Deep Learning Based Framework for Hand Keypoints Detection from a Monocular RGB Images

Thesis submitted in partial fulfillment
for the Award of Degree

Doctor of Philosophy

by

Purnendu Mishra

DEPARTMENT OF ELECTRONICS ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY
(BANARAS HINDU UNIVERSITY)
VARANASI - 221005

Roll No. 16091004

Year 2022

Chapter 5

Conclusion and Future Scope

The primary objective of the dissertation " A Deep Learning Based Framework for Hand Keypoints Detection from a Monocular RGB Image "is to design algorithms that will facilitate the estimation of the positions of hand joints (often referred to as hand keypoints) from monocular RGB images. The estimation of the hand joints' position in the image helps in discerning the hand pose and understanding the hand gesture. The valid hand poses carry some pieces of information that humans have been using along with speech or independently for human-to-human communications. Humans generate certain signs by giving a unique form to hand joints and a set of these hand forms or poses create a language known as sign language. A group person knowing sign language can use it for one-to-one communication.

In the current century, computers and digital cameras have become an essential part of our daily lives. In the computer vision domain, it is being tried to mimic human vision. And, hand pose estimation is a computer vision task where it is attempted to detect hand keypoints and decipher or translate the information carried in a hand pose. The motivation behind hand pose estimation through computer systems comes from the capability it can offer to Human-Computer Interaction (HCI). A human hand has a very high degree of freedom (DOF) which enables it to take numerous poses. By

learning the information in each of these hand poses, a control system can be designed that would replace traditional input sources used with computers like keyboard, mouse, joystick, etc. This type of system is easily adoptable by humans as hand actions are more intuitive and natural.

The first step in estimating the hand pose in a computer-vision system is to capture the hand image using an imaging source. The use of RGB cameras is more predominant than the other sources as they are widely available commercially and are comparatively cheaper. However, the estimation of hand pose using images captured by an RGB camera comes with certain challenges. A subject of interest in an image can appear differently under different lighting conditions. Additionally, there is an influence of background objects. The major disadvantage is that it is very difficult to get the depth information of a subject in the image as it is not stored at the time of image capturing. However, the advancement in machine learning technology has paved the way to utilize RGB images in a pose estimation task. Especially, many tasks that otherwise seem challenging are being solved using Convolutional Neural Networks (CNN) in the computer-vision domain. The proposed dissertation is focused on solving the problem of hand keypoint detection from monocular RGB images. The algorithms are designed to estimate hand keypoint positions (either partial or complete) using a deep learning model. It has been tried to remove dependencies on hand localization. Moreover, attempts are made to perform hand keypoints detection on multiple hands simultaneously. Moreover, attempts are made to perform hand keypoints detection on multiple hands simultaneously.

Chapter 1 dealt with the introduction of the dissertation. The definition of hand keypoints detection is presented. The different types of hand pose estimation and the ways to differentiate one from the other are provided. The details of possible applications of hand keypoints detection are briefly described. The chapter covered the details of different ways that are being followed to estimate hand keypoints. The

dissertation is aimed at using RGB images is the primary input source. As there are some challenges associated with the RGB image, those are described in Chapter 1. Since the algorithms used for hand keypoints detection described in this dissertation use CNN as the major tool. Therefore, a brief description of the working principle of CNN architecture is presented. Furthermore, the motivation to take on this research topic and research problems are presented in this chapter.

Chapter 2 started the discussion on the algorithm proposed for hand keypoints detection. The approach started with first performing the partial hand keypoints detection. The partial keypoints detection can be considered as a process where instead of trying to estimate the position of twenty-one hand keypoints (most commonly used to define the hand pose), a subset of these keypoints is detected. The fingertips detection process is considered to be part of partial hand keypoints detection. In Chapter 2, three different algorithms are used for fingertips detection. The first two algorithms followed the traditionally used two-step approach where the hand is localized and then fingertips are detected. Algorithms based on multi-label classification and local regression using a set are anchors are used for fingertips detection in the two-step approach methodology. The third proposed algorithm performs the estimation of keypoints directly on the full-size image and does not require the hand localization step. The performance of the algorithms is comparatively better than the methods available.

In **Chapter 3**, we presented two different CNN-based algorithms for complete hand keypoints detection from a monocular RGB image. Both of the presented algorithms are independent of the hand localization step. However, an approach to localize the hands in the RGB image was discussed which was integrated with the single-stage keypoints detection step. The hand localization output is used to improve the keypoints detection accuracy of the small hands. Out of the two processes for hand keypoints detection presented in Chapter 3, the first method uses a grid structure to locate the fingertips. The probability for the presence of one or more hand keypoints in each grid cell was

computed. Based on a threshold value high probability grid cells are identified and the position of a keypoint is estimated. The second approach presented also makes use of grid cells but instead of a single cell of the grid, multiples are used to estimate the position of hand keypoints. The second approach can be extended to hand keypoints for both the hand of the user if they are used in making a gesture.

Chapter 4 presents a method of hand keypoints detection from multiple hands simultaneously. The methodology is based on the use of an intermediate feature generated by a CNN architecture during an object detection process. Since hand ROI information is already present in the feature maps, each hand ROI can be processed separately and independently to detect the hand keypoints of multiple hands. The process of training and inference is described in detail in this chapter. The model's performance was tested on datasets that provide ground-truth annotation for multiple hand keypoints. The results obtained are according to the hypothesis made.

Contributions

The original contributions of this dissertation are

- Three separate algorithms were proposed for fingertip detection from the monocular image. The first two algorithms were dependent on hand localization and worked effectively in estimating the fingertips' position in a certain hand gesture. Another algorithm proposed that robustly estimates fingertips from a full-size image. The experimental results show that the algorithm works effectively in varying hand gesture scenarios. The accuracy of the model is comparatively better than the state-of-the-art method available for fingertips detection.
- Two algorithms for complete hand keypoints detection were proposed. The algorithms were designed to work in a single stage and are effective for multiple hand poses. An approach to improve the hand keypoints detection for small hands was also proposed.
- A new method for simultaneous estimation of hand keypoints from multiple hands

is proposed. The methods are designed to work directly on full-size RGB images. The additional computation resources and time required during the hand localization (if a separate process is used) are saved. The experimental results prove that the algorithm effectively detects keypoints directly on multiple hands.

Future Scope

In this dissertation, an attempt has been made to develop algorithms for hand keypoints detection from RGB images in a monocular camera setup. However, several challenges are still left to be dealt with to continue this work. Some of those are

- *Estimation of hand keypoints position in 3D spaces*: The algorithm described in this dissertation provides only the 2D coordinate of the detected hand keypoints. These detected points are in the image space. To obtain the position of these points either in the camera coordinate system or the world coordinate system, intrinsic camera parameters are required. This makes the system dependent upon the camera used for capturing the image. Therefore, there will be a need for calibration from system to system for an algorithm to work accurately. Additionally, in the monocular RGB image, the depth information is not present which makes it difficult to estimate the depth of a keypoint from the imaging source. There is a need to tackle these challenges to have a robust solution for hand keypoints detection in three-dimensional space.
- *Inference time of deep learning model*: In a deep learning model, there are a lot of parameters that need to be optimized to have an accurate working system. In general, there are a lot of multiplication and summation operations involved in a neural network's inference process. Due to a large number of these operations, the inference of time a model gets affected. Additionally, to facilitate these many calculations specialized hardware like Graphical Processing Units (GPUs) are used which can parallelize the mathematical operation. GPUs are costly and cannot be used in all use cases. The low parameters model with low hardware requirements

is generally less accurate. Therefore, a deep learning architecture is to be designed in such a way that it should have the best accuracy with the shortest possible inference time and be suitable for real-time applications.

- *Occlusion issue*: The occlusion of the hand by itself or by an object in the image creates a major hurdle in the process of hand keypoint detection. Due to occlusions, there can be both false negatives and false positives. Therefore, to have a reliable system both the false positive and negative should be as minimal as possible.
- *Interacting as well as small hands challenge*: Often in many applications like VR or AR, the hand interacts with virtual objects and performs many manipulative tasks. To understand the behavior of hand motion in these scenarios a system can have robust pose estimation even under the occlusion required. Also, there are use cases where to understand human behavior when they are interacting a robust hand keypoints detection system is required. Moreover, the accuracy of hand keypoints detection systems is not accurate when the fractional area occupied by the hand is small. And, the detection of a small object in computer vision is a major issue which is needed to be solved.

In summary, this dissertation presents algorithms for detecting hand keypoints and addresses some of the challenges in this area. The goal of the research was to improve a system that would allow visually impaired individuals to interact with computers in the author's laboratory. If the solution presented here can be used for this purpose, the author will consider it a modest success.