Vision-based Hand Gesture Recognition

1st Huan Nguyen Dinh
Advance Program in Computer Science
Falcuty of Information Techonology
University of Science, VNU-HCM
ndhuan@apcs.vn

2nd Duy Ho Cong
Advance Program in Computer Science
Falcuty of Information Techonology
University of Science, VNU-HCM
hcduy@apcs.vn

3rd Khang Nguyen Phuc
Advance Program in Computer Science
Falcuty of Information Techonology
University of Science, VNU-HCM
npkhang@apcs.vn

***Abstract*—Many efforts have been made in increasing Human-Computer Interaction in the 21st century, specifically through voice like Cortana or via touch screen. The goal is to get more people being able to use the power of computer as there are old people that are not familiar with devices like mouse and keyboard. Motivated by the trend, the authors re-implement and modify many vision-based Hand Gesture Recognition algorithms for the purpose of studying the matter. This is a fundamental step before choosing the right tools and finally being able to develop a well fully functional system for Hand Gesture Recognition. Such a system would have a wide range of applications like Sign Language Translator, building blocks for Gesture-based User Interface and be very beneficial to the human society.**

I. Introduction

Computers are powerful tools for humans to use, but we speak different languages. Sometimes it becomes unnecessarily complicated just to tell the computer to do some simple tasks. Also carrying a computer around along with a mouse and keyboard can be very tiresome. This idea has also existed in many since the beginning of the twentieth century. To sum up, the way that we have been using the power of the computer are still very limited and inefficient.

From this point of view, it would be great if we can communicate with the computer in the most natural way possible. Mouse and keyboard are just too rough for old people to get familiar with and there are still a lot of means for communication that humans have been using for thousands of years. For example, through voice, touch or gestures. This is certainly more convenient and more efficient than mouse and keyboard in many aspects. After a lot of great achievements in building Human Voice Recognition Systems like Cortana (Developed by Microsoft) , Siri (Developed by Apple), Google Assistant (Developed by Google) and the market getting flooded with touch screen phones and computers in 2019, many computer scientists are gradually moving toward the field of Hand Gesture Recognition [3] [6] [7].

We humans have been using hand gesture as a form of non-verbal communication for a very long time. A system that is able to recognize hand gestures has a wide area of application in Human Computer Interaction and also Sign Language [8], which can greatly benefit mute people. Systems like this can also make users feel comfortable and make a lot of money.

Two popular known hand gestures are static hand gesture and dynamic hand gesture. The number of static hand gesture is bigger than that of dynamic, that's why we choose static hand gestures to work with. There are three known approaches to this problem: kinetic sensors, computer vision and sensors gloves, with the simplest way to approach this problem is by wearing sensor gloves [9], usually known as the state-of-the-art method. Even though this method has its own advantages and is able to brings good results, one can argue that it can make users feel uncomfortable and awkward [3]. A pair of gloves can also be very expensive. This pretty much explains why this approach is not popular. Kinetic sensors are also very expensive so not many people can afford it.

Noticing this drawback, the authors try to dig in the methods of Recognizing Hand Gestures using Computer Vision. Compared to the state-of-the-art method, this approach has more nasty problems: Cluttered Background, Lighting Condition, Camera Resolution, Skin Color, etc... However, the authors strongly believe that this approach's potential can well reach far beyond what we can imagine and thus, worth the effort in pursuing to the end.

In this paper, the authors try to go through as much methods as possible and try to nudge a bit further to see if the results can be improved. If not, give logical reasons for what might have happened.

As far as the authors know, not many data sets are available, so for the sake of studying the authors create their own data set. Further details on this dataset will be mentioned in section 4 of the paper.

This paper's content is as follows: in Section 2, the authors present the related works, including related or direct solutions of the problem of hand gesture recognition. In section 3, the authors give further details of the methods in the section of related works and their contribution to these method. In section 4, the author present the Experiment Result and details of Data Sets.

II. Related Works

These are the related works that from them the authors re-implement and modify.

A. ColorHandPose3D network

This is the work of Christian Zimmermann et al from the University of Freiburg. ColorHandPose3D is a Convolutional Neural Network estimating 3D Hand Pose from a single RGB Image. This work has been tested on many challenging data sets and has been able to give back good results. Even though

the performance is limited by the lack of pose diversity in the data sets, ColorHandPose3D can still provide a good base for Prasad Pai, a computer scientist at Medium innovation lab to develop a fully functional Hand Gesture Recognition System.

From the return 3D hand pose estimation, he introduces 3 ways to classify hand gestures :

1) Directional Orientation and Curls of fingers: This method required no training nor learning network. Using Cartesian Geometry to estimate curl and direction of fingers and search for the pose that match. There are 3 categories of curl fingers : straight, half curl and full curl. There are 8 categories for direction: vertically up, vertically down, horizontal left, horizontal right, diagonally up right, diagonally up left, diagonally down right and diagonally down left. There is a threshold level for which the confident for each pose below that will be unidentified.

2) Neural Network: The two layers fully connected network is trained.

3) Support Vector Machine: The classifier was able to classify various hand gesture poses.

B. HGR-Net

This is the work of Amirhossein Dadashzadeh et al from the Department of Computer Science, Shahid Beheshti University, Tehran, Iran. HGR-Net is a two-stage convolutional neural network (CNN) architecture. The first stage locate the hand and segmentate it at pixel level. The second stage classify the hand gesture from the segmentation. This is the model of how the system works:

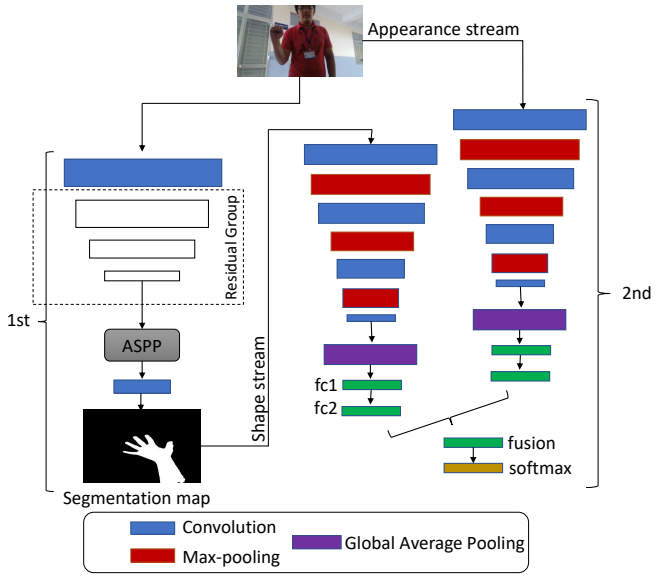

Fig. 1. The proposed HGR-Net by Amirhossein Dadashzadeh et al.

III. Methods

A. Skeleton

1) OpepPose library: The authors use the OpenPose library to collect 42 hand keypoints from input image, each keypoint has coordinate x and y in 2D-dimension [1] [2] [10] [11]. The internal mechanism of the library is then modified so that it can output the result in text files. The output is a long list of coordinates with confident rate for each coordinate. Then the authors try to think of some rules to classify gestures based on the given pairs of coordinate.The author's idea focuses on the keypoints of each fingers and try to specify if a certain finger of the subject in the image is curl, half-curl or straight.

2) ColorHandPose3D network:

B. Segmentation

1) HGR-Net: This is the work of Amirhossein Dadashzadeh et al from the Department of Computer Science, Shahid Beheshti University, Tehran, Iran. HGR-Net is a two-stage convolutional neural network (CNN) architecture. The first stage locate the hand and segmentate it at pixel level. The second stage classify the hand gesture from the segmentation.

2) YOLO: The experiment result tested on the authors' dataset yield high result, around 70 percent up to 100 percent accuracy for each gesture. In the three, using Support Vector Machine provides the highest accuracy, especially for OK pose. After that, we try to extend the numbers of gestures using the Directional way but fail to keep the accuracy consistent. The idea to combine his work with the OpenPose library also doesn't work out because of the complexity of the code. After trying to use a KNN classifier, the result is unsatisfactory.

Finally, from the authors experience, HGR-Net Fusion Network is a simple but efficient framework to work with. The most noticeable thing about this is the hand segmentation model. After re-implement the algorithm on our machine, the authors proceed to modify it so that the system can run and recognize hand gesture in real time at 10fps. The result is also satisfactory with accuracy around 80.4 percents for each gesture.

IV. Experiment Result

As far as the authors know, there are not many publicy available datasets for static hand gestures [4]. So the authors create their own dataset. This dataset contain 6000 RGB images of 10 hand gestures from 3 subjects. Each subject performs 10 hand gestures in 200 RGB images.

Later on, the authors also use the OUHANDS dataset for training and evaluation of the HGR-Net Fusion Network [5]. This dataset contains 10 hand gestures from 23 subjects, and is split into training, validation and testing sets with 1200, 400 and 1000 images respectively. All sets come with corresponding segmentation masks and are captured in very challenging situation. [4]

V. Conclusion

Summary and Open Problems (to be updated...)

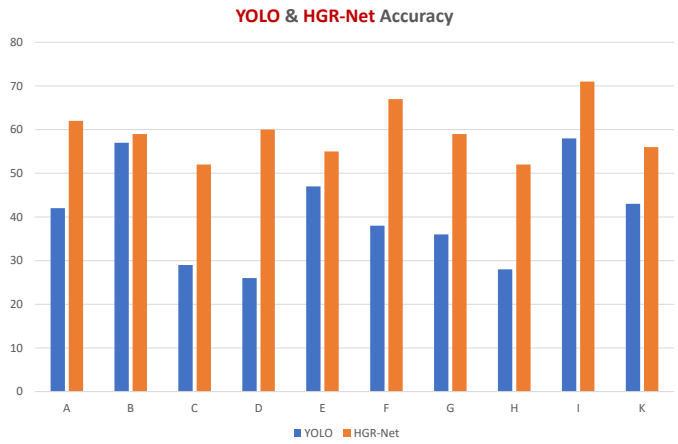

Fig. 2. Comparison of Accuracy HGR-Net YOLO.

REFERENCES

[1] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. OpenPose: realtime multi-person 2D pose estimation using Part Affinity Fields. In *arXiv preprint arXiv:1812.08008*, 2018.

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[3] Q. Chen, N. D. Georganas, and E. M. Petriu. Real-time vision-based hand gesture recognition using haar-like features. In *2007 IEEE Instrumentation Measurement Technology Conference IMTC 2007*, pages 1–6, May 2007.

[4] Amirhossein Dadashzadeh, Alireza Tavakoli Targhi, and Maryam Tahmasbi. Hgr-net: A two-stage convolutional neural network for hand gesture segmentation and recognition. *CoRR*, abs/1806.05653, 2018.

[5] M. Matilainen, P. Sangi, J. Holappa, and O. Silvn. Ouhands database for hand detection and pose recognition. In *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–5, Dec 2016.

[6] J. Nagi, F. Ducatelle, G. A. Di Caro, D. Cirean, U. Meier, A. Giusti, F. Nagi, J. Schmidhuber, and L. M. Gambardella. Max-pooling convolutional neural networks for vision-based hand gesture recognition. In *2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, pages 342–347, Nov 2011.

[7] E. Ohn-Bar and M. M. Trivedi. Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations. *IEEE Transactions on Intelligent Transportation Systems*, 15(6):2368–2377, Dec 2014.

[8] M. Panwar and P. Singh Mehra. Hand gesture recognition for human computer interaction. In *2011 International Conference on Image Information Processing*, pages 1–7, Nov 2011.

[9] Lam T. Phi, Hung Dinh Nguyen, T. T. Quyen Bui, and Thang Vu. A glove-based gesture recognition system for vietnamese sign language. *2015 15th International Conference on Control, Automation and Systems (ICCAS)*, pages 1555–1559, 2015.

[10] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *CVPR*, 2017.

[11] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016.