Gesture Recognition for Human-Robot Interaction

Internship Report

by

Rahul Kumar, IIT Patna
(Empl. No. **1832587**)

under the guidance of

Dr. Chayan Sarkar

TCS Reseach & Innovation,
Newtown, West Bengal 700135

January 24, 2020

Abstract

Human-robot interaction is becoming an important aspects of robotics as we envisage robots in our daily surroundings. As gesture is one of the common and natural modality of interaction, the robots should be able to identify the gesture and their contextual meaning.This project deals with the detection and recognition of hand gestures necessary for human-robot interaction.

In our day to day life we convey messages by using different gestures. Gestures may be nodding of head, waving of hand, facial expressions, eye movements and any body part movement. In this report, I have collected all the possible gestures which is used in human-robot or natural human-human interaction.

Acknowledgements

Through this acknowledgement, I express my sincere gratitude to all those people who have been associated with this project and have helped me with it and made it a worthwhile experience.

A special gratitude I give to my advisor, Dr. Chayan Sarkar, Scientist at TCS research & innovation whose contribution in stimulating suggestions and encouragement, helped me to coordinate my project especially in writing this report.

I would like to thank Mr. Sayan Paul for his assistance throughout the project and whole TCS community in achieving my goal. Also I would like to thank my parents who always keep me motivated.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

In today's world, computer science and related technology have provided a value added services to the user. In everyday life, physical gestures are a powerful means of communication. They can economically convey a rich set of facts and feelings. For example gestures may be used for nodding of head, waving of hand, facial expressions, eye movements and any body part movement. Use of the full potential of physical gesture is also something that most human computer dialogues lack.

This project aims to build a gesture identification network in order to achieve a higher accuracy. To enlighten the project it has been decided that the identification would consist of those gestures which are commonly used while interaction between human to human and as per identified gesture command will be given to robot for its movement.

1.2 Problem Statement

Our aim is to build a monocular vision based robust gesture recognition system for natural human robot interaction that can take a gesture as an input and automatically output the command that will help in further communication. I have implemented different algorithms on different kinds of gesture datasets.

1.3 The Gesture Datasets

I have collected some gesture datasets that are required for human-robot interaction. Some are publicly available and some can be available on request. In chapter 3, I give detailed description about the datasets. Different kinds of gestures used are–

- **Head and Face Gestures:** When people interact with one another, they use an assortment of cues from the head and face to convey information. These gestures may be intentional or unintentional, they may be the primary communication mode or back channels, and they can span the range from extremely subtle to highly exaggerate. Some examples of head and face gestures include: nodding or shaking the head, direction of

eye gaze, raising the eyebrows, opening the mouth to speak, winking, flaring the nostrils and looks of surprise, happiness, disgust, anger, sadness, etc [13]

- **Hand And Arm Gestures:** These two parts of body (Hand & Arm) have most attention among those people who study gestures in fact much reference only consider these two for gesture recognition. The majority of automatic recognition systems are for deictic gestures (pointing), emblematic gestures (isolated signs) and sign languages (with a limited vocabulary and syntax). Some are components of bimodal systems, integrated with speech recognition. Some produce precise hand and arm configuration while others only coarse motion.
- **Body Gestures:** This section includes tracking full body motion, recognizing body gestures, and recognizing human activity. Activity may be defined over a much longer period of time than what is normally considered a gesture; for example, two people meeting in an open area, stopping to talk and then continuing on their way may be considered a recognizable activity.

1.4 Applications

There are various applications of this gesture recognition system like human-robot interaction, virtual and augmented reality interaction, sign language recognition, and robotics. In many futuristic movies, characters perform human-robot interactions using hand gesture.

It is also useful as an assistive system. The development of a vision-based assistive system can help patient in their ambient lifestyle by analyzing their daily activities. A large part of natural interaction happens through hand gestures, especially if a robot is designed to help humans with everyday tasks (e.g. bring something by pointing at the object, go somewhere, etc). This requires a system that allows the robot to detect an interacting human and, obviously, make it understand hand gestures.

1.5 Our Contributions

Gestures come naturally to human beings, as they are an important part of the way we communicate, so the use of gestures for human-robot interactions is very easy to learn. The main contribution is first to collect all possible gestures that occurs in natural human-human interaction like waving hand, clapping hand, thumbs up or down, etc. There are different publicly available datasets that contains these kind of gestures which I have mentioned in chapter 3.

The success of CNNs in object detection and classification tasks [15],[11] has created a growing trend to apply them also in the other areas of computer vision. So, I tried different CNN Architecture with the given gestures. I described about all the algorithm and shown the result and accuracy in chapter 4.

1.6 Summary

This chapter gives a brief introduction about this report. I also provided the description of gesture datasets.

Chapter 2

Literature Survey

Considerable effort has been put towards developing intelligent and natural interfaces between users and robot systems. This is done by means of a variety of modes of information (visual, audio, pen, etc.) either used individually or in combination. The use of gestures as means to convey information is an important part of human communication. The basic goal of Human Robot Interaction is to improve the interaction between users and robots by making the robot more receptive to user needs. Interaction between humans comes from different sensory modes like gesture, speech, facial and body expressions. Being able to interact with the system naturally is becoming ever more important in many fields of Human Robot Interaction.

In this chapter, I first briefly discuss the human-robot communication. Next, I give a more detailed description of the gestures and gestures recognition. Later then, I described the challenges and use of deep learning in the gesture recognition. Finally I will summarize this chapter in this report.

2.1 Gestures

It is required to provide a way to explore the use of gestures in human-robot interaction(HRI) so that it can be interpreted by robot. The static and/or dynamic form of gestures of human arm, hand and even some other body parts require to be measurable by machine for the HRI interpretation. To facilitate and accomplish the advanced interaction between humans and robots, the designing of some special input devices has been found to be of great care in this area.

Theoretically gestures can be classified in two types Static and Dynamic Gestures. Static gestures can be defined as the gestures where the position and orientation of it in space does not change for an amount of time. If there are any changes within the given time, the gestures are called dynamic gestures. Dynamic gestures include gestures like waving of hand while static gestures include joining the thumb and the forefinger to form the “Ok” symbol.

2.2 Gesture Recognition

In computer science and language technology, gesture recognition is an important topic which interpret human gesture through computer vision algorithms. There are various bodily motion which can originate gesture but the common form of gesture origination comes from the face and hands. The entire procedure of tracking gesture to their representation and converting them to some purposeful command is known as gesture recognition.

A computer or other machine which recognises human gestures is one that can be effectively controlled by the movement of the hands and arms of the user [12]. Gestures come naturally to human beings, as they are an important part of the way we communicate, so the use of gestures for human-robot interactions is very easy to learn. Ideally, users can command machines to complete complex tasks using a single posture or relatively simple, continuous, dynamic hand gestures.

2.3 Human-Robot Interaction (HRI)

Human-Robot Interaction is a multidisciplinary arena which draws on the fields of computer science, psychology, cognitive science, and organisational and social sciences in order to understand how people use and experience interactive technology. This multidisciplinary field makes use of qualitative and quantitative research methods either to gather or to analyse information to be used. We are convinced that this is a trivial matter as research methods for HRI.

The motivation behind this research is to make an interaction between human and computer using various applications running on computer by aiming basic shapes made by hand. Our hand movements have an important role while interacting with other people, as they convey very rich information in many ways. According to this thought hand gestures would be an ideal option for expressing the feelings, or controlling the dynamic applications of computers through easier hand gesture.

2.4 Related Work

Gesture recognition was first proposed by Myron W. Krueger as a new form of interaction between human and computer in the middle of seventies. It has become a very important research area with the rapid development of computer hardware and vision systems in recent years. Freeman and Roth [9] introduced a method to recognize hand gestures, based on a pattern recognition technique developed by McConnell employing histograms of local orientation. Naidoo and Glaser [21] developed a system that is recognized static hand gesture against complex backgrounds based on South African Sign Language (SASL), a (Support Vector Recognition) system used to classify hand postures as gestures. Chang and Chen [2] presented a new approach for recognizing static gestures based on Zernike moments (ZMs) and pseudo-Zernike moments (PZMs). Triesch and Malsburg [29] employed the Elastic-Graph Matching technique to classify hand postures against complex backgrounds, hand postures

were represented by labelled graphs with an underlying two dimensional topology, attached to the nodes were jets.

The success of CNNs in object detection and classification tasks [15], [11] has created a growing trend to apply them also in the other areas of computer vision. For video analysis tasks, CNNs have been initially extended to be applied for video action and activity recognition and they have achieved state-of-the-art performances [27], [5]. The real-time systems for hand gesture recognition requires to apply detection and classification simultaneously on continuous stream of video. There are several works addressing detection and classification separately. In [23], authors apply histogram of oriented gradient (HOG) algorithm together with an SVM classifier. The authors in [19] use a special radar system to detect and segment gestures.

2.5 Challenges

Human-computer interaction (HCI) is the study of communication between computer and humans. It is a challenging field because the system needs to be able to perceive, understand and react to human activity in real time. Challenges include:

- There is a limitation on the size of the gesture recognition system, it must be able to fit on the computer.
- As the human are mobile, static backgrounds cannot be used for segmentation and a fixed camera location cannot be assumed.
- The computer need to adapt to drastic changes in environmental conditions, such as lighting or background color specifically.
- The system must be able to work in real-time. Ideally, there must not be a perceivable lag between the user performing a gesture and the computer response.

2.6 The Use of Deep Learning

Like many research areas in pattern recognition, including the hand pose estimation field, deep learning approaches have shown particularity good performance for hand gesture recognition. Their ability to learn relevant spatial and/or temporal features in addition to play the role of classifier, has been studied last years.

Convolutional neural networks [69] designed to take images as input has been used for static hand gesture recognition using RGB data [17] [20] and/or depth maps [16] . Neverova et al.[22] designed a multi-modal deep learning framework which takes as inputs: RGB, depth, audio stream and body skeleton data. Their network captured several spatial information, such as motions of the upper body or the hand, at three distinct spatial scales in order to perform dynamic sign language recognition. Their framework classified each frame and the final label of a sequence was computed using a majority vote.

Garcia et al. [10] used a two-stacked Long-Term Short Memory (LSTM) network as a baseline for their hand action dataset. LSTM has shown better performance over all previous traditional methods. Du et al. [6] proposed to divide the human body skeleton in five meaningful parts and fed each one into a distinct RNN network. They used a bidirectionnal variant [26] of the LSTM in order to use past frames but also future one to model each time step of a sequence. The recurrent layers are then fused step by step to be inputs of higher layers.

2.7 Summary

In this chapter, we review the literature necessary for our report. We looked at different literature on using deep learning and challenges of gesture recognition.

Chapter 3

Data Collection

This chapter provides a brief overview of the publicly available gesture recognition datasets, namely hand pose and body gesture datasets.

3.1 Publicly Available Datasets

There are datasets that have been made available for download allowing authors to compare results with a common benchmark. Section 3.1.1 presents the static gesture sets available and Section 3.1.2 presents the dynamic hand pose datasets available.

3.1.1 Static Gesture Sets

The majority of publicly available hand gesture sets are recorded using only an ordinary camera. Thus, there is no depth information available. The datasets that have both colour and depth information are summarised in Table 3.2

Source	Datasets
Ren at al. [42]	NTU Microsoft Kinect Hand Gesture dataset
Tompson et al. [88]	NYU Hand Pose dataset
Qian et al. [70]	Microsoft Research Asia Hand Pose dataset

Table 3.1: *Publicly available static gesture datasets*

The NTU dataset [25] contains both colour and corresponding depth from the Kinect sensor. It contains ten gestures from ten subjects with ten repetitions per gesture. Therefore there are one thousand gesture instances. However, no skeleton tracking information is provided and the dataset cannot be used for evaluating the hand gesture algorithms

The NYU dataset [28] consists of colour and depth images from three different views, one frontal and two side. There are only two users recorded, and rather than defining specific gestures the dataset contains a wide range of hand poses. As there is no ground truth, this dataset is hard to use for effective evaluation.

The MSRA dataset [24] is recorded using an Intel Creative depth camera. This dataset consists of six subjects performing various rapid gestures, with ground truth data for 2400 frames. Rather than a gesture label, the ground truth is the position of the five fingertips and the wrist. However, as there are no gesture labels this dataset cannot be used to evaluate the designed gesture recognition system.

3.1.2 Dynamic Gesture Sets

There are a number of publicly available datasets for gesture recognition, captured using a variety of sensors. These are summarised in Table 3.1.

Source	Datasets
Fothergill et al.	Microsoft Research Cambridge-12 Kinect Gesture
Lin et al.	Keck Gesture dataset
Chen Chen et al	UTD Multimodal Human Action Dataset
Escalera et al.	ChaLearn dataset 2014
Celebi et al.	VisApp2013 Gesture dataset
Bernstein et al.	Cornell Military Gesture Dataset
Christian et al.	FreiHAND

Table 3.2: *Publicly available dynamic gesture datasets*

The MSRC-Kinect dataset [8] consists of 12 full-body gestures that relate to gaming, music, and dance. The main purpose of the work was in investigating methods for instructing users for gesture performance. The dataset consists of 30 subjects recorded using a Microsoft Kinect. There are total of 6244 gesture instances, approximately 500 per class.

The Keck dataset [18] consists of 14 military signals performed using only the upper body. The dataset was collected using a standard colour VGA camera with a resolution of 640 x 480. There are 21 instances of each gesture: nine training and twelve testing. Each gesture was performed seven times by three different participants. This dataset has no depth information available; therefore, it could not be used for testing the proposed system.

The UTD-MHAD dataset [3] was collected using a Microsoft Kinect sensor and a wearable inertial sensor in an indoor environment. The dataset contains 27 actions performed by 8 subjects (4 females and 4 males). Each subject repeated each action 4 times. After removing three corrupted sequences, the dataset includes 861 data sequences. Four data modalities of RGB videos, depth videos, skeleton joint positions, and the inertial sensor signals were recorded in three channels or threads.

The ChaLearn2014 dataset [7] focused on recognising gestures from several instances performed by different users. The gesture lexicon consists of 20 Italian cultural signs performed using the upper body. The data is collected using a Kinect sensor and includes the depth map,

RGB image and skeleton information for 7754 gesture instances. There are approximately 388 samples per gesture.

The VisApp2013 dataset [1] consists of eight dynamic gestures performed using the upper body. There are four unique gestures, performed using the left and right sides of the body. The skeleton from the Kinect SDK is recorded. There are only 28 samples of each gesture, eight samples for training and twenty for testing. This dataset is used to evaluate the performance of the body gesture recognition system.

The CMU dataset consists of fifteen dynamic gestures, namely, "action", "advance", "attention", "charge", "cover", "crouch", "rally", "shift fire", "point of entry", "confused", "hurry", "sneak", "out of action" and "come". The gestures are performed using the right arm and the skeleton from the NiTE SDK is recorded.Three users perform each of the 15 gestures ten times giving a total of 450 gesture instances. The dataset is recorded using a Kinect sensor and the joint positions of the body are saved. This dataset is used to evaluate the performance of the body gesture recognition system.

FreiHAND dataset [30], a dataset for hand pose and shape estimation from single color image, which can serve both as training and benchmarking dataset for deep learning algorithms. It contains 4*32560 = 130240 training and 3960 evaluation samples. Each training sample provides: - RGB image (224x224 pixels) - Hand segmentation mask (224x224 pixels) - Intrinsic camera matrix K - Hand scale (metric length of a reference bone) - 3D keypoint annotation for 21 Hand Keypoints - 3D shape annotation The training set contains 32560 unique samples post processed in 4 different ways to remove the green screen background. Each evaluation sample provides an RGB image, Hand scale and intrinsic camera matrix

3.2 Prepared Own Datasets

I have also collected datasets with 4 classes only which are Thumbs Up, Thumbs down, Pointing and Palm gestures. The datasets contain 2000 samples for each gestures for 5 different subjects out of which 3 male and 2 female. I have collected it in different background, different lighting condition and with change in distance.

Chapter 4

System Design

In this chapter, I will discuss the different approaches of doing gesture recognition. We can see the pictures in two different ways i.e as a traditional way and deep learning way.

4.1 Traditional Way

There are numerous way to perform gesture recognition over the years. In this way, we need to take care about the pre-processing steps, since image taken from camera is highly important. Apart from that, numerous factors such as lights, environment, background of the image, hand and body position and orientation of the subject, parameters and focus the of camera impact the result dramatically. There are few steps which we were taken while doing gesture recognition as a traditional way.[14]

- **Color Segmentation:** The main purpose of Color segmentation is to find particular objects for example lines, curves, etc in images. In this process every pixel is assigned in an image in such a way that pixels with the same label share certain visual characteristics. The goal of color segmentation is basically to simplify and increase the ability of separation between skin and non-skin, and also decrease the ability of separation among skin tone.
- **Skin Detection:** There are several techniques used for color space transformation for skin detection. This process involves classification of each pixel of the image to identify as part of human skin or not by applying Gray-world Algorithm for illumination compensation and the pixels are categorized based on an explicit relationship between the color components YCbCr.
- **Image Segmentation:** To reduce the computational time needed for the processing of the image, it is important to reduce the size of the image and only the outline of the sign gesture has to be process.
- **Image Filtering:** In Image Filtering technique, the value of the pixel of any given image is determined by applying algorithm to the value of the pixels in the neighborhood.

4.2 Using Deep Learning

Recently, many applications of the Computer Vision (CV) field shown a change of paradigm. From human activity recognition to speech recognition, image classification and labeling, CV areas see the emergence and successful arrival of the machine learning technology called deep learning. Since 2010, researchers migrate from traditional handcrafted features to learned-based features also called data-driven algorithm. There are many learned-based feature methods for vision recognition tasks such as dictionary-based approaches or genetic programming. Nevertheless, we focus on deep learning as, in recent years, it changes the game in computer vision.[4] We also surveyed some deep learning algorithms. In next section, I will explained which algorithms outperform than others. I will also demonstrate which is the best for other gestures as well.

4.3 Implementation

After collecting all the gestures datasets, it's time to implement those gestures on some real time gesture recognition algorithm.

4.3.1 Model Architecture

It is a very light weight model where input image of size 200 x 200 is passed to conv2d layers and later followed by 3 dense layers.

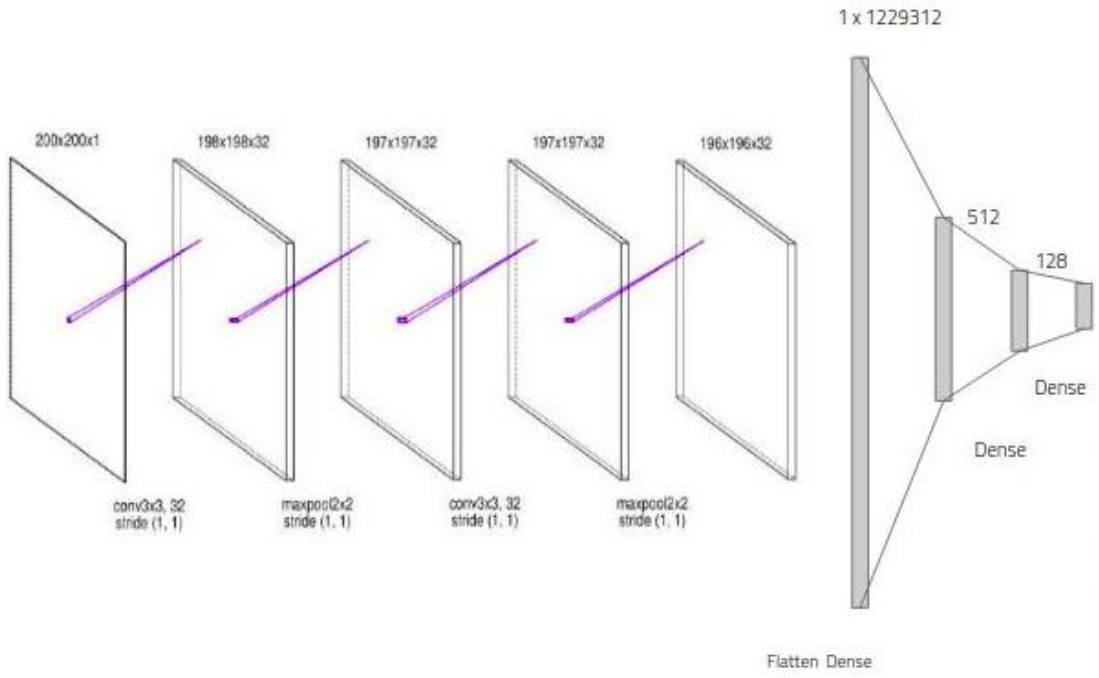

Figure 4.1: *CNN based Hand Gesture Recognition Model*

4.3.2 Result

After running 100 epoch following result, I obtained. The final training accuracy is 0.9487 and validation accuracy is 0.6205. And the training loss is 1.3826 and validation loss is 2.6172. Since the accuracy is not very much. We can improve this accuracy by making datasets more generic considering three factors i.e background, lighting and distance while taking images.

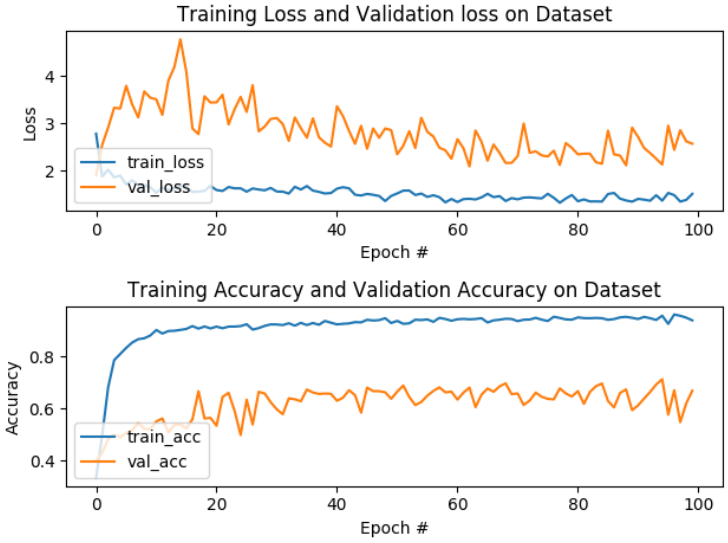

Figure 4.2: *Training Loss/Accuracy and Validation Loss/Accuracy*

4.4 Summary

This chapter is all about our approach to multi-task learning. In next chapter 5, we have concluded the report and discussed the future work and also explained the reason why we are getting low values.

Chapter 5

Conclusions and Future Work

Human–robot interaction is a growing field of research and application. For real life applications it is a very challenging because of its robustness, accuracy and efficiency. The field includes many challenging problems and has the potential to produce solutions with positive social impact. Research and development of gesture controls for over 30 years made it possible that we can now interact with devices in a more natural and intuitive way than in the past. Gesture controls already have many application possibilities and will play an even bigger role in our technological future.

We can improve the model by choosing an appropriate classifier. Since the results are not good now. In future, we have to work on hand tracking system i.e localization of hand so that there is no restrictions to show hand inside a particular block in front of camera. We need to speed up the real time predictions. Although some algorithm is giving much appropriate result but still we need to work on it. The system should detect multiple gesture at a time using less hardware as possible. The system should respond accurately on partial occlusions.

Bibliography

[1] Sait Celebi, Ali Selman Aydin, Talha Tarik Temiz, and Tarik Arici. Gesture recognition using skeleton data with weighted dynamic time warping. In *VISAPP*, 2013.

[2] Chin-Chen Chang, Jiann-Jone Chen, Wen-Kai Tai, and Chin-Chuan Han. New approach for static gesture recognition. *J. Inf. Sci. Eng.*, 22:1047–1057, 09 2006.

[3] C. Chen, R. Jafari, and N. Kehtarnavaz. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *2015 IEEE International Conference on Image Processing (ICIP)*, pages 168–172, Sep. 2015.

[4] Quentin De Smedt. *Dynamic hand gesture recognition - From traditional handcrafted to recent deep learning approaches* . Theses, Université de Lille 1, Sciences et Technologies; CRIStAL UMR 9189, December 2017.

[5] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. *CoRR*, abs/1411.4389, 2014.

[6] Yong Du, Wei Wang, and Liang Wang. Hierarchical recurrent neural network for skeleton based action recognition. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1110–1118, 2015.

[7] Sergio Escalera, Xavier Baró, Hugo Jair Escalante, and Isabelle Guyon. Chalearn looking at people: Events and resources. 01 2017.

[8] Simon Fothergill, Helena Mentis, Pushmeet Kohli, and Sebastian Nowozin. Instructing people for training gestural interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, page 1737–1746, New York, NY, USA, 2012. Association for Computing Machinery.

[9] William T. Freeman and Michal Roth. Orientation histograms for hand gesture recognition. Technical report, Mitsubishi Electric Research Labs., 201, 213.

[10] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 409–419, 2017.

[11] Ross B. Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CoRR*, abs/1311.2524, 2013.

[12] C. Harshith, Karthik R. Shastry, Manoj Ravindran, M. V. V. N. S. Srikanth, and Naveen Lakshmikhanth. Survey on various gesture recognition techniques for interfacing machines based on ambient intelligence. *CoRR*, abs/1012.0084, 2010.

[13] J. Heinzmann and A. Zelinsky. Robust real-time face tracking and gesture recognition. In *In Proc. of the Int. Joint Conf. on Artificial Intelligence, IJCAI'97*, pages 1525–1530, 1997.

[14] Belas Ahmed Khan and Amir Hassan Pathan. Hand gesture recognition based on digital image processing using matlab. 2015.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, page 1097–1105, Red Hook, NY, USA, 2012. Curran Associates Inc.

[16] Shaozi Li, Bin Yu, Wei Wu, Songzhi Su, and Rongrong Ji. Feature learning based on sae-pca network for human gesture recognition in rgbd images. *Neurocomputing*, 151:565–573, 2015.

[17] Hsien-I Lin, Ming-Hsiang Hsu, and Wei-Kai Chen. Human hand gesture recognition using a convolution neural network. *2014 IEEE International Conference on Automation Science and Engineering (CASE)*, pages 1038–1043, 2014.

[18] Zhe L. Lin, Zhuolin Jiang, and Larry S. Davis. Recognizing actions by shape-motion prototype trees. *2009 IEEE 12th International Conference on Computer Vision*, pages 444–451, 2009.

[19] Pavlo Molchanov, Shalini Gupta, Kihwan Kim, and Kari Pulli. Multi-sensor system for driver's hand-gesture recognition. *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 1:1–8, 2015.

[20] Jawad Nagi, Frederick Ducatelle, Gianni A. Di Caro, Dan C. Ciresan, Ueli Meier, Alessandro Giusti, Farrukh Nagi, Jürgen Schmidhuber, and Luca Maria Gambardella. Max-pooling convolutional neural networks for vision-based hand gesture recognition. *2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, pages 342–347, 2011.

[21] S. Naidoo, C. W. Omlin, and M. Glaser. Vision-based static hand gesture recognition using support vector machines, 1998.

[22] N. Neverova, C. Wolf, G. Taylor, and F. Nebout. Moddrop: Adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(8):1692–1706, Aug 2016.

[23] 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.

[24] Chen Qian, Xiao Sun, Yichen Wei, Xiaoou Tang, and Jian Sun. Realtime and robust hand tracking from depth. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, page 1106–1113, USA, 2014. IEEE Computer Society.

[25] Zhou Ren, Jingjing Meng, Junsong Yuan, and Zhengyou Zhang. Robust hand gesture recognition with kinect sensor. pages 759–760, 11 2011.

[26] Mike Schuster and Kuldip K. Paliwal. Bidirectional recurrent neural networks. *IEEE Trans. Signal Processing*, 45:2673–2681, 1997.

[27] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, page 568–576, Cambridge, MA, USA, 2014. MIT Press.

[28] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. Real-time continuous pose recovery of human hands using convolutional networks. volume 33, 08 2014.

[29] Jochen Triesch and Christoph von der Malsburg. Robotic gesture recognition. In Ipke Wachsmuth and Martin Fröhlich, editors, *Gesture and Sign Language in Human-Computer Interaction*, pages 233–244, Berlin, Heidelberg, 1998. Springer Berlin Heidelberg.

[30] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images, 09 2019.