Taylor & Francis
Taylor & Francis Group

Applied Artificial Intelligence
An International Journal

ISSN: 0883-9514 (Print) 1087-6545 (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

# Gesture and Speech Recognizing Helper Bot

Kailash Gogineni, Akhil Chitreddy, Anirudh Vattikuti & Natarajan Palaniappan

**To cite this article:** Kailash Gogineni, Akhil Chitreddy, Anirudh Vattikuti & Natarajan Palaniappan (2020) Gesture and Speech Recognizing Helper Bot, Applied Artificial Intelligence, 34:7, 585-595, DOI: 10.1080/08839514.2020.1740473

**To link to this article:** https://doi.org/10.1080/08839514.2020.1740473

APPLIED ARTIFICIAL INTELLIGENCE
2020, VOL. 34, NO. 7, 585–595
https://doi.org/10.1080/08839514.2020.1740473

Taylor & Francis
Taylor & Francis Group



# Gesture and Speech Recognizing Helper Bot

Kailash Gogineni[a], Akhil Chitreddy[b], Anirudh Vattikuti[c], and Natarajan Palaniappan[d]

[a]The George Washington University, Washington, DC, USA; [b]Virtusa, Chennai, Tamilnadu; [c]University of New South Wales, Sydney, Australia; [d]Vellore Institute of Technology, Vellore, Tamilnadu

**ABSTRACT**

In industries, difficult work is being decreased everywhere scale to expand effectiveness and exactness, and gain benefit by introducing robots that can do repetitive works at lesser expense of preparing. A onetime establishment of such a gadget may cost an enormous sum at first, yet in the more drawn out run, will end up being more beneficial than difficult work. Out of the part, a basic robotic arm is a standout amongst the most generally introduced machines. Robotic arm is one of the significant undertakings in the present computerization industry. Automated arm is a piece of the mechatronic industry which is a quickly versatile and developing industry today. Distinctive changes and extra features are being associated with the first kind of straightforward robotic arm to upgrade its ease of use under various conditions. In this paper, we are building up a robotic arm which will have a free rotation around multiple axes and we are including the technology of Image Processing with it to make it a visual signal based working robotic arm. The model is a pick and place robotic arm you can take the desired article starting with one place and carry it to another place with its gripper claw. The operations will be constrained by a visual processing framework that reads the gestures and will give directions to the Arm for performing various types of movements and tasks. We are making use of low torque servos to lift light weight.

## Introduction

In this paper, we will learn diverse parts of Mechatronics, Image Processing, sound processing Mechanics of Machines and their modern execution. We will have a hands-on involvement on working with Python, coding, and execution. We will find out about the usage of Artificial Intelligence in the field of Image Processing. The proposed framework comprises two modules. The first module is Robotic arm control utilizing static hand signals and the second module deals with Maneuverability control utilizing voice commands.

Raspberry Pi acts as a master in perceiving the signals and Arduino will go about as slave by moving the mechanical technology according to the direction

**CONTACT** Natarajan Palaniappan pnatarajan@vit.ac.in Department of Computer Science and Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India

issued by Raspberry Pi. The gestures are prepared utilizing Machine learning procedure utilizing Open CV and SKlearn (essential python libraries utilized for machine learning). At first, the motion is distinguished utilizing background subtraction and after that skin segmentation is made and the hand portion is alone segmented and further the features are extracted using Histogram of Oriented gradients and further it is classified using Random Forest. The final model will be deployed in the Raspberry Pi based on the detected gesture it will send a command to Arduino to control servo/DC motors of the robotic to perform custom gestures using pulse width modulation.

Further utilizing Google API clients voice inputs are captured and prepared in such an approach to get the string of the specific command given by the client. The strings are further given to Arduino such that the mobility is controlled.

Utilizing Histogram of Oriented Gradients and Random Forests makes the venture increasingly exact in perceiving motions at any rate and parity of exchange off among speed and exactness is kept up.

## Literature survey

Robots have turned into a substitute for people in unsafe and risky conditions. They have turned out to be all the more socially intuitive and this clears path for improvement of algorithms to influence them to perform the tasks productively. Making machines progressively interactive with people and to improve the assignment execution of the robot with less human intervention are the principle goals of HRI. The robot must be equipped for translating, breaking down and understanding a few correspondence components engaged with human to human interaction. The past works in the field of HRI are talked about as for the region of utilization (Chandrasekaran and Conrad 2015).

Sign language is a vital instance of open gestures. Since gesture-based communications are exceptionally basic, they are truly reasonable as testbeds for vision calculations (Wisenburn and Higginbotham 2009). In the meantime, they can likewise be a decent method to assist the crippled with interacting with PCs. Communication via gestures for the deaf is a precedent that has got critical consideration in the gesture literature (Hawley 2007).

The applications like Telepresence and telerobot are regularly arranged inside the area of room investigation and military-based research ventures. The gestures used to connect with and control robots are like completely submerged computer-generated experience collaborations, anyway the worlds are regularly genuine, giving the administrator video feed from cameras situated on the robot (Todman et al. 2008). Gestures can control a robot's hand and arm developments to go after and control genuine articles, its development through the world.

Desktop and Tablet PC Applications: In work area figuring applications, gestures can give an elective cooperation to the mouse and console (Ferrier

et al. 1995). Numerous gestures for work area registering assignments include controlling illustrations, or clarifying and altering archives utilizing pen-based signals (Thomas-Stonell et al. 1998).

A player's hand or body position has been followed by Freeman et al. (Bloor, Barrett, and Geldard 1990) to control the development and introduction of intelligent amusement articles, for example, vehicles. Konrad et al. (Sandler and Sonnenblick 1998) utilized gestures to control the development of symbols in a virtual world, and Play Station 2 has presented a camera named Eye Toy, which tracks hand developments for intelligent amusements (Wisenburn and Higginbotham 2008).

Gestures for virtual and augmented reality applications have encountered one of the best dimensions of take-up in registering. Augmented reality connections utilize signals to empower reasonable controls of virtual items utilizing one's hands, for 3D presentation communications (O'Keefe, Kozak, and Schuller 2007) or 2D presentation that mimic 3D associations (Murphy 2004).

Needs for senior individuals at home are being filled by social-interactive robots. One genuine model is HOBBIT which takes guidelines from the user as motions and facilitates the user with a contact screen interface to make a call, stimulation, web, etc. Thus, MOBILEROBOTS PeopleBot helps the impaired and old individuals in family unit exercises. The association is through Phantom Omni, a haptic interface mounted over a robot (Vincze et al. 2014). Kinect sensor goes about as interface to recognize arm signal for shopping in Mobile Shopping Cart (MSC) which executes the directions inside a large portion of a second time (Gai, Jung, and Byung-Ju 2013).

Recognition of the feelings being a piece of human-robot association, wherein physiological signals such as tension are registered and sustained through the bio-sensors mounted on human administrator. Moving nearer or casual banter with the user are the extra tasks performed by the robot alongside wall following, obstacle identification dependent on the data sources sustained to the robot. The Robot-based Basketball task (RBB) is an exploratory proving ground where consistent adjustment of its conduct happens as for the human physiological states (Bekele and Sarkar 2014).

Kindergarten children were made to communicate with robots so as to advance the geometrical thinking while at the same time playing. Thus, it improved their geometrical information and appreciated learning. Over the time, the robot is equipped for giving out execution insights of the kids. KindSAR (Chandrasekaran and Conrad 2015), Kindergarten social assistive robots, give entertainment to youngsters by narrating.

Recognition of the Hand Gesture was executed utilizing the parameters based on the shape (Panwar 2012). The downside is that extreme parameters are assumed for this approach.

## Technology stack

The paper deals with identifying the gestures and responding to the speech with the use of Google API and projecting the actions a bot can perform to lift an object. Technology stack includes Machine learning, Image processing, and Robotics.

- Machine Learning: It is the advanced technology in computer science, which helps the system to behave as a human. The base of the Machine learning is pure Mathematics, which helps to perform wonderful tasks. Basically, four algorithms such as SVM, logistic regression, Random forest, and KNN are used and their accuracy, precision, recall, support are tested and the highest accuracy supported algorithm is used to embed into the Raspberry Pi and further process of loading the data set is done. Now, the machine is trained to recognize some actions. Recognized data is loaded into the program and test against the known and unknown fingers.
- Image Processing: Image processing helps to do some processing techniques which help to improve the accuracy of recognition of the fingers through the mask. While taking the images of the gestures for dataset collection hog algorithm is used to extract features from the images.
- Robotics: To perform the actions of picking a block or an object a bot is needed which can walk and turn in different directions with respect to the gestures and speech, mostly in the form of a best method called Artificial Intelligence. The bot runs from calculating angles one by one, more likely they are powered by AI blocks that were trained to stand/walk/balance

## Algorithms

### *Implementation of KNN*

(1) At the outset, we will generate the data set and send the data.
(2) Then, the value of k is initialized (the nearest neighbor so that we can make boundaries of each class).
(3) Now, for obtaining the class of predicted output, the iteration is to be done from 1 to total number of training data arguments.
    - Analyze the test data and each row of the training data and certainly compute the distance between them with the most popular method called Euclidean distance.
    - The next step is, we will be using sorting method to the premeditated spaces in ascending order depending on their distance values.

- Now, obtain the top K rows from which the sorting is done.
- Now, just obtain the most recurrent class of the rows.
- Finally, the predicted class is returned and the subsequent output is given.

### *Random forest algorithm procedure*

The procedure consists of two phases which is "creation" and "perform the prediction" from the obtained random forest classifier.

### *Random forest pseudocode*

(1) Arbitrarily pick "k" structures from a whole of "m" structures, only when $k \ll m$.
(2) Amid the different features of "k," compute the node "d" using best split approach.
(3) Next, we need to separate the node into offspring nodes using the best split approach.
(4) Replicate the above 1 to 3 steps until "l" no.of nodes have been occurred and then construct forest by replicating all 1 to 4 steps for "n" no.of times to generate "n" no.of trees.

### *Prediction technique*

Prediction is performed using the trained random forest algorithm, the steps are

(1) Using the test features and the arbitrarily created decision tree perform the prediction and store it in the form of a target. Then, the votes for each predicted target are computed.
(2) Finally, the high voted target is certainly fixed as final prediction.

Voting: Tree one and two would vote that she subsisted, but tree three votes that she expires. If we take a vote, it is 2 to 1 in favor of her existence, so we would classify this commuter as a stayer.

### *SVM*

Given a data set containing of features set and tags set, an SVM classifier dimensions a prototype to predict classes for innovative occurrences. It allocates different data arguments to individual classes. If at all, there are only two classifiers means, then the classifier is called Binary SVM classifier.

In the SVM algorithm, the plotting of each data element as unique identity in the "n"- dimensional space, where "n" is defined as the no. of features with

every feature having a precise coordinate. Then, in the next step the classification is done by recognizing the hyperplanes that categorize the two classes. SVMs are briefly defined as the cutting edge which separates the two classes (hyperplane/line).

### *Logistic regression*

For logistic regression, we can take any kind of data with a wide range. It will run the algorithm on the training set and the model is used to decide the class of the test data.

#### *Pseudocode*

The equation needed to predict the class of the test data is prepared and the data is supplied to the equation to get regression value.

The regression value is used by the activation function to predict the class of the data. In the paper, sigma function is used as the activation function.

## Methodology

The procedure starts by taking images of hand gestures and creating a dataset. The features from the images taken are extracted by using the HOG (histogram of oriented gradient) algorithm. The dataset is used to train various models such as KNN, SVM, random forest, logistic regression. The best among the algorithms is chosen to create an efficient prediction model using the test data. The model is then fed into the Raspberry Pi module. When the program starts running the camera starts to detect hand gestures using the random forest model, as soon as the gesture has been identified the output is sent in the form binary numbers to the Arduino module. The Arduino module makes use of relays to increase the power voltage from 5 V to 12 V and it will be sent to the motors along with the signals from the Arduino board. As per the signals received the robotic arm performs actions in order to pick and place objects. The speech recognizing module includes a chassis which is controlled using the same Arduino module which has Bluetooth module connected to it. The Bluetooth module can be operated by using the google API which is already present in an app, then the signals are sent to Arduino for movement controlling. The movements include forward, reverse, right, left, stop. The power source for the robot is given through plug points and the movement is enabled till 2 m.

## Design of the system

The helper bot is designed in such a way that the gesture and speech modules are independently implemented. The gesture recognition system uses

machine learning techniques to identify gestures whereas the speech recognition system uses google API to pass the voice commands as signals to the Arduino board.

### Gesture recognition

There are two phases in the gesture recognition module. They are the training and the testing phase. From the above diagram, we can see that during the training phase, dataset is collected and then the edge detection algorithms start to act to identify the gestures from our hand. After completing these steps, the training of images takes place to classify the gestures. Here we chose random forest algorithm to classify the gestures. Once the model is created we start to take video or an image input to predict the class of the image which will be sent to the Raspberry Pi for the robotic arm movement as given in Figure 1.

### Speech recognition

The speech recognition module works by taking voice input through an android device. The voice commands are converted into signals by making use of a google API. When the Arduino microcontroller receives the signals, it controls the chassis of the robot by making it move front, back, left, right as given in Figure 2.

## Results

The dataset images of the hand gestures are collected and fed into all the algorithms and then training and tested parts are carried out for calculating the various parameters required to calculate the f1 score which will be helpful for finding accuracy. The results from the various algorithms are as follows:

| Algorithms | Precision (avg) | Recall (avg) | F1-score (avg) | Support (avg) |
|---|---|---|---|---|
| Random Forest | 1 | 1 | 1 | 1296 |
| SVM | 0.99 | 0.99 | 0.99 | 1296 |
| KNN | 0.88 | 0.88 | 0.88 | 1296 |
| Logistic Regression | 0.80 | 0.80 | 0.80 | 1296 |

So, from the derived results we can say that the algorithm of Random Forest and Support vector machine (SVM) gives us the highest values for the parameters precision, recall, and f1-score. Among the top two, it is obvious

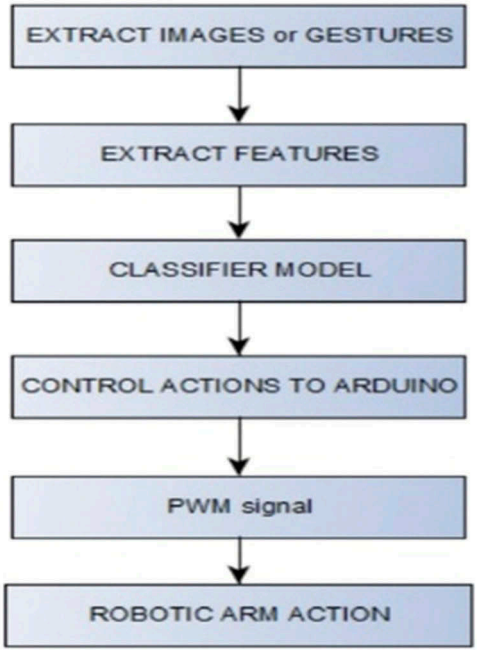


**Figure 1.** Gesture recognition module flowchart.

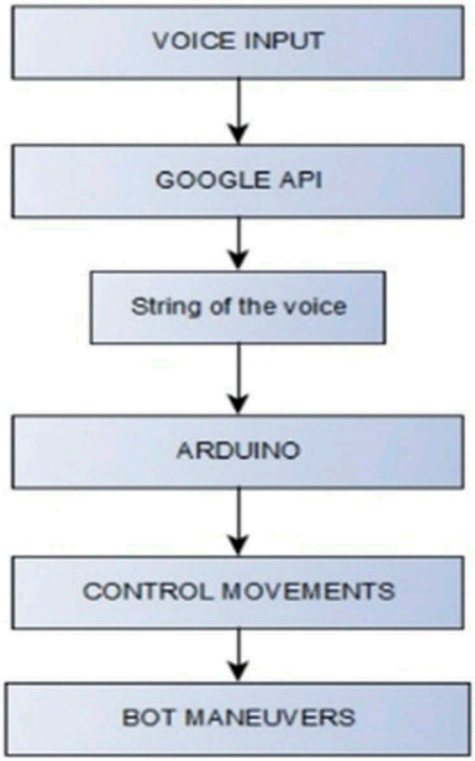


**Figure 2.** Speech recognition module flowchart.

that random forest which is an ensemble learning classifier gives the best results. This will now be embedded into the Raspberry Pi module for recognizing gestures.

After training the model with a dataset of 5000 images, we then test it on some random test data to classify the gesture. The gesture number appears on the top left corner as given in Figure 3.

The program successfully identifies the gestures when embedded in Raspberry Pi. The Raspberry Pi sends the signals to the robot to perform the corresponding action as given in Figure 4.

## Scope

- Robotic devices combined with machine learning is the trending area for developing applications such as automatic plant cutting in a given period of time.
  - These can also be used as assistant bots.
  - When integrated with large-scale equipment, which can send signals in a high range, it will be a great help for the military in detecting the mines.

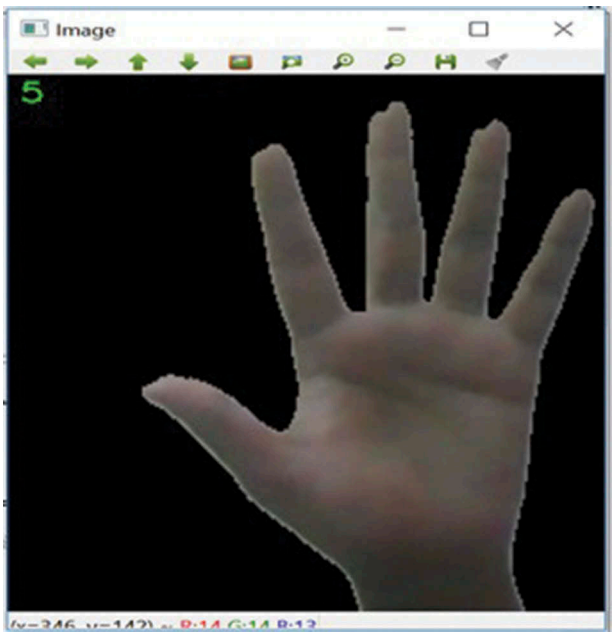


**Figure 3.** Random forest algorithm classifying the gesture on the top left corner.

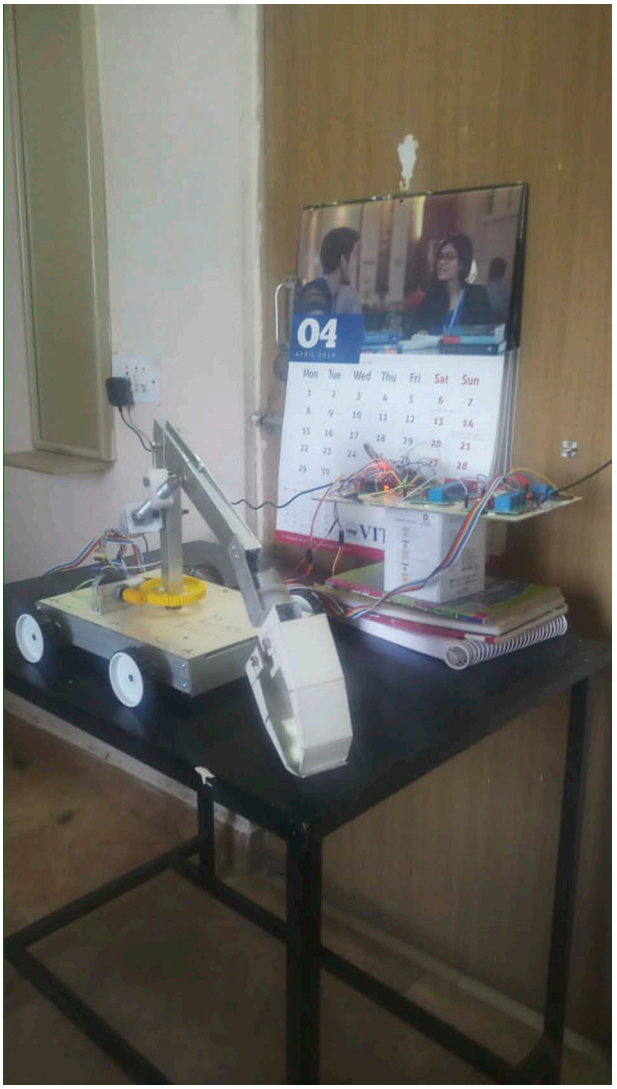


**Figure 4.** Robotic arm picking up light objects such as paper balls, etc.

## Conclusion

We can imply that the random forest performs better compared to the rest of the algorithms. Therefore, that will be fed into the Raspberry Pi module for recognizing gestures. When the data set is collected without noise, the gestures can be classified with higher accuracy. The robot picks light objects perfectly with a grip by following the gestures and voice commands.

## References


Bekele E., and N. Sarkar. 2014. Psychophysiological Feedback for Adaptive Human–Robot Interaction (HRI). In *Advances in Physiological Computing. Human–Computer Interaction Series*, ed. by S. Fairclough and K. Gilleade. London: Springer.

Bloor, R. N., K. Barrett and C. Geldard. 1990. "The clinical application of microcomputers in the treatment of patients with severe speech dysfunction," *IEE Colloquium on High-Tech Help for the Handicapped*, p. 9. London, UK.

Chandrasekaran, B., and J. M. Conrad. 2015. Human-Robot Collaboration: A Survey. Proceedings of the IEEE SoutheastCon 1–8. Fort Lauderdale, Florida, USA.

Ferrier, L. J., H. C. Shane, H. F. Ballard, T. Carpenter, and A. Benoit. 1995. Dysarthric speakers' intelligibility and speech characteristics in relation to computer speech recognition. *Augmentative and Alternative Communication* 11 (3):165–75. doi:10.1080/07434619512331277289.

Gai, S., E.-J. Jung, and Y. Byung-Ju, "Mobile shopping cart application using kinect", *10th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI)*, pp. 289–91, Jeju, South Korea, 2013 "Face recognitioin with local binary patterns" by Timo Ahonen, Abdenour Hadid, Matti Pietikainen.

Hawley, M. S. 2007. A speech-controlled environmental control system for people with severe dysarthria. *Medical Engineering & Physics* 29 (5):586–93. doi:10.1016/j.medengphy.2006.06.009.

Murphy, J. 2004. I prefer contact this close: Perceptions of AAC by people with motor neurone disease and their communication partners. *Augmentative and Alternative Communication* 20 (4):259–71. doi:10.1080/07434610400005663.

O'Keefe, B., N. Kozak, and R. Schuller. 2007. Research priorities in augmentative and alternative communication as identified by people who use AAC and their facilitators. *Augmentative and Alternative Communication* 23 (1):89–96. doi:10.1080/07434610601116517.

Panwar, M. 2012. "Hand gesture recognition based on shape parameters," *2012 International Conference on Computing, Communication and Applications*, Dindigul, Tamilnadu.

Sandler, U., and Sonnenblick, Y. 1998. A system for recognition and translation of the speech of handicapped individuals. In *MELECON'98. 9th Mediterranean Electrotechnical Conference. Proceedings,* vol. 1, 16–19. IEEE.

Thomas-Stonell, N., A. L. Kotler, H. A. Leeper, and P. C. Doyle. 1998. Computerized speech recognition: Influence of intelligibility and perceptual consistency on recognition accuracy. *Augmentative and Alternative Communication* 14 (1):51–56. doi:10.1080/07434619812331278196.

Todman, J., N. Alm, J. Higginbotham, and P. File. 2008. Whole utterance approaches in AAC. *Augmentative and Alternative Communication* 24 (3):235–54. doi:10.1080/08990220802388271.

Vincze, M., W. Zagler, L. Lammer, A. Weiss, A. Huber, D. Fischinger, T. Koertner, A. Schmid, and C. Gisinger. 2014. Towards a robot for supporting older people to stay longer independent at home. Proceedings of 41st International Symposium on Robotics 1–7. Munich, Germany.

Wisenburn, B., and D. J. Higginbotham. 2008. An AAC application using speaking partner speech recognition to automatically produce contextually relevant utterances: Objective results. *Augmentative and Alternative Communication* 24 (2):100–09. doi:10.1080/07434610701740448.

Wisenburn, B., and D. J. Higginbotham. 2009. Participant evaluations of rate and communication efficacy of an AAC application using natural language processing. *Augmentative and Alternative Communication* 25 (2):78–89. doi:10.1080/07434610902739876.