VACATION RESEARCH EXPERIENCE SCHEME
FINAL REPORT

Student: Tran Quang Huy – 10069275
Supervisor: Nicole Robinson
Topic: Social robot interaction and interactivity through enhanced verbal/non-verbal

Science & Engineering Faculty
Queensland University of Technology

Research Article

Real-Time Hand Gestures Recognition Using Cartesian Geometry and Deep Neural Networks For Human-Robot Interaction

Quang Huy Tran – Vacation Research Experience Scheme
Supervisor: Nicole Robinson
Research Topic: Social robot interaction and interactivity through enhanced verbal/non-verbal communication
School of Electrical Engineering and Computer Science
Queensland University of Technology

Abstract: Nowadays, many robots are in service for the people, such as Pepper, Paro, HSR and so on. To be a good partner, a robot needs to have a better model of feedback, such that it can interact naturally. Moreover, in order to design a robot that can communicate verbally and non-verbally, we need to design the human gesture recognition for the robot. In this research, we proposed a method of hand gesture detection based on extracted features from Deep Neural Networks and Cartesian geometry to recognize different types of hand gestures. With this approach, the robot's response will change to interact with human in the conversation according to the hand gesture recognized.

1. Introduction

For a successfull physical human-robot interaction (pHRI), the capability of a robot to understand its environment is imperative. More importantly, the robot should extract from the human operator as much as possible. A well-known study [1] shows that 93% of the human communications is non-verbal and 55% of this accounted for elements like facial expressions, posture, etc. In this perspective, capabilities like gesture recognition and human behavior understanding may be extremely useful for a robotic system in pHRI scenarios [2]. Gesture recognition is an active field of research in computer vision and is an effective way of communicating with robot [3]. In this paper, we propose a pHRI framework based on Convolutional Neural Networks (CNN) and Cartesian geometry, which enables robot to understand the response given by the human operator in the form of hand gestures, being a natural means of a non-verbal communication for people. Our research focuses on the static hand gestures, particularly between thumbs up and other gestures in this paper.

The paper is organized as follows: background and related work are described in Section 2. In Section 3, we presented the method using Cartesian geometry to classify the hand gesture. A complete model pipeline and experimental results are provided in Section 4. Conclusions are drawn and some future works are presented in Section 5. Section 6 acknowledges for contributors and supporters.

2. Background and Related Work

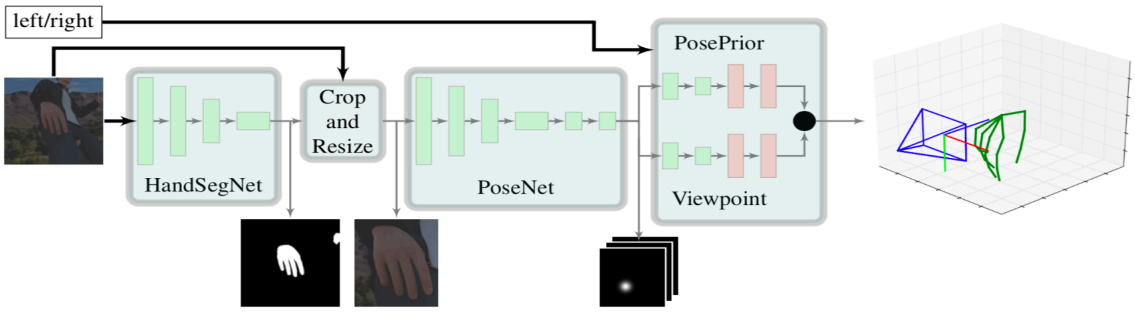

Figure 1: A pipeline for the 3D Hand Pose Estimation consists of three building blocks mentioned below (HandSegNet, PoseNet, PosePrior) [4]

Estimation of 3D Hand Pose:

In this section, we provide a review on the architecture of a hand pose estimation from single color images (RGB images) without the need for any special equipment introduced by C.Zimmerman and T.Brox [4]. The hand keypoints detected by this paper are used as extracted features for the input data of classifier methods. The reason for using this paper method is that the input data is RGB images instead of relying on depth data (RGB-D images) in several works [5,6,7], which are more convenient to test, experiment and develop a small-scale robotics applications. The resulting hand pose estimation yields very promising results, both quantitatively and qualitatively on existing small-scale datasets. This method consists of three deep neural networks (Figure 1) that cover important subtasks on the way to the 3D pose.

Hand segmentation with HandSegNet: The first network (HandSegNet) provides a hand segmentation to localize the hand in the image. The architecture is a smaller version of the network from Wei et al.[8] which is trained on the hand pose datasets. The hand mask output from this network allows us to crop and normalize the input data, which simplifies the learning task for PoseNet.

Keypoint scoremaps with PoseNet: From the cropped and normalized input images from HandSegNet, PoseNet localizes hand keypoints in the 2D images. To explain, PoseNet network predicts 21 score maps, where each map contains information about the likelihood that a certain keypoint is present at a spatial location. The network uses an encoder-deconder architecture similar to the PoseNet network by Wei et al.[8] using the initialized weights and retraining the network for hand keypoint detection.

3D hand pose with the Pose Prior Network: the Pose Prior Network derives the 3D hand pose from the 2D keypoints by estimating the most likely 3D structure conditioned on the score maps. The network architecture (Figure 2) for the pose prior has two parallel processing streams. Two almost symmetric streams estimate cannonical coordinates (w^c) and viewpoint estimation (R) relative the coordinate system. Combination of the two predictions yields an estimation for the relative normalized coordinates (w^{rel}), which is a set of 3D coordinates to learn a translation invariant representation of hand poses.

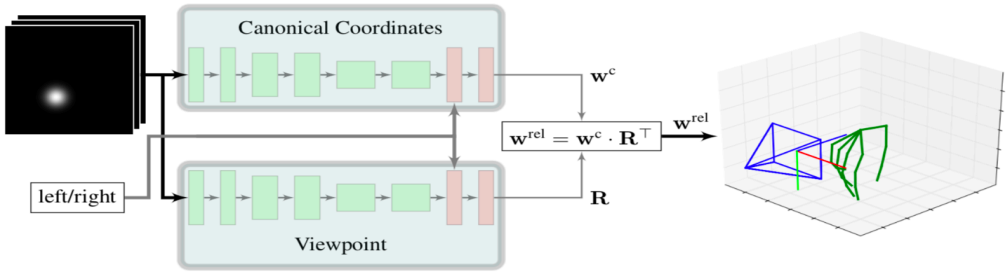

Figure 2: Proposed architecture for the PosePrior Network [4]

Related Work

The idea of using a model to extract the information from data and a method to classify the extracted features has been used widly in the field of hand gesture recognition. Antón-Canalís L. et al. [9] introduced a fast and robust hand pose detector that integrates temporal coherence information and a wrist detector using a continuously operational system. The authors of [10] proposed a real-time hand gesture recognition using hand segmentation using background subtraction and localization of each finger to recognize the gesture. Additionally, the real-time human-computer interaction system based on hand gesture used CNN and Kalman Filter for smoothing hand positions , has been studied in [11]. All of these approaches above have mostly used the image data with the palm hand facing in front of camera, thus limiting the use of recognition capability with multiple views

3. Hand Gesture Recognition

In this section, we present some categories of curl and directional orientation of fingers introduced by P. Prasad [12] that allow generating the appearances of the hand poses in custom configurations. With these categories, we present a mechanism using the Cartesian geometry to classify the pose based on 21 landmarks in 3D detected by hand pose estimation. This method absolutely requires no training data and no networks involved, thus serving a real-time application with reducing the computation time.

3.1. Categories of Curl of Fingers and Directional Orientation

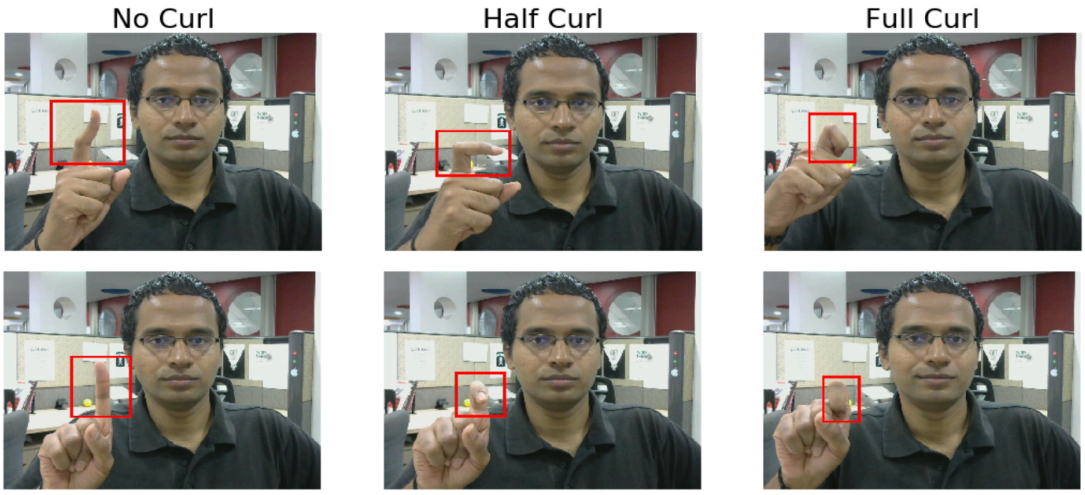

Figure 3: Curls from front and side view [12]

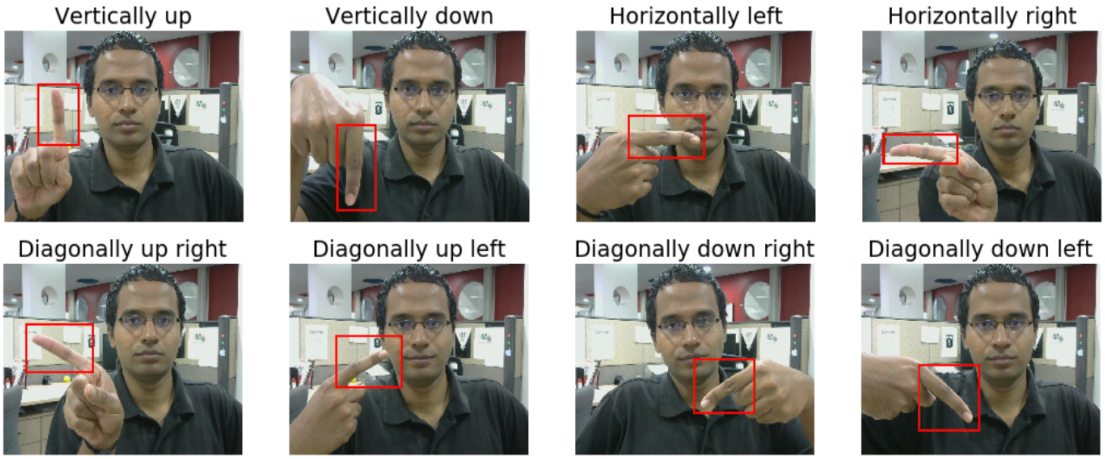

Figure 4: Different directional orientation of finger [12]

As shown in Figure 3, curls of each hand finger has been categorised into three different types: no curl, half curl and full curl. Similarly, in directional orientation of fingers, there are six different types illustrated in Figure 4. These categories mentioned can be use to define a particular hand gesture and each finger can have many predefined curls and directional orientations . In case of thumbs up gesture, the thumb finger can be recognized with no curl and vertically up, diagonally up left or diagonally up right orientation while other fingers have full curls and horizontally left or horizontally right orientations.

3.2. Classify Hand Pose

Figure 5: Types of defined hand gestures

Having those categories, six different types of hand gestures illustrated in Figure 5 are clearly defined. Defining lots of hand gestures is intended work efficiently in case of binary classification between thumbs up gesture and not thumbs up gesture. Additionally, in order to improve the tolerance in distoritions of hand gesture during prediction, each of predefined hand gesture pose has a certain confidence level for a particular curl and directional orientation.

Based on the keypoints provided by hand pose estimation model, the Cartesian geometry is used to calculate the angles of keypoints to predict the type of curl and directional orientation for each hand finger. If the predicted category of is included in any given hand gesture, the finger score is obtained weighted by the confidence level. The total score of is the sum of all finger scores, the highest total score predefined pose. To get the better result for hand gesture recognition, a thumbs up threshold level is needed. If the score is above this threshold, the predicted pose is highly likely to be the pose of the input image. However, if the score is below the threshold level, the gesture output is estimated to be one of other gestures.

4. Methodology & Exerimental Results

4.1. Methodology

The proposed pHRI pipeline for hand gesture recognition which integrates the 3D hand pose estimation and Cartesian geometry mentioned above is illustrated in Figure 5.

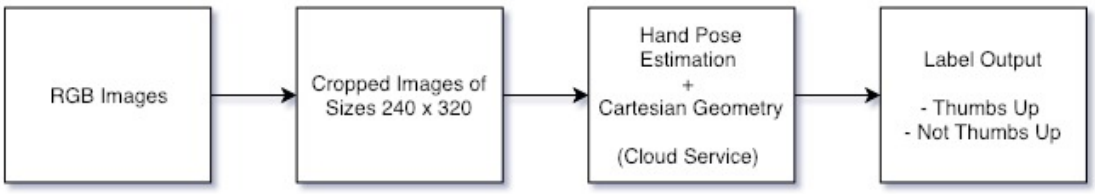

Figure 6: A complete pipeline for hand gesture detection

As shown in the figure, the hand pose estimation and Cartesian geometry are integrated into a container for cloud deployment using Docker and Tensorflow backend to achieve a real-time application. Having this cloud service, the predicted label output for hand gesture can be obtained by calling API request to the model. Additionally, the input size of this cloud model requires single color images to crop the size into 240 x 320 before being processed further. Finally, the label output shows the predicted result, which is either thumbs up or not thumbs up in our research.

4.2. Experimental and Results

A. Datasets: In the experiments, a dataset of hand gestures are used to evaluate the perfomance of the proposed method. The dataset is included an RGB image collection of thumbs up gesture and other random hand gestures with ground-truth labels. The dataset consists of 50 images which are limited to one hand per image and collected online for hand gesture recognition. Sample of the dataset are shown in the Figure 7. Both data sets are challenging for hand gesture recognition with variations in hand orientation, scale, articulation, etc.

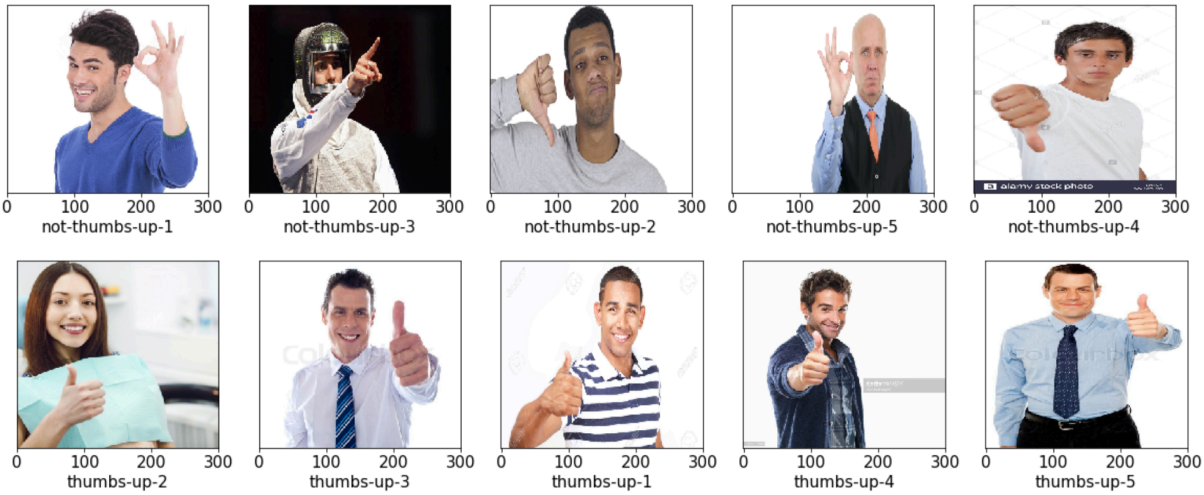

Figure 7: Sample of testing datasets

B. Performance Evaluation on Datasets

Figure 8: Result of sample tesing datasets

In order to measure the performance of the proposed hand gesture recognition method, the binary classification accuracy is evaluated in the experiments. Using the methodology described, he results of sample datasets can be obtained and are demonstrated in Figure 8. The accuracy of dataset is about 80%, showing a good result in determining whether the image consists of thumbs up gesture or not. It is obvious that the proposed method is highly efficient and can meet the requirements of the real-time applications. However, most of the wrong predicted labels are thumbs down gesture, which has a similarity in 3D shape with thumbs up gesture. Therefore, this model does not robustly work well in terms of distinguishing between them.

5. Conclusion and Future Work

A method for hand gesture recognition with a real-time application is introduced in this paper. The hand keypoints in the image ar detected by hand pose estimation network. Then, the Cartesian geometry is used to classify the hand gesture followed by predefined curls and directional orientations.

The experimental results show that our approach yields a very promising result on a small-scale dataset and is fit for a real-time applications. Despite that, the model still misclassify the thumbs down gesture as the thumbs up gesture, demonstrating a point to improve in the future research.

As the performance of the proposed method highly depends on the consistency of image background, the position of hand gesture and lighting conditions plus with using the simple rule classifier, this method is not robust to detect the hand in many real-life scenarios. Therefore, future research is focused on collecting training datasets and using robust classifier methods such as neural networks or support vector machine (SVM) in order to develop a gesture-based human-robot interaction system.

6. Acknowledgements

This work is supported by the technical leader Gavin Suddrey (School of Electrical Engineering and Computer Science – Queensland University of Technology) in deploying the hand gesture recognition model on the cloud with a web interface demo.

References

[1] A. Mehrabian. *Nonverbal Communication*. Aldine Publishing Company, 1972.

[2] G. Canal, S. Escalera, and C. Angulo. A real-time human-robot interaction system based on gestures for assistive scenarios. *Computer Vision and Image Understanding*, 149:65-77, 2016.

[3] Hongyi Liu and Lihui Wang. Gesture recognition for human-robot collaboration: A review. International Journal of Industrial Ergonomics, 2017.

[4] C. Zimmerman and T. Brox. Learning to Estimate 3D Hand Pose from Single RGB Images. In *International Conference on Computer Vision*, 2017.

[5] X. Zhou, Q. Wan, W.Zhang, X. Xue, and Y. Wei. Model-based deep hand pose estimation. pages 2421-2427, 2016.

[6] M. Oberweger, P. Wohlhart and V. Lepetit. Hand deeps in deep learning for hand pose estimation. *arXiv preprint arXiv: 1502.06807*, 2015.

[7] N. Sarafianos, B. Boteanu, B. Ionescu, and I.A.Kakadiaris. 3d Human pose estimation: A review of literature and analysis of covariates. *Computer Vision and Image Understanding*, 152:1-20, 2016.

[8] Shin-En Wei, Varun Ramakrishna, Takeo Kanade and Yaser Sheikh. Convolutional Pose Machines. In *Proc. of the IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 4724-4732, 2016.

[9] Antón-Canalís L., Sánchez-Nielsen E., Castrillón-Santana M. (2005) Fast and Accurate Hand Pose Detection for Human-Robot Interaction. In: Marques J.S., Pérez de la Blanca N., Pina P. (eds) Pattern Recognition and Image Analysis. IbPRIA 2005. Lecture Notes in Computer Science, vol 3522. Springer, Berlin, Heidelberg

[10] Chen, Zhi-Hua & Kim, Jung-Tae & Liang, Jianning & Zhang, Jing & Yuan, Yu-Bo. (2014). Real-Time Hand Gesture Recognition Using Finger Segmentation. TheScientificWorldJournal. 2014. 267872. 10.1155/2014/267872.

[11] Pei Xu. A real-time hand gesture recognition and Human-Computer Interaction System. *arXiv: 1704.07296*

[12] P. Prasad. Prasad9/Classify-HandGesturePose: Classification of Hand Gesture Pose Using Tensorflow. https://github.com/Prasad9/Classify-HandGesturePose