# Parsing the Hand Gesture of Traffic Police Officers by Using OpenPose

Maogen Fu

*University of Edinburgh, School of Informatics, Edinburgh, Scotland, United Kingdom, EH75EA*
**Corresponding author. Email: S2102187@ed.ac.uk*

**ABSTRACT**
Nowadays, as self-driving cars become more and more popular, more features should be added to the whole function of self-driving cars. One of them should be the ability to parse the hand gestures of traffic police officers. To fully bring self-driving cars to reality, such a feature should be discussed and considered in their functionality. In recent decades, real-time 3D hand pose estimation combined with depth cameras has contributed to the successful launch of virtual reality and augmented reality applications. This essay focuses on using OpenPose to parse the hand gestures of traffic police officers based on the expected functionality of self-driving cars. Based on the given background, this article will use comparison to separate the traditional methods and the new method. The result generated by OpenPose is faster and more accurate (99% accuracy), so that it can be a choice for the future algorithm of recognition of hand gesture recognition of traffic officers in self-driving cars.

***Keywords:*** *Hand gesture parsing, traffic hand gesture, 3-D movement Tracker, OpenPose.*

## 1. INTRODUCTION

In reality, traffic signs are not the only indicators for cars to move. However, the human factor should be considered—the hand gestures of traffic police officers. Nowadays, the mainstream method is to use depth image to parse hand gestures. The problem that scholars aim to figure out is how to keep higher efficiency while not losing accuracy during the process of parsing hand gestures. In a real-time scenario, there are more factors to take into account—the whole environment, pedestrians, and the useless hand gestures of targets. So, a real-time tracker of hand gestures is essential. However, much of the research in this field mainly focus on the theoretical aspect of parsing hand gestures and this article will mainly concentrate on the realistic aspect of parsing hand gestures. The current studies of hand gesture parsing are mainly focused on the parsing of static pictures and transforming them into forms that can be analyzed by a computer directly or through a machine learning model. There are two branches: firstly, transforming images into depth images that can be directly analyzed by machine learning; or using new introduced method such as 3-D tracker to capture movements.

In this article, it explores one of the most important functions of self-driving cars and compares 3 methods to implement such a functionality. Meanwhile, it discusses the possibility to utilizing a 3-D real-time tracker to analyze the hand gestures received by devices in self-driving cars. This article provides a new idea to deal with the realistic problem of the implementation of self-driving cars and could contribute to an improvement of the algorithm that parses the hand gestures that can affect self-driving cars' movement.

## 2. TRADITIONAL DEPTH IMAGE TRANSFORMATION

A complex topology of the hand skeleton with high similarity among fingers and a small size could pose challenges in parsing hand gestures. The traditional method of parsing hand gestures is to transform the hand gesture into a depth image. The way to transform hand gestures into depth image in this article would be by using a kinect sensor. Kinect sensor can directly extract color information and depth information flow. Meanwhile, combining two information flows, the required palm image can be drawn out from complex background and figure 1 shows the structure of kinect sensor [1].

 

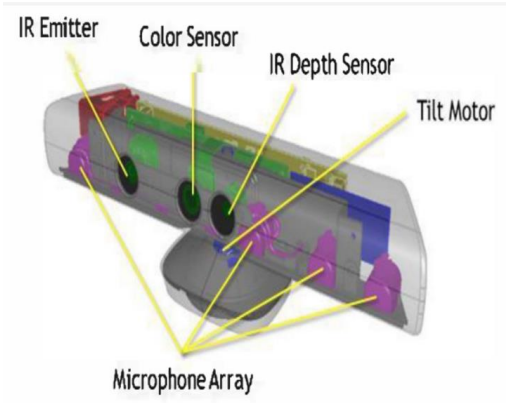


**Figure 1** Kinect hardware structure

Color information flow can be set with several levels of resolution ratio and distinct formats. High resolution ratio can provide more information, while the update rate is low [1]. In the next operation, the palm will be roughly segmented based on depth information we get before. Spatial-temporal filtering (STF) [2] can be utilized to track the hand position. With this tracking result, the initial hand depth will be automatically marked. Then, the initial depth is used to calculate the depth threshold value, depending on which the palm images are segmented. Moreover, the initial edge of the hand gesture can be achieved using Sobel method [3]. The next step is to employ a model called Superpixel-Markov Random Field (SMRF) to enforce the spatial smoothness and the label co-occurrence prior to remove the misclassified regions [4].

Although the best average accuracy of hand gesture measured by per-pixel is 90.37% [5], since those depth images are segmented from a set of hand gestures, there is a problem of parsing dynamic hand gesture in our real life.

What the paper needs is to find out a model/algorithm to figure out the best time efficiency and accuracy of a combo of hand gesture, or dynamic hand gesture. Obviously, the traditional method of transforming hand gestures into depth images cannot satisfy the article's requirements and this paper needs a method of real-time tracker. Moreover, since the method is transforming the image captured by the sensor into a depth image, there will be some distractions that will result in loss of depth information. So, a more accurate and real-time tracker for dynamic hand gesture is needed.

## 3. REALSENSE DEPTH IMAGE TRANSFORMATION

The problem the method shall resolve is that with staying the high level of efficiency and accuracy, the method could extract key information from hand gesture. The method can provide a more accurate way to deal with the depth image drawn from a new device and get a better analysis of hand gestures.

### *3.1. The way the method solves the problem*

The way the method provides to solve the problem is to use RealSense Front-Facing camera SR300 and transform the image to depth image. However, the problem is that it may lose key information from different hand gestures. As a result, a mechanism should be considered that can capture the complete hand gesture and finish further analysis.

The RealSense Front-Facing camera SR300 is a cheaper but more efficient device on detecting hand gesture. The fundamental working scheme of RealSense camera is quite similar to the traditional depth image transformation. What has been different is its ability to gain pixels that are not one-one in color images [6]. This property limits the use of traditional depth image processing algorithm and a new algorithm based on RealSense camera can be implied.

A double-channel CNN algorithm is employed with RealSense camera and it can better combines depth information with color information for improving the recognition accuracy [7]. The figure 2 below shows how the whole system works to parse hand gesture [8]. By using both depth-based CNN and RGB-based CNN, the final result of parsing dynamic hand gesture can be incredibly accurate—the average recognition accuracy on 16 kinds of gestures reached 99.8% [9].

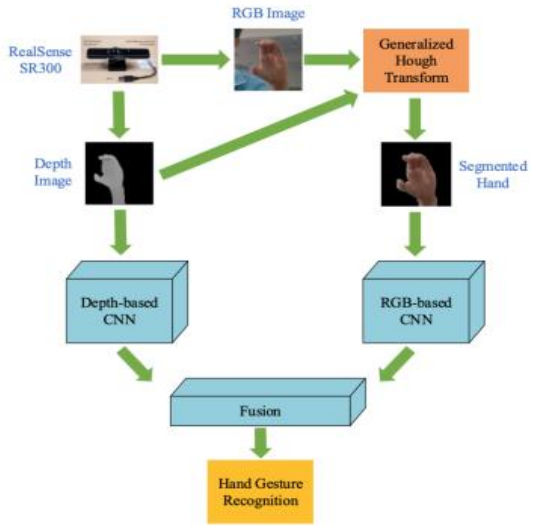


**Figure 2** General framework of hand gesture parsing

### *3.2. Limitation of the method*

The limitation of this method is the difficulty to deal with dynamic hand gestures. The reason is that RealSense camera can only work on the static images of hand gesture. The results show a good trend for RealSense camera to process static hand gesture to a better depth image that can make the whole recognition

process better. The problem left is that RealSense camera is still far from reality, which means that it is not possible for realistic scenarios such as hand gesture of traffic police officers.

## 4. REAL-TIME 3-D TRACKER

### *4.1. Method*

The new introduced method is supported by OpenPose and it can help researchers solve problem of real-time 3-D dynamic movement. Therefore, it is possible for self-driving cars to recognize the hand gesture of traffic police officer by using this new model to deal with dynamic hand gesture.

The new method utilized a library called OpenPose. OpenPose is a real-time human pose detection that can jointly detect the human body, including hand gestures on single images [10]. By using this new library, there could be new generated features that can help us in dynamic hand gesture recognition: OpenPose can do both real-time 3D and 2D key points detection; single-person tracking that can facilitates detection; and a Calibration toolbox for the estimation of extrinsic, intrinsic, and distortion camera parameters [11].

### *4.2. Working mechanism*

The working mechanism of OpenPose is that OpenPose library initially can pull out essential features from a picture using the first few layers and those features can be put into two parallel divisions of CNN layers. The first division predicts a set of 18 confidence maps. The next branch predicts another set of 38 Part Affinity Fields (PAFs) that denotes the level of association between parts [12]. The next step is to clean the predictions made by the branches and with confidence maps, bipartite graphs are made between pairs of parts. By calculating PAF values, weaker links are pruned in the bipartite graphs. Combining those procedures, human pose skeletons can be estimated and allocated to every person in the picture.

### *4.3. Data*

The basic model this article used here is OpenPose and the dataset is chosen from online sources, which contains 10 videos of Chinese traffic police officers' hand gestures. Those videos are consistent of nearly 1000 pieces of pictures if taking each movement as a picture. The purpose is to check whether those pictures can be successfully defined by critical points that are generated by OpenPose and further analysis can be applied to those critical points. Running the dataset with OpenPose, a process in which OpenPose splits a captured picture into 38 critical points is shown. Those critical points can be then dircetly used by further analysis.

In terms of gesture data set, this article uses the Chinese traffic police gesture data set. The Chinese traffic police have eight gestures: stop, go straight, turn left, wait left, turn right, change lanes, slow down and stop. The data set is divided into two parts: training set and test set. The training set contains 11 videos, a total of 72843 frames, including 1789 groups of actions. The test set contains 10 videos, a total of 61581 frames, including 1565 groups of traffic police gesture actions. The total time is up to 2 hours, and each video lasts about 6 minutes on average, The video frame rate is 15fps, the aspect ratio is also fixed at 1:1, and the resolution is 1080 * 1080 except that the 003.mp4 video in the training set is 540 * 540.

The shooting environment of the video is rich and diverse, including indoor scenes such as classrooms and laboratories and outdoor scenes such as parks, campuses, roadsides, parking lots and woods. The lighting environment is complex and changeable, and the characters and their clothes are different. Many people wear black sportswear or traffic police reflective vests, traffic police hats and white gloves to participate in the recording. There are 3354 groups of traffic police gesture actions in the data set. The interval between each two groups of different gesture actions is at least 2 seconds. During the interval, the traffic police are standing at attention. At this time, the traffic police are idle, and the data set of the corresponding frame is marked as '0'.

### *4.4 Training and Result*

**Figure 3** Dynamic hand gesture parsing

After running OpenPose with a data-set that contains 10 videos of hand gestures of Chinese traffic police officers, it shows that OpenPose can successfully get all the key points of the dynamic hand movement as well as body movement. It transformed dynamic hand gesture to a set of images that contain key points that can be recognized by computer through further deep learning or analysis. The following Figure. 3 shows the key points recognition and the accuracy is nearly 99% for key points recognition. After key points recognition, OpenPose can combine those single pictures with key points into a consecutive picture and then generally analyze the hand gesture as well as body movement. As the hand gestures changes, OpenPose can still capture pictures and transform them to pictures with critical points.

## 5. DISCUSSION

The result of openpose is expected because it does not convert the target into a depth image (which will lose accuracy in the process). Openpose can divide the target into recognizable key points. Those three methods have their own advantages. Traditional depth image transformation has a high accuracy when parsing hand gestures and it is easy to implement its model by a kinect sensor. RealSense camera can provide a better depth recognition (99.7% vs 90.37%) but it cannot be a mainstream method for dynamic hand gesture parsing. OpenPose is a library that provides a way to parse dynamic hand gesture by key points transformation of human body and movement and then do key points analysis to parse those movements or hand gestures. The recognition of hand gesture of traffic police officers is a real-time problem and those hand gestures should be dynamic and flexible. That is why this paper needs a light network to deal with those movement. With OpenPose, a crucial and essential functionality of self-driving cars may be successfully implemented.

## 6. CONCLUSION

Compared with two methods (traditional depth image transformation and RealSense depth image transformation), OpenPose has an advantage in dynamic hand gesture parsing by applying two branch CNN to jointly predict human body movement. The results of accuracy show that under experimental circumstances, OpenPose can be equipped with high efficiency and accuracy. In the future, it is important to explore the possibility of implementing OpenPose as the main supporting library for self-driving cars to parse hand gestures of traffic police officers. Future work will mainly focus on the tests that whether OpenPose can still recognize key points in different cultural backgrounds. Also, there would be a larger sample of data-set to ensure the robustness of OpenPose.

## ACKNOWLEDGMENTS

I want to show thanks to Henry Zhang, one of my best friends at the University of Edinburgh. He helped me run the OpenPose model and gave me several ideas on how to interpret those results. Also, thanks to my tutor, Ruyi, for her brilliant guidance on my article.

## REFERENCES

[1] S, Kajal. Kinect Sensor based Object Feature Estimation in Depth Images. International Journal of Signal Processing, Image Processing and Pattern Recognition Vol.8, No.12 (2015), pp. 237-242

[2] H, Liang. Parsing the hand in Depth Images. IEEE TRANSACTIONS ON MULTIMEDIA, VOL.16, NO.5, 2014, pp 1241

[3] J.Lin, W.Ying, and T.S. Huang. “Modeling the constraints of human hand motion” in Proc. HUMO, 2000

[4] H.Liang, J Yuan, D.Thalmann, and Z.Zhang, “Model-based hand pose estimation via spatial-temporal hand parsing and 3D fingertip localization”. Visual Comput. J., vol.29, no.6-8, Jun. 2013, pp. 837-848

[5] A.Hernandez-vela, N. Zlateva, A. Marinov, M. Reyes, P. Radeva, D. Dimov, and S. Escalera, “Graph cuts optimization for multi-limb human segmentation in depth maps” in Proc. CVPR, 2012

[6] Bo, L. Hand Gesture Recognition with Generalized Hough Transform and DC-CNN Using Realsense, 2018. https://pure.port.ac.uk/ws/portalfiles/portal/11506771/2018_ICIST.pdf

[7] Q. Gao, J. Liu, et al. “Static Hand Gesture Recognition with Parallel CNNs for Space Human-Robot Interaction” In: Huang Y., Wu H., Liu H., Yin Z. (eds) Intelligent Robotics and Applications. ICIRA 2017. Lecture Notes in Computer Science, vol. 10462. pp. 462-473, 2017

[8] K. Fukushima. “Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position”. Biological Cybernetics, vol. 36, no. 4, 1980, pp. 193-202.

[9] Z.Cao, G.Hidalgo, T.Simon, S.Wei, and Y.Sheikh. PoenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, vol. 43, Jan. 2021, pp. 172-186

[10] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping" in CVPR, 2017.

[11] N. Dinesh Reddy, M. Vo, and S. G. Narasimhan, "Carfusion: Combining point tracking and part detection for dynamic 3d reconstruction of vehicles" in CVPR, 2018.

[12] G. Papandreou, T. Zhu, L.-C. Chen, S. Gidaris, J. Tompson, and K. Murphy, "Personlab: Person pose estimation and instance seg-mentation with a bottom-up, part-based, geometric embedding model" in ECCV, 2018.