HT=HT_{blurred} / max\ (HT_{blurred}) \quad (2)$$

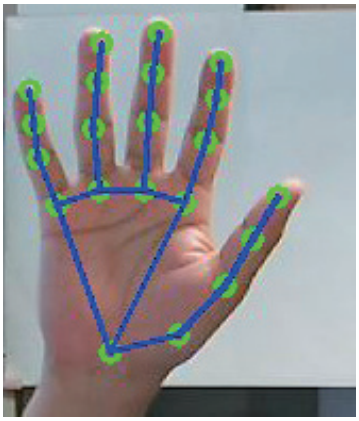

Fig. 2 Keypoints of a hand image.

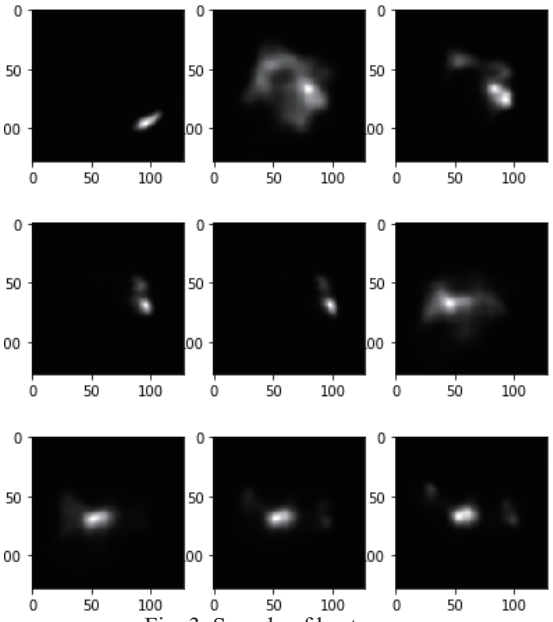

Fig. 3 Sample of heatmaps.

A total of 21 heatmap points was normalized. Figure 3 depicts the visualization of heatmaps of a random sample. Furthermore, the pixel coordinates were normalized and flat in the form of 2D vectors. Moreover, we predicted 3D coordinates using normalized representations [11] as vector size 3x21. The normalization was performed among the position of the middle finger and distance between wrist and middle fingers.

C. Predictive Model

We implemented the proposed model to predict hand pose as shown in Fig. 4. Here, the 3D keypoints coordinates were predicted through the vector of size 3 by 21 [11]. The proposed predictive model created both of 2D heatmap and intermediate image features, then unifying these features. Therefore, we applied convolutions, and the final properties were flattened and fully connected to predict the coordinates of the 3D pose.

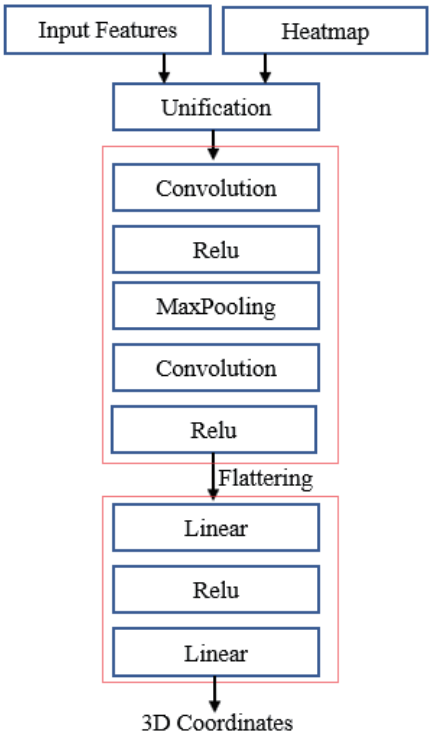

Fig. 4 Predictive model of 3D coordinates of hand pose.

Experimental Results and Evaluation

A. Dataset

For the training and testing of our proposed model, we used Synthhand's Benchmark dataset [10]. The data set of 4000 images was randomly selected for validation and 1500 images for testing, and the rest of the images for training. The image size was 128 x 128 pixels. We trained our model for both 2D and 3D locations.

B. Results and Evaluation

The evaluation of 3D hand pose estimation is considered through vector representation using the predictive method. However, the model was trained with heatmaps and the locations of the keypoints were estimated. The range of the keypoints locations was between 0 and 1 for both (x, y) coordinates that made an average error of 3% for keypoints joints. Figure 5 shows the example of the performance of 2D and 3D hand pose from the benchmark dataset. The performance became better when the hand image was a normal appearance and the key points were visible. Table 1 shows the model performance based on the validation sets. However, some images were similar in viewpoints and difficult to separate from the dataset. Therefore, these types of images were considered in both the train and the validation process. For this, the model showed high efficiency in validation datasets.

ISBN: 978-1-7281-9333-5

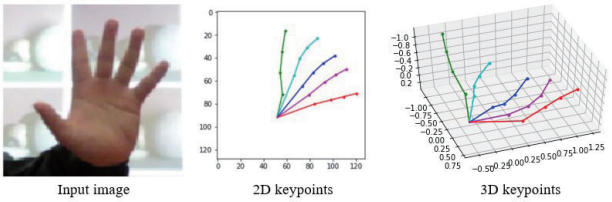

Fig. 5 An example of the input image of 2D and 3D keypoints.

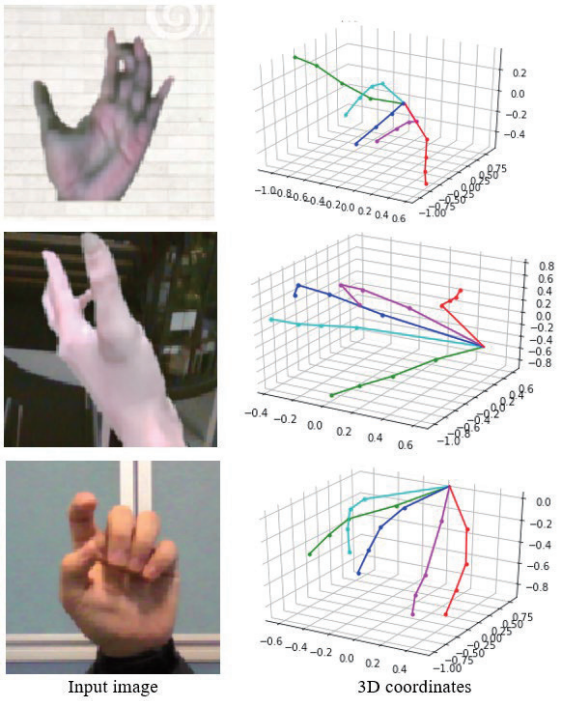

Fig. 6 Examples of 3D coordinates of input images.

Table II presents the average performance accuracy between the keypoints joints and fingertips. The predictive model showed an average accuracy of the keypoints of the fingers and joints at 97%. Figure 6 illustrates examples of 3D coordinates from input images.

TABLE I
VALIDATION PERFORMANCE OF PROPOSED MODEL

Dataset	Before trained		After trained	
	2D	3D	2D	3D
Non-augmented	0.030	0.23	0.09	0.19
Augmented	0.21	0.70	0.06	0.26

TABLE II
PERFORMANCE OF KEYPOINTS USING HEATMAPS

Keypoints		Model Performance (%)
Fingers	Thumb	97.6
	Index	96.2
	Middle	96.2
	Ring	96.9
	Little	96.8
Joints	Fingertip	96.3
Wrist		99
Average		97

Conclusion

This paper estimated 3D hand poses from 2D images. We considered hand images from benchmark datasets. The proposed CNN model created a 2D heatmap and image features. The convolution was then applied to obtain the normalized 3D hand pose. We trained and evaluated the proposed method for 2D and 3D estimations of heatmap and vector representation poses. The results show that the keypoints error of the model performance was about 3% for normalized 3D hand poses.

For future study, we may consider Graph Convolutional Network (GCN) for better hand-held non-touch interface performance such as character input, alphabet recognition.

References

[1] Krejov P. Real time hand pose estimation for human computer interaction. University of Surrey (United Kingdom); 2016.
[2] Ahmad A, Migniot C, Dipanda A. Hand pose estimation and tracking in real and virtual interaction: A review. Image and Vision Computing. 2019 Sep 1; 89:35-49.
[3] Rahim MA, Shin J, Islam MR. Gestural flick input-based non-touch interface for character input. The Visual Computer. 2019 Oct 1:1-4.
[4] Mandikhanlou K, Ebrahimnezhad H. Multimodal 3D American sign language recognition for static alphabet and numbers using hand joints and shape coding. Multimedia Tools and Applications. 2020 May 19.
[5] Shotton J, Fitzgibbon A, Cook M, Sharp T, Finocchio M, Moore R, Kipman A, Blake A. Real-time human pose recognition in parts from single depth images. In Computer Vision and Pattern Recognition, pp. 1297-1304, 2011.
[6] Gomez-Donoso F, Orts-Escolano S, Cazorla M. Accurate and efficient 3D hand pose regression for robot hand teleoperation using a monocular RGB camera. Expert Systems with Applications. 2019 Dec 1; 136:327-37.
[7] Shin J, Kim CM. Non-touch character input system based on hand tapping gestures using Kinect sensor. IEEE Access. 2017 May 11; 5:10496-505.
[8] Lee J, Kunii TL. Model-based analysis of hand posture. IEEE Computer Graphics and applications. 1995 Sep;15(5):77-86.
[9] Chua CS, Guan H, Ho YK. Model-based 3D hand posture estimation from a single 2D image. Image and Vision computing. 2002 Mar 1;20(3):191-202.
[10] Mueller F, Mehta D, Sotnychenko O, Sridhar S, Casas D, Theobalt C. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In Proceedings of the IEEE International Conference on Computer Vision Workshops 2017 (pp. 1284-1293).
[11] Mueller F, Bernard F, Sotnychenko O, Mehta D, Sridhar S, Casas D, Theobalt C. Ganerated hands for real-time 3d hand tracking from monocular rgb. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018 (pp. 49-59).

ISBN: 978-1-7281-9333-5