Hand Gesture & Human-Drone Interaction

Bilawal Latif, Neil Buckley and Emanuele Lindo Secco

Robotics Laboratory, School of Mathematics, Computer Science & Engineering, Liverpool Hope University, Hope Park L16 9JD, UK
20210503@hope.ac.uk, bucklen@hope.ac.uk, seccoe@hope.ac.uk

Abstract. Human computer Interaction is a wide domain which includes different ways of interaction i.e., using hand gestures and body postures. Gestures Detection relate to non-verbal ways to deliver information to the system for control. The aim of gesture recognition is first recording the gestures and then these gestures are read and interpreted by a camera. Gesture recognition has wide range of applications. It can be used by disabled persons to communicate.

This paper focuses on detailed research of controlling drones with hand gestures. The presented system is made of three main blocks i.e. (1) the detection of gestures, (2) translating the gestures and (3) controlling the drone. Deep learning algorithm is used in the first module to detect the real-time gestures of hands. Secondly gesture translator uses some image processing techniques to identify gestures. Control signals are then generated for the drone. Third part shows the implementation of the algorithm using Tensorflow. The accuracy of system is 95.7%.

Keywords: hand gesture, human robot interaction, human drone interaction.

1 Introduction

1.1 Background

Human uses different ways to communicate with machines and some common forms of communication are by body postures. Drones are widely used in most of the applications i.e., coverage of sports event, ariel photography, fast way of transporting equipment to emergency area. Nowadays researchers are planning to use it as a mode of transportation. Much research had been brought forward in finding the most precise way to interact with drones such as, for example, through Hand gesture-controlled [1].

Human-Computer Interface (HCI) can be referred as methods to interact with the machines. The basic example of interacting with machine is a keyboard and a mouse that are used to give input to a personal computer. Advancements in HCI has created interest in researchers. The most significant beliefs in HCI are usability and functionality [2]. Functions are services or tasks offered by a system while usability will be using the function appropriately.

The increasing growth of drones has prompted researchers to establish a new field of study known as *Human Drone Interaction* (HDI). Without a remote controller, it was previously difficult to interact with drones [3-6]. Many studies suggest that drones can be controlled by gestures and postures. In some of the experiments that controlled drones by gesture, a front camera connected to the drone was utilized, and in others, a camera was used to control drones from a ground station.

1.2 Problem Formulation

Drones' fast growth and expansion make them an attractive topic for researchers in a variety of fields, whether for commercial or personal use. Most of the current study has been conducted utilizing third-party data. Many research organizations have also been aroused in controlling the drone with different and efficient means. The author in [2], for example, uses Kinect camera to detect the body and gestures and then translating the gestures according to their requirement and send to drone which was connected to microprocessor via a wi-fi communication protocol, which finally connects to a Leap Motion controller. Leap motion controller do track and recognize gestures using its cameras and infrared LEDs.

On the other hand, author in [3] uses different technology, such as a front camera. The images taken from the video streaming of the camera were then processed. In other literature contributions, such as in [4], for example, a Kinect camera to detect gestures and gestures are predefined. For the hand gesture detection, a lot of work has been performed by Nvidia as well: they use depth, color, and IR sensor to gather the data [5].

1.3 Gesture Recognition

Usually, one human gesture has many meanings: when we raise our hand above head or wave the hand some people are unclear about what we mean and take it as stop. This problem is not only with gestures, but it also affects written and speaking languages.

Focusing on the human upper limb and, in particular, on the human hand, we could provide the following definitions: a movement of the hand could be defined as an *hand motion*, whereas single pose of hand is called *hand posture*. We will focus on the latter one and try to provide an overview on how we can detect and monitor hand posture in terms of the available technologies and software.

Earlier many of researchers are attracted towards senor-based hand glove. This glove consists of sensors that is used to detect the motion of fingers and hand. Now a days vision based is new concept that is in attention of many researchers it is simply defined as to detect motion by using camera.

There are two main techniques of vision based i.e., *model based*, and *image based*. Model based technique is making a model of human hand and then using that for recognition while in image-based approach they use camera to capture the image and then recognizing the human gesture by certain algorithms.

1.4 Devices for Gesture Recognition

Hand gesture is key for the interaction between human and computer and is very convenient. The first section gives detailed discussion about different types of data gloves and their functionalities. The second section gave a bit touch to the image processing techniques.

Wearable Glove

Wearable gloves have been designed since 1970. Each glove has its own special capability and functionality. For example, Sayre Glove was the first ever invented hand glove for detection of hand gestures. Data entry glove was then introduced in back 1980 which is used to enter data into computers.

Data glove has given birth to hand sensor recognition techniques. Many researchers think that sign language is inspired by gestures, and it can be used to interact with computer. Data glove consists of some sensors, wires attach with it. The position of hand (open or closed) is determined by resistive sensors at joints. It detects he joints are straight or twisted. These data are conveyed to computer and then interpreted to information. The advantage of these devices are they do not need many resources and limited processing power. It is bit difficult to manage because of huge number of wires otherwise it's a worthy invention in back 90's [7].

After advancements in technologies, wireless sensors that transmit the data to the computer wirelessly have been also introduced.

We may classify two main types of data glove i.e., active, and passive data gloves. The glove with many sensors to monitor finger movement and accelerator and sends data to computer is an active glove. Gloves with marker of colors on it is passive glove with no sensors [7].

MIT AcceleGlove

In the gloves' context, MIT gloves represent an interesting solution since they have wide capabilities as compared to other systems. The MIT glove was developed by AnthroTronix, a MIT company. It is reprogrammable glove and user can reprogram it according to its need and usage. It is widely used in sports video games etc. An accelerometer sensor is placed under each fingertip and at the back which detect the position of the finger in 3D space and then predict the motion with reference to default position. It is a very user-friendly device where users can do their tasks after wearing it on their hands.

Other devices

There are also many other gloves as well that are widely used, such as, for example the Cyber Glove III (and Cyber Glove II), the 5D sensor Glove, the X-IST Data Glove and the P5 Glove. These devices well represent a mix of methods for gesture recognitions that are widely used before the introduction of vision-based methods.

1.5 Algorithms for Gesture Recognition

Hand Gesture recognition is a demanding task and – from an image processing viewpoint - a crucial computer vision task. Detection of hand from a congested scene is moreover a challenging task as every human skin has its own color type and many diversities may occur in the scene. Therefore, to detect the hands present in each frame, we can use various methods, with extreme precision in identifying the hands. It is also important to mention that it is challenging to get high accuracy in real time behavior.

Some methods of hand detection using camera interface are based on the use of:

- Artificial Neural Networks
- Fuzzy Logic
- Genetic Algorithm

These methods can be combined with

Sensor Glove based hand detection - Gloves consists of multiple sensor that detect the position and motion of the fingers and palm of hand. These techniques are very accurate and easy to use. The connection of sensors with computer is very complex and sluggish. This system itself is not cost effective

Color Marker based Glove - This technique used color markers mark on the glove which gave separate colors to the palm and fingers. Extracting the geometric features gives the actual shape of the hand. This approach is not so costly and is very straightforward as well but its not the feasible interaction with the machine.

Appearance Based - Fingertip technique is widely used as a technique in image generation. Here, for example, Nolker [8, 9] proposed a system called GREFIT which generates the image of hand using the fingertip. In this study the author shares some important points for locating fingertip i.e., using different images for prototype and by marking fingertip as colored. Most of the authors have propose the study that reconstruct the hand with the help of fingertip, contour, and vectors.

Skin color Thresholding - The most basic way to detect the hand is using color range thresholding. Setting the color range of human skin all the elements other than color range should be removed and only color ranged object remains. Some geometric calculation is applied on the extracted hand to extract the fingers. The method fails for some systems due to many reasons. For example, the skin color range for humans are different (i.e. the objects with human skin color also exists in the picture and therefore it is difficult to extract the hand from the image); environmental changes also effect the skin color, and corrupts the whole perception; the hand is placed in front of object with the same color range.

1.5.1. Hand Detection using Deep learning

One of the most accurate way to detect the hand is applying deep learning techniques. Many researchers have been working in Computer vision domain in deep learning and many studies has been abrupted. According to some researchers some architectures give good accuracy in image processing, such as, for example, AlexNet [6], VGG [10] and ResNet [11]. There are network architectures that can work as 10 detectors, but they differ in speed and accuracy. Most of neural networks are very accurate but cannot be used in real time. We need high accuracy in real time and therefore YOLO [12] and SSD [13] can be used as a solution to these problems.

1.5.2. Hand Detection using Tensor Flow

TensorFlow is a framework defined by Google Inc to ease the implementation of machine learning models and to optimize the training algorithms [14]. TensorFlow offers a wide range of operations, from numerical computations to neural network components. TensorFlow is a backend library which is use as a base for *Keras library*. It allows developer to create ML applications by utilizing different tools, Libraries, and resources. Keras is basically an API which is built over TensorFlow which ease the complex commands and instruction of TensorFlow. It eases the test train and save the CNN model.

Tensor Flow design also enables simple compute application over a wide range of platforms. It permits to define flow graphs and topologies to indicate the flow of data over a graph by recognizing inputs as a multidimensional array. It supports on designing a flowchart of processes that may be done on these inputs, which goes at one end and returns as output. The TensorFlow architecture is basically organized in three parts, namely

- Preprocess the data
- Build the model
- Train the model

Graphs contain multiple nodes and each node act as a calculator. All the calculators are connected to each other by stream of data packets. Data path is then set by these calculators and stream and the Mediapipe is built on three different models

- Evaluating the performance
- Sensor data gathering framework
- Component collection

Mediapipe have built-in models and are ready to use. Developers can amend it according to their need and modify it accordingly. Hand detection is carried out very smoothly and easily without consuming many resources. Previously real time object detection with a camera at 30 fps with limited resources is not possible but Mediapipe

achieve this by tracking and detecting in parallel. Mediapipe detects the hand and its key point.

Based on this background, we selected TesnorFlow as our tool in order to detect human hand gesture and apply it to our system.

Figure 1 shows detected hand. This shows all the key points on the hand. To detect the hand in real time Mediapipe used single shot detector. First this module is trained by palm detector model as it is easy to train palm. Furthermore, fingers and joints are detected.

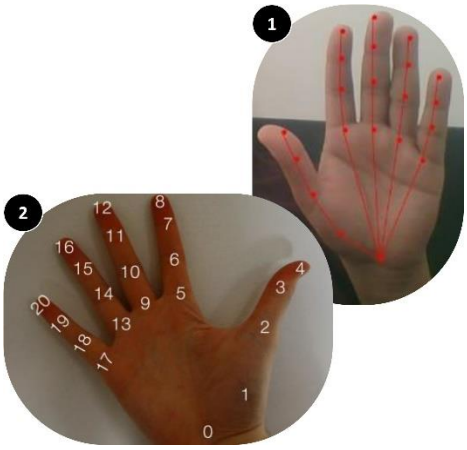

Fig. 1. The hand posture detection by using the Mediapipe library and the 20 key points of the hand on panel 1 and 2, respectively.

2 Implementation

This paragraph provides an overview of HCI and the background of gesture detection, and its types. Here we provide an overview of hand gesture detection and its many forms, as well as the various cameras used for 2D and 3D pictures.

A gesture is a nonverbal means of communication used in HCI systems, according to one basic definition. The primary goal of gesture recognition system is to create a system that can recognize human motion and utilize them to transmit information and establishing interface between user and machine.

HCI has recently grown in importance as its use expands over a variety of applications, including human movement tracking. It must first establish the concept of human motion acquisition, which is the recording of a human or an object's motions and transmission of those movements as 2D or 3D information. Developing a 3D digital image requires the use of software and technologies that are deemed proprietary to such organizations [9, 15]. One of the key aspects is the synchronization between the technology and the actual world, which guarantees that

the system uses the human body motion while adhering to real-world standards and presenting information in a simple and reasonable sequence.

The techniques used and some of the vision-based gesture detection

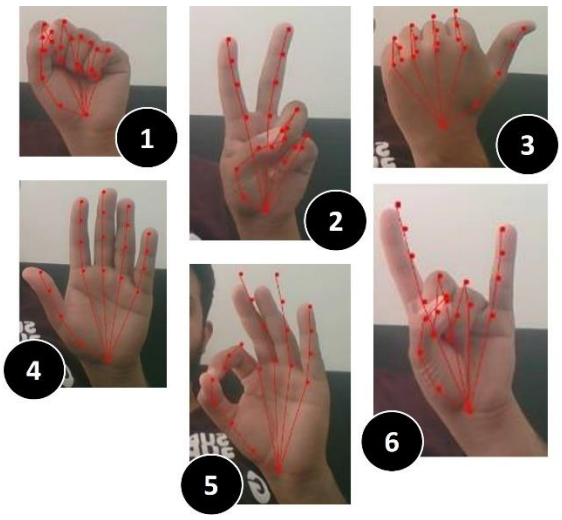

Fig. 2. Fist gesture (1-flying drone), peace gesture (2-moving forward), opposite thumb (3-turning right), open hand (4-landing drone), thumb and index (5-turning left), rock and roll (6-moving backwards)

2.1 Image Processing

TensorFlow provides many libraries that already have some trained models. Mediapipe, in particular, has the trained model of the hand and some of the gestures are recognized in it as well. Therefore we integrate all the libraries according to requirement and how to use and access the functions of TensorFlow and Mediapipe. Mediapipe had already Tensorflow pre trained model saved. And we just must load that model. OpenCV module will allow us to read frames from the camera over which we will perform landmark estimation this functionality is offered by Mediapipe module and then we must convert image into RGB. The function takes the input in RGB format. After that we must predict the gesture by calling a function by Keras library. Mediapipe performs the SSD at the backend. Landmarks are basically the key-points on the hand that tells the actual posture of it and track it. After that we must set the output from a file on getting the output we gave signal to the drone i.e., move forward, move backward, takeoff, land etc.

2.2 Mapping hand gesture with the drone behavior

In order to have the drone flying, some conditions have to be met by the drone and if one of those condition failed drone will not accept the arm command. If drone is

ready and user doesn't respond any gesture, then it will set guided mode. It will fly at the altitude of 1 m. A fist gesture is applied to takeoff the drone.

- ✓ *Moving Forward* - To move the drone in the forward direction you should use the peace gesture to move the drone in the forward direction. The gesture is parsed by the camera and then passed to the drone controller module.

- ✓ *Turn Right* - To move the drone in the right direction you should use the gesture as shown in Figure 2. The gesture is passed to the drone controller.

- ✓ *Landing* - Gesture is shown in the same figure is used to land the drone. Gesture is parsed and then forwarded to drone controller.

- ✓ *Turn left* - Gesture shown in panel 5 (Figure 2) is used to move the drone in left direction.

- ✓ *Moving Backward* - Gesture shown in panel 6 (Figure 2) is used to move the drone in backward position.

2.3 Code

Importing the necessary package OpenCV, NumPy, Mediapipe, Tensorflow, loading the model from Keras. These are the libraries that are commonly used while doing mathematical analysis OpenCV is used for the computer vision where Mediapipe is the aforementioned library that runs over Tensorflow.

Mp.solution.hand it is the function that performs the hand detection algorithm. So, object is defined. *Mp.hands.hands* is a function that is used for the configuration of model max_num_*hands* means that number of hands we want to be detected so we set it to 1 whereas *mp.solution.drawing.utils* will draw and connect the key points.

Initializing TensorFlow and loading the pre trained models. Opening the file that contains the string data of the name of the gestures that we will perform. *ClassNames* will read and splits all the gestures name in numerical order.

In the first line we created the object *videocapture* and we passed 0 as an argument because if we have got more than 1 camera then we must pass a different value we left it default. Inside the loop we are reading every frame. Similarly flipping and showing the frame on new window.

The basic technique in image processing starts with drawing the landmarks of the object that we must recognize after that those landmarks are passed to the predict function which returns an array as shown in Figure 3 (panel 6). As shown that the *classID* is displayed below the predicted classes which is the index of the gesture. After that taking gesture into the frame.

Fig. 3. Code implementation: (1) importing packages, (2) *Mp.solution.hand*, (3) *TensorFlow* initialization and loading of the pre-trained models, (4) definition of the *videocaputre* object, (5) landmarks, (6) output array of the prediction function

3 Conclusion

Gesture Recognition is widely used in all the necessities from smart home automation system to medical field. Mostly it deals with the interaction of human and machines. We have discussed the evolution of hand detection and shared the study of many researchers.

The block diagram of three main module i.e., hand detector, gesture detector and drone controller. We pass image as the input to the system via camera connected to the device.

The main objective of this report is to propose a robust system that can work with high accuracy in real time. System consists of three main modules.

1. Hand detection
2. Gesture Recognition system
3. Drone controller

First module uses deep learning models and dataset was gathered and model is trained. Mediapipe python library which uses SSD is used to detect the hand. Second module uses TensorFlow library to detect the gesture of hands and it was dynamic system which means if we want to add more gestures, we can add it without retraining

the model. The last module drone controller than takes the signals from the drone and then parsed it accordingly. These three modules interact together very friendly. Deep learning method of hand detection is easiest solution that can replace any method of gesture recognition.

Clearly, in this context, other technologies and approaches may be considered where the end-user interacts with external devices in an intuitive way [16-19].

Acknowledgements

This work was presented in dissertation form in fulfilment of the requirements for the MSc in Robotics Engineering for the student Bilawal Latif under the supervision of N. Buckley and E.L. Secco from the Robotics Lab, School of Mathematics, Computer Science and Engineering, Liverpool Hope University.

References

1. Faa.gov. (2018). FAA Releases Aerospace Forecast | Federal Aviation Administration. [online] Available at: https://www.faa.gov/news/updates/?newsId=89870 [Accessed 27 Sep. 2021].
2. Bin Abdul Mutalib, M.K.Z. (2020). Flying Drone Controller by Hand Gesture Using Leap Motion. International Journal of Advanced Trends in Computer Science and Engineering, 9(1.4), pp.111–116.
3. Brown-Syed, C. (2011). Library and Information Studies and Open-source Intelligence. Library & Archival Security, 24(1), pp.1–8.
4. Cheng, X., Ge, Q., Xie, S., Tang, G. and Li, H. (2015). UAV Gesture Interaction Design for Volumetric Surveillance. Procedia Manufacturing, 3, pp.6639–6643.
5. Molchanov, P., Yang, X., Gupta, S., Kim, K., Tyree, S. and Kautz, J. (2016). Online Detection and Classification of Dynamic Hand Gestures with Recurrent 3D Convolutional Neural Networks. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).
6. Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), pp.84–90.
7. Palacios, J., Sagüés, C., Montijano, E. and Llorente, S. (2013). Human-Computer Interaction Based on Hand Gestures Using RGB-D Sensors. Sensors, 13(9), pp.11842–11860.
8. Sacchi, C., Granelli, F., Regazzoni, C.S. and Oberti, F. (2002). A real-time algorithm for error recovery in remote video-based surveillance applications. Signal Processing: Image Communication, 17(2), pp.165–186.
9. Kofman, J. and Borribanbunpotkat, K. (2014). Hand-held 3D scanner for surface-shape measurement without sensor pose tracking or surface markers. Virtual and Physical Prototyping, 9(2), pp.81–95.

10. Savidis, I., Vaisband, B. and Friedman, E.G. (2015). Experimental Analysis of Thermal Coupling in 3-D Integrated Circuits. IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 23(10), pp.2077–2089.
11. MMBIA IEEE Computer Society Workshop on Mathematical Methods in Biomedical Image Analysis In conjunction with Computer Vision and Pattern Recognition (CVPR) 8-9 December 2001 Kauai, Hawaii, USA http://ipagwww.med.yale.edu/mmbia2001. (2001). Medical Image Analysis, 5(2), pp.171–171.
12. Redmon, J., Divvala, S., Girshick, R. and Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [online] Available at: https://arxiv.org/pdf/1506.02640.pdf.
13. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y. and Berg, A.C. (2016). SSD: Single Shot MultiBox Detector. Computer Vision – ECCV 2016, [online] pp.21–37. Available at: https://arxiv.org/abs/1512.02325.
14. Rampasek, L. and Goldenberg, A. (2016). TensorFlow: Biology's Gateway to Deep Learning? Cell Systems, 2(1), pp.12–14.
15. Ramachandra, P. and Shrikhande, N. (2007). Hand gesture recognition by analysis of codons. Intelligent Robots and Computer Vision XXV: Algorithms, Techniques, and Active Vision.
16. Buckley N, Sherrett L, Secco EL, A CNN sign language recognition system with single & double-handed gestures, IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC), 1250-1253, 2021
17. E.L. Secco, D.D. McHugh, N. Buckley, A CNN-based Computer Vision Interface for Prosthetics' application, EAI MobiHealth 2021 - 10th EAI International Conference on Wireless Mobile Communication and Healthcare,
18. D. McHugh, N. Buckley, E.L. Secco, A low-cost visual sensor for gesture recognition via AI CNNS, Intelligent Systems Conference (IntelliSys) 2020, Amsterdam, The Netherlands
19. A.T. Maereg, Y. Lou, E.L. Secco, R. King, Hand Gesture Recognition Based on Near-Infrared Sensing Wristband, Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020), 110-117, 2020 - DOI: 10.5220/0008909401100117