DOI: 10.15849/IJASCA.230320.10

Int. J. Advance Soft Compu. Appl, Vol. 15, No. 1, March 2023
Print ISSN: 2710-1274, Online ISSN: 2074-8523

Intelligent Gesture Recognition System for Deaf People by using CNN and IoT

Abdallah Abualkishik[1]*, Wael Alzyadat[2], Marwan Al Share[1], Sara Al-Khaifi[1], Mojtaba Nazari[1]

[1] Faculty of Computing and IT, Sohar University, Oman
e-mail: aabualkishik@su.edu.om, e-mail: mshare@su.edu.om,
e-mail: Sara.sk1994@hotmail.com, e-mail: nerfan15@gmail.com,
[2]Faculty of Science and IT, Al-Zaytoonah University of Jordan, Jordan
e-mail: wael.alzyadat@zuj.edu.jo

Abstract

Communication with the hearing impaired is a great challenge in the society today. Sign language is a significant communication method between deaf persons and their societies. However deaf and dumb people have an issue in this field as sign language may cause a lot of misunderstanding. Today, many new technologies are becoming more applicable and cheaper. The intelligent algorithms and techniques are improved and becomes more accurate. Gesture recognition is one of the key technologies that facilitate many people life. Translating signs language into text and speech is an attractive field for researchers from past few years. However, it did not receive adequate interest in Sultanate of Oman. The proposed recognition system utilizes the Convolution neural network (CNN) method. It aims to convert the dynamic deaf signs from live video to text and speech using raspberry pi device and normal camera. The dataset for this project was created by the researchers. It contains 62000 (64x64 pixel) images of the 30 letter of Alphabets and 1 Word. Each pattern has 2000 images that are divided into 1750 images for training and 250 images for testing. The proposed system achieved 99.8 % accuracy.

Keywords: Sign Language Recognition, Deep learning, image processing, American sign Language, Hand gesture detection.

1 Introduction

This means that technology is improving, and researchers are able to do more with multimedia, computer vision, and human computer interaction. Technology is expanding dramatically and acting as a critical actor in discovering solutions to all

Received 4 November 2022; Accepted 2 January 2023

world's needs and challenges. People communicate by talking to each other. This is a natural way for people to interact and it helps us to do things together. However deaf-mute people have a special circumstance make them not able to communicate naturally. Some countries have special schools and learning centers for people with disabilities. These centers help the disabled learn and become active members of their community. [1]. But unfortunately, this concern is absent in most of the countries. In [2], around 466 million people worldwide have disabling hearing loss, and 34 million of these are children. One of the main impacts of hearing loss is on the individual's ability to communicate with others. Exclusion from communication can have a significant impact on everyday life, causing feelings of loneliness, isolation, and frustration. Deaf-mute person throughout the world uses sign language for the communication. They are specially trained to use this sign language between each other's. Normal people mostly are not able to understand deaf-mute people sign language [3]. In [4], Nowadays, technology provides users with different services and capabilities. This sentence is saying that there are many opportunities to make the world smarter and more connected in different areas. A hand gesture recognition system is a computer program that can identify, and track hand gestures made by a person. This technology has received a lot of attention in recent years because it has many potential applications, such as helping people with disabilities to communicate or interact with machines more efficiently [5].

Sign language is not a language that is spoken universally. There are various countries that have their own standardized sign language. American Sign Language (ASL) is different than British Sign Language (BSL) even though both countries speak English. There are no published studies that are concerned with hand gesture recognition systems in the Sultanate of Oman. Since that there is a need to progress in this direction. Currently, there is a massive growth in technologies. This progress should improve and positively affect the lives and learning of people with disabilities, providing them with the chance to lead, create, and introduce initiatives. This project is about making a system that can recognize hand gestures and turn them into commands. This will be done by using a camera and special software that can track movement and interpret it as a command. The app aims to help deaf-mute people communicate with normal people. Hand gesture recognition is a cutting-edge technology that can adopt and manipulate conditions by two aspects: human and technology.

2 Related works

Hand gesture recognition is a field of research that has been active for a long time. This means that there are many different approaches to hand gesture recognition in the literature. Some people have tried to use cameras and computer programs to understand sign language. Other sign language recognition systems are mostly old and out-of-date, written before deep learning was created. Most reviews focus on a specific subarea,

such as the software and hardware used to recognize sign. In [6], Some researchers are working on a way to recognize people's hands in vision-based systems. They recorded the movement of the hand using video camera(s). The videos were turned into a set of features that look at individual frames. While in [7], the project works on a real-time hand gesture recognition and human-computer interaction system. In this study, 19852 images were collected from five people and classified into 16 different gestures. The researchers use a CNN from letNet-5 to recognize gestures. A CNN is a type of neural network that is designed to work with images. In this case, the CNN is being used to classify images based on whether they contain a hand. The reason the CNN is being fed a binary image (one that only contains two colors) is so that the classifier Images need to be pre-processed before they are fed into the CNN classifier [8]. This means that the images need to be edited or changed in some way so that the classifier can better understand them. A group of researchers has created a system that uses a range camera. This camera is able to take measurements of objects and then create a three-dimensional image of them. The range camera is a type of camera that can see the depth of objects. This means that it can see how far away an object is at every pixel. The distance of 3D points is determined by a Time-of-Flight (TOF) approach using modulated infrared light. They used a finite state machine to recognize the 3d trajectories [9]. Some researchers at al-Azhar University used a Kinect camera to try and recognize signs language for 42 different medical words. They did this by using two different methods, which are called the K-Nearest Neighbors method and the Support Vector Machine method. This means that the system makes sure everyone is the same size and in the same position so that there are no arguments. The system was successful in recognizing 89% of the images for the KNN classifier with majority voting. The segmentation accuracy reached 91%. The system was trained on 840 samples and tested on 420 samples.

[10] Developed a system that uses a gesture recognition pipeline. The article is discussing a system that uses different image processing techniques like segmentation, tracking, and filtering to normalize orientation and extract features for classification. After the hand region is extracted, it is compared to a set of learned templates that are stored in a template database. The system then outputs the result of the comparison, which is used to determine what gesture the user is performing. The hand gesture recognition system invention helps. This means that the vectors (which are a type of mathematical information) are being used to figure out what hand gestures are being made. In Germany, a group of researchers [12] concerned to recognize German sign language using Dynamic Time Warping (DTW) algorithm for sequence matching, and Visual Gesture Builder (VGB) along with DTW to avoid comparing every gesture in the DTW dataset. The system detected most of the gestures (65.45%), but also had a lot of false positives (10. 91%). In [13], The researchers are investigating the Kinect camera for making an approach that converts the sign language to spoken language. The dataset has 2890 entries. The system incorporates hand movement, form, position

and orientation. They are used to convert Kosovo Sign Language into speech. This means that the computer was able to correctly identify static signs (ones that were not moving) 90% of the time. While in [14] researchers used Kinect to recognize Indian sign language and the 90% accuracy was for 13 out of 16 signs. The used dataset for this study has 36 classes and 5041 images.

3 Convolutional Neural Network Design

Through this study, the used research method is adopted from [15] and [16]. The used CNN mode utilizes keras library. Convolution is a special type of linear operation used for feature extraction, where a small array of numbers, called a kernel, is applied across the input, which is an array of numbers that is called a tensor. The developed CNN contains 10 layers that designed with LeNet and AlexNet styles (see figures 1 & 2). These layers as following are consequences (1)first conv2d layer with relu, (2) first maxpooling2d, (3) second conv2d layer with relu, (4) second maxpooling2d, (5) third conv2d layer with relu, (6) third maxpooling2d, (7) flattening, (8) first dense layer with relu, (9) dropout, and (10) second dense layer with softmax.

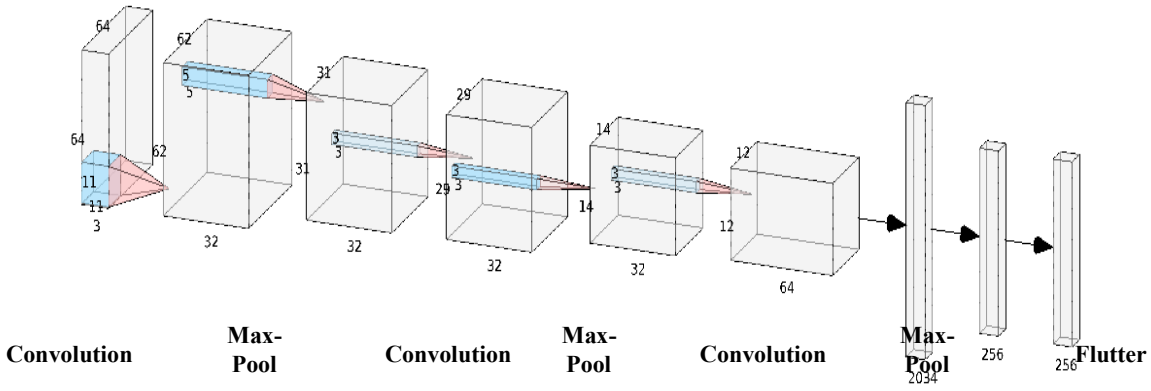


Fig 1: CNN Model (AlexNet Styl

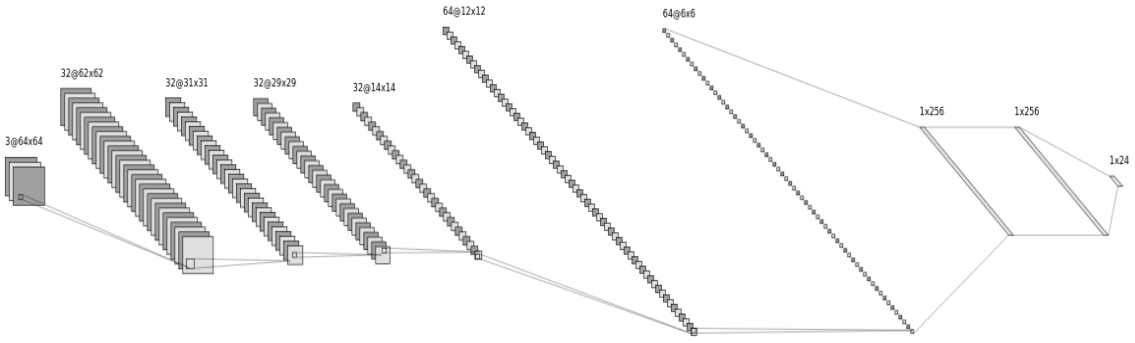
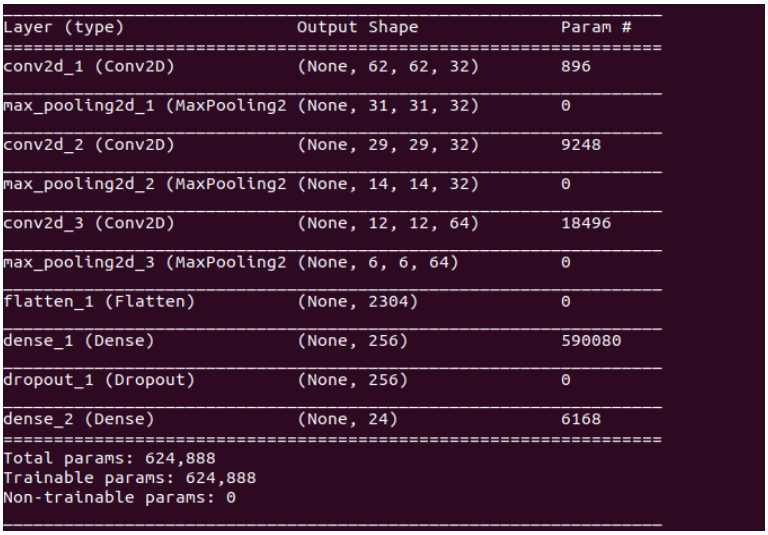


Fig 2: CNN Model (LeNet Style)

The first perform is set feature extractor to be hierarchically composed with multiple learned layers and conducted learnable parameters for obtaining from high layer to lowest layer in Alex Net style. Meanwhile. The LeNet architecture is a simple design that uses fewer resources, making it a good choice for teaching the basics of CNNs. It is composed of three parts: convolution, pooling, and nonlinear activation functions.

3.1 CNN Summarized & Visualized Models

After the network model is defined, the model is summarized using summary function. Figure 3 shows the summary model. It indicates the layers and their order, output shape of each layer, number of weights and total number of weights in model.



Layer (type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 62, 62, 32)	896
max_pooling2d_1 (MaxPooling2	(None, 31, 31, 32)	0
conv2d_2 (Conv2D)	(None, 29, 29, 32)	9248
max_pooling2d_2 (MaxPooling2	(None, 14, 14, 32)	0
conv2d_3 (Conv2D)	(None, 12, 12, 64)	18496
max_pooling2d_3 (MaxPooling2	(None, 6, 6, 64)	0
flatten_1 (Flatten)	(None, 2304)	0
dense_1 (Dense)	(None, 256)	590080
dropout_1 (Dropout)	(None, 256)	0
dense_2 (Dense)	(None, 24)	6168

Total params: 624,888
Trainable params: 624,888
Non-trainable params: 0

Fig 3 Summarized Model

Through the summary model, the first convolution provides 62 and 62 as a size for the output image, and 32 to show number of filters in this stage. The Nonparametric is related to the batch size which it is None in the beginning. The Max-pooling layer aims to decrease the size of the feature map; the input of the first layer is divided by two and the filters remain 32.the flatten layer is hen applied to transform pooled feature matrix into single vector that is passed to the fully connected layer. Dense layer will add the fully connected layer to the neural network. Dropout process is added to promote redundancy in the weight matrix and regularize deep neural networks. Keras API also come up with another function to facilitate the understanding of complex neural network. In figure 4, a visualize plot model was created to show the order of the layers in the model. The visualize model helps to add a new layer in the wrong order, emphasize the output shape of any layer, and confirm parameters [17].

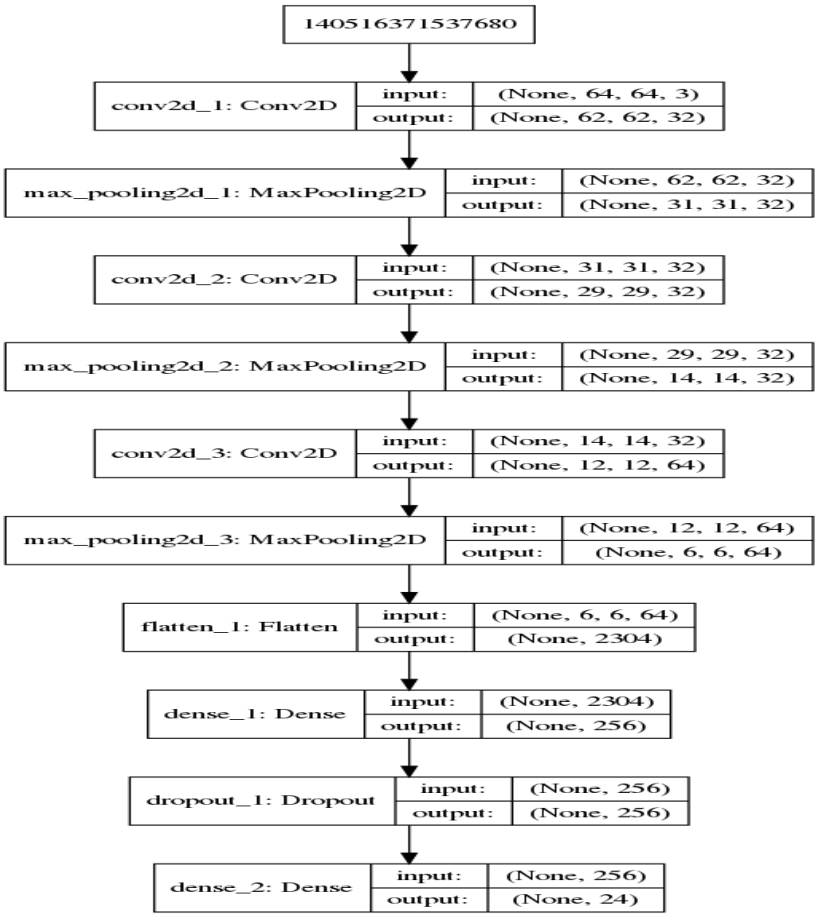


Fig 4: CNN Visualized Model

3.2 Training & testing processes

The dataset for this project is created by researchers based on American Sign Language. The data includes 62000 (64x64 pixel) images of the 30 letter of Alphabets and 1 Word. It is split into training and test sets, and each letter or word contains 1750 training images and 250 testing images in png format and HSV color model. Figures 5 & 6 show the processes of capturing 2000 images of letter C. However, capturing images for dataset was one of the hardest parts of creating our system.

```
mojtaba@mojtaba-GL553VD:~$ cd Desktop
mojtaba@mojtaba-GL553VD:~/Desktop$ cd SignL
mojtaba@mojtaba-GL553VD:~/Desktop/SignL$ ls
1.png  CNN_Train.py  dataset  DatasetCapture.py  Model.h5  Start.py
mojtaba@mojtaba-GL553VD:~/Desktop/SignL$ python DatasetCapture.py
Enter gesture name: "C"
```

Fig 5: Process of creating gesture of letter C

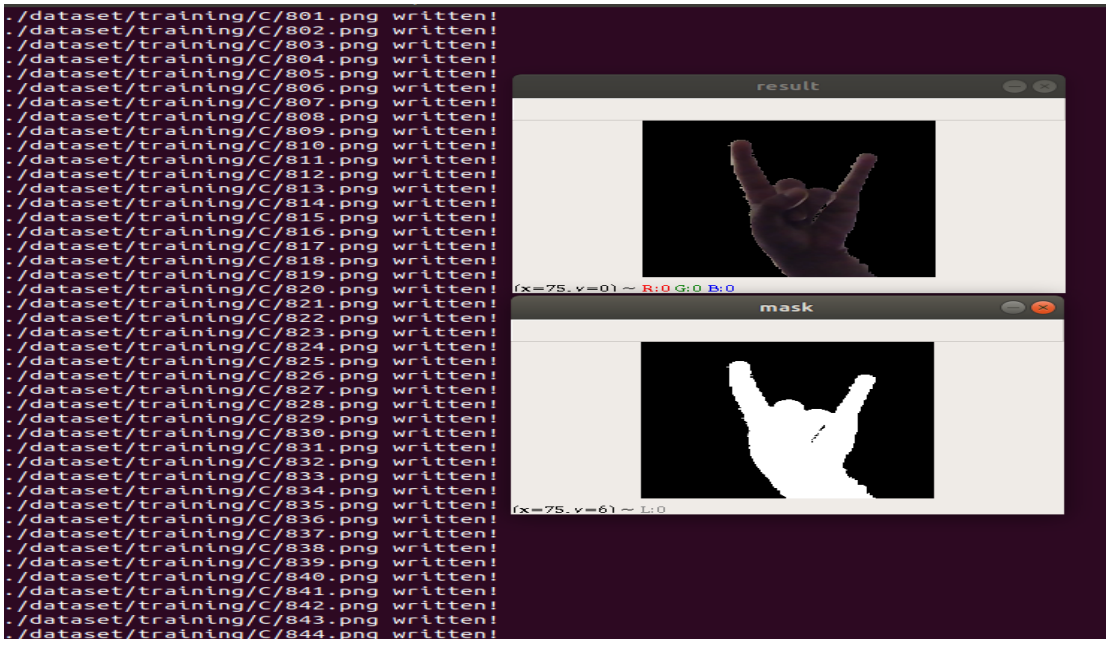


Fig 6: Capturing Gesture of letter C

To solve skin color and background issue for creating and performing sign language dataset, HSV Color Model (figure 7) is utilized to gain exact sign.

Fig 7: HSV vs RGB Color Model

4 Proposed System

The proposed system is detected the deaf hand sign using camera from online video, the detected RBG sign image is converted to HSV (Hue, Saturation, Value) image, the HSV image is verified, and the features are extracted and recognized based on the developed database (The database was built by researchers) and CNN model. Finally, the sign is showed as a text and speech. The system can run using raspberry pi 3 and above. Figure 8 illustrates the use case diagram for the proposed system.

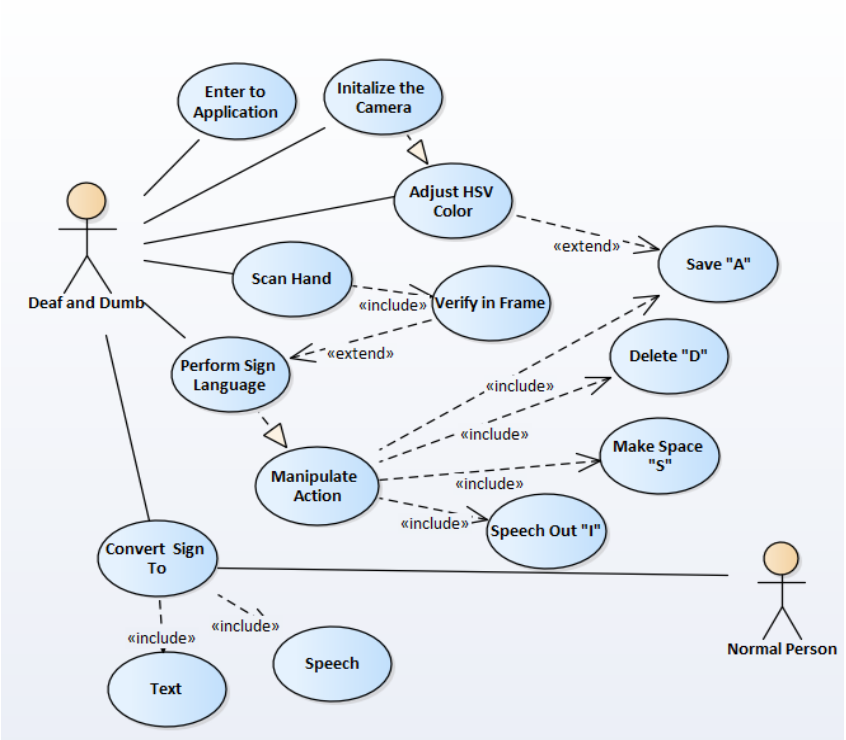


Fig 8 uses Case Diagram for Sign Language Recognition System

The use case diagram illustrates the behavior described for the sign language recognition system; it includes two actors and fourteen use cases. The type of actor is human, with the core called deaf and dumb, who interacts with all functions of the system as well as for the purpose the system provides him with the solution, the second actor is called normal person who can write text and/or speech. The life cycle of the deaf and Dumb through the system is as follows:

1- Enter the Application: the interface of application is compatible with actor conditions with easy to access camera.
2- Initialize the camera: Turn on the camera is a role to capture the hand sign, it presents as input to system, the sub-function of this function is Adjust HSV color carry on by three aspects are hue, saturation, and lightness.

3- Adjust HSV color is modified refer to time and location for each experiment request specific setting up color then save it.
4- The value of input to the camera is a scan hand that belongs to the frame boundary that must be verified to conduct accurate results unless the rest function will leak.
5- To capture sign language from the hand of the deaf and dumb based on HSV color to recognize it as a series of images.
6- The manipulate Action consist of save actions that interlink the image, catch up to sign, or spilt it by make space, speech out present by voice, or delete if decide the image ambiguity.
7- Convert to text or speech in multifaceted use between deaf and dumb, and normal people, which function able to present two ways forward and backward.

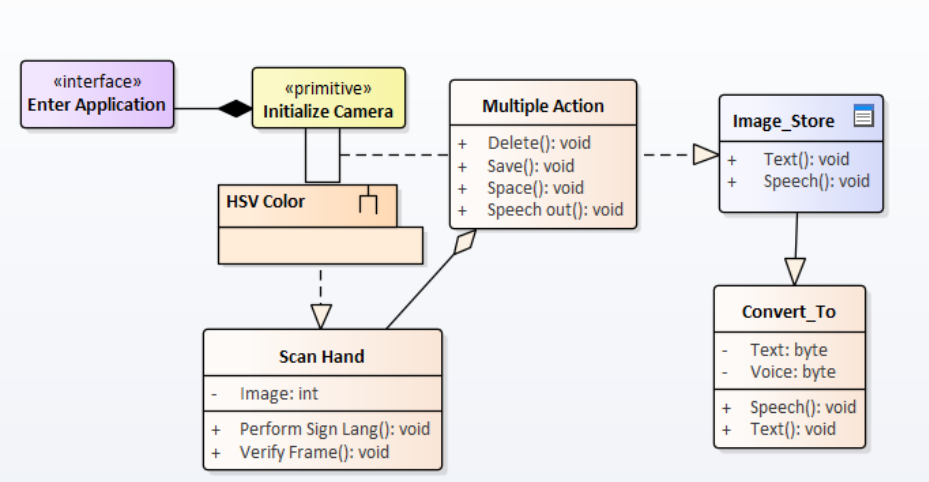


Fig 9 Component Model of Sign Language Recognition System

The Sign Language Recognition System comprise package, classes, interface, and pipeline, the interface role treats input and output image specifically the input and output type multiple types as image, and voice text, where the forward process determine those type that request composite relationship with package and classes; the image capture request camera need adjust HSV color to determine the frame of the image and acquire the features of an image then multiple action class involve the primary function is save that association relationship with data base belong to conditions rules the convert to class able to apply.

The proposed system falls into four main components which is compromises an HSV color space changer, main frame, sign frame and black sequence. Through the system executing, the user needs to run the terminal from start menu or press Ctrl + Alt + T, chooses the system directory and run Start.py into Terminal, and then enters the camera index to initialize the camera. Check figure 10.

```
mojtaba@mojtaba-GL553VD:~$ cd Desktop/SignL
mojtaba@mojtaba-GL553VD:~/Desktop/SignL$ ls
1.png  CNN_Train.py  dataset  DatasetCapture.py  Model.h5  Start.py
mojtaba@mojtaba-GL553VD:~/Desktop/SignL$ python Start.py
Start.py:103: SyntaxWarning: name 'is_voice_on' is assigned to before global dec
laration
  global is_voice_on
Using TensorFlow backend.
2020-05-23 16:03:12.693438: I tensorflow/core/platform/cpu_feature_guard.cc:137]
 Your CPU supports instructions that this TensorFlow binary was not compiled to
use: SSE4.1 SSE4.2 AVX AVX2 FMA
Enter Camera Index : 0
```

Fig 10: System Execution

The user also has the option to check the available camera index in case there is another video device by running the command "ls –ltrh /dev/video*", "lsusb" to list devices attached to USB, or "lspci" to lis devices attached to PCI connector (This option support Arduino devices). ColorSpace Changer utilized to filters out a range of color and leaving you with range of color needed to detect hand. Different time and locations with different lightning need different setup. As mentioned in figure 10, usually the Upper HSV and Lower H&V are kept as it is and Lower S is chanced to detect the hand [18].

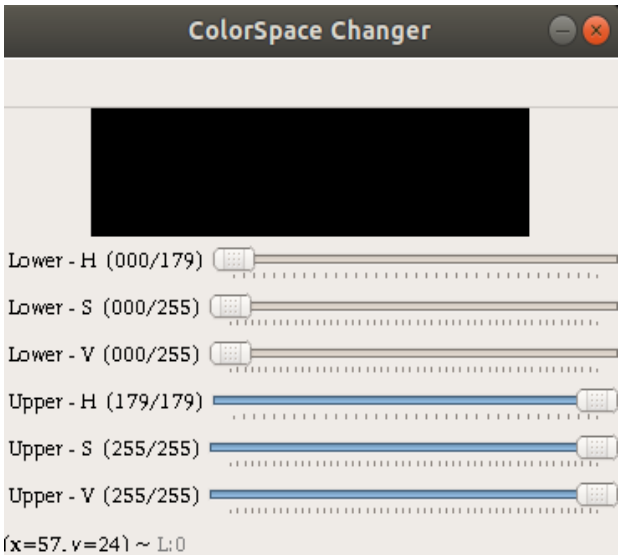


Fig 12: ColorSpace Changer to setup HSV

After HSV is setup, the hand is located inside the green box (as shown in figure 11), then the sign language is performed in the main frame. Letter or word related to each sign is showed at downside of left corner.

Fig 12: Zoom in process

Fig 13: Sign Frame

5 Results and Validation

After Defining CNN architecture, researchers fit the model into training and test sets to optimize the model performance. The number of epochs is a hyperparameter that defines the number times that the learning algorithm will work through the entire training dataset, [19], [20] & [21]. The value of epoch is set to equal to 25 in this project. The model achieves 0.9988 accuracy, and the loss was almost 0.0015. see figure 14.

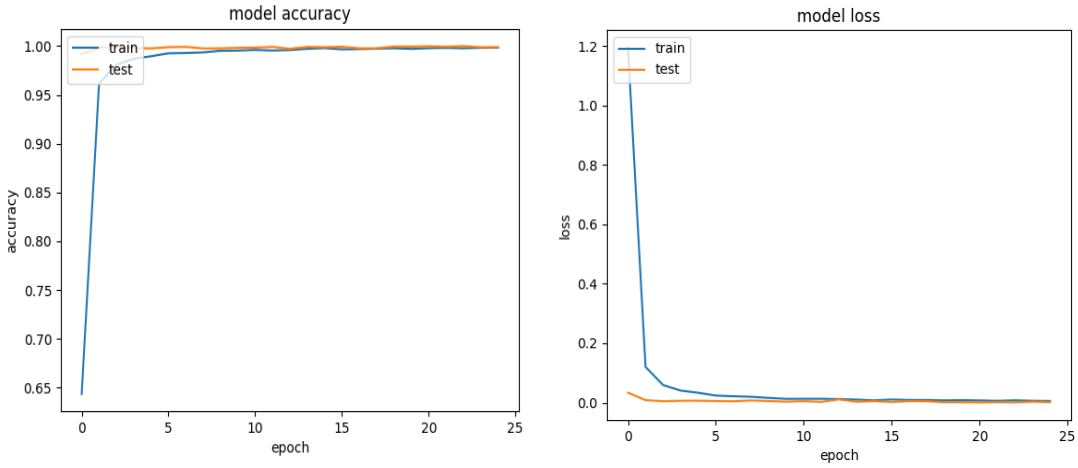


Fig 14: Model Accuracy

As a result, for the proposed system, the signs are then shown in the output board. Figure 15 A, B & C show a sample of different converted examples, where a combination of signs formulates the fully English sentence.

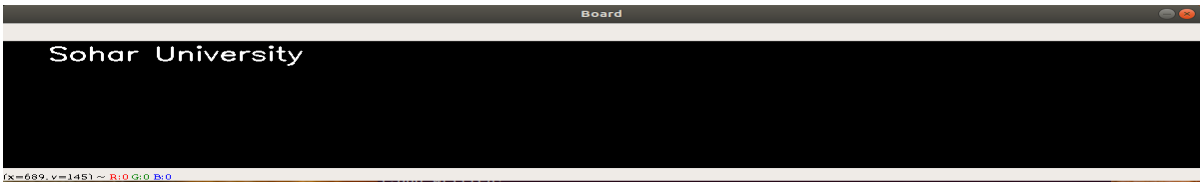


Fig 15A: Output board

Fig 15B: Output board

Fig 15C: Output board

The user can use the following keys in the output board:

“a” to save and shown letters into board.

“s” to make space between words.

“d” to delete the words.

“f” to speech out the text on the board using pyttx .

The researcher also developed a special remote that can control the main system commands. The proposed system reads the text on the output board by using pyttsx (text to speech) function [22]& [23], the user can choose the preferred sound for the speech whether male or female. See figure 16.

```
is_voice_on = True

def say_text(text):
        if not is_voice_on:
                return
        while engine._inLoop:
                pass
        engine.say(text)
        engine.runAndWait()
```

Fig 16: Pyttsx function

6 Conclusion

A recognition hand gestures for American sign language is presented in this paper. The characteristics recovered from the sign image are used to train a sign recognition using CNN method. The system is using raspberry pi and normal camera to detect and convert deaf sign. The system could identify 31 hand gestures, including the letters A to Z and a unique word "I love". Through this project, a dynamic recognition system was proposed to recognize the sign pattern from live video that is taking from camera. The researchers built their own dataset to test the proposed CNN model. The proposed CNN model used 25 epochs to provide high image training quality. The Model achieves 0.9988 accuracy and almost 0.0015 loss.

7 Acknowledgment

This project was supported by Award Number SU/ BFP/ URG12 from Ministry of Higher education, research and innovation (MoHESRI)in Sultanate of Oman under TRC program. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of MoHESRI.

8 References

[1] World Health Organization. Deafness and hearing loss. Geneva: WHO. (2019).
[2] Umarani, C., J. Naskath, M. Banu, and J. Uhaya Shanthi. "Embedded Based Deaf Mute Communication System." (2017).
[3] Abualkishik, A.M., Omar, K. and Odiebat, G.A. QEFSM model and Markov Algorithm for translating Quran reciting rules into Braille code. Journal of King Saud University-Computer and Information Sciences, 27(3), pp.238-247.2015.
[4] Alsinglawi, Belal, Mahmoud Elkhodr, Quang Vinh Nguyen, Upul Gunawardana, Anthony Maeder, and Simeon Simoff. "RFID localisation for Internet of Things smart homes: a survey." arXiv preprint arXiv:1702.02311 (2017).
[5] Rouse, Martyn, and Natallia Yakavets. "10 TOWARDS INCLUSIVE EDUCATION: Swimming Against the Tide of Educational Reform." Education reform and internationalisation: the case of school reform in Kazakhstan (2014): 196.
[6] Chakraborty, Biplab Ketan, Debajit Sarma, Manas Kamal Bhuyan, and Karl F. MacDorman. "Review of constraints on vision-based gesture recognition for human–computer interaction." IET Computer Vision 12, no. 1 (2018): 3-15.

[7] Xu, Pei. "A real-time hand gesture recognition and human-computer interaction system." arXiv preprint arXiv:1704-07296. 2017.
[8] AlKishri, Wasin, Abdallah Abualkishik, and Mahmood Al-Bahri. "Enhanced Image Processing and Fuzzy Logic Approach for Optimizing Driver Drowsiness Detection." Applied Computational Intelligence and Soft Computing 2022 (2022).
[9] Abualkishik, Abdallah M., Khairuddin Omar, and Ghadeer A. Odiebat. "QEFSM model and Markov Algorithm for translating Quran reciting rules into Braille code." Journal of King Saud University-Computer and Information Sciences 27, no. 3 (2015): 238-247.
[10] Kurakin, Alexey, Zhengyou Zhang, and Zicheng Liu. "A real time system for dynamic hand gesture recognition with a depth sensor." In 2012 Proceedings of the 20th European signal processing conference (EUSIPCO), pp. 1975-1979. IEEE, 2012.
[11] Smith, Anthony Vernon Walker, Alistair Ian Sutherland, Arnaud Lemoine, and Sean Mcgrath. "Hand gesture recognition system and method." U.S. Patent 6,128,003, issued October 3, 2000.
[12] Amatya, P., Sergieieva, K. & Meixner, G. Translation of Sign Language Into Text Using Kinect for Windows v2. The Eleventh International Conference on Advances in Computer-Human Interactions, Rome, Heilbronn University. ISBN: 978-1-61208-616-3. 2018.
[13] Edon, M. & Dimopoulos, K., 2014. Sign Language Recognition using Kinect. Thessaloniki, The University of Sheffield.
[14] Velugula, D. S., "An IOT Integrated Gesture Recognition Using Image Processing For Speech Impaired People". International Journal of Advance Research, Ideas and Innovations in Technology, 2(6). 2017
[15] Browniee, J., 5 Step Life-Cycle for Neural Network Models in Keras. [Online]
Available at: https://machinelearningmastery.com/5-step-life-cycle-neural-network-models-keras/. 2016. [Accessed 16 May 2021].
[16] Al-Kasasbeh, Basil. "Artificial Intelligence Scheme for Medical Images Classification and Prediction." International Journal of Advances in Soft Computing and its Application 14, no. 2 (2022).
[17] Zreik, M., Leiner, T., De Vos, B.D., van Hamersvelt, R.W., Viergever, M.A. and Išgum, I., April. Automatic segmentation of the left ventricle in cardiac CT angiography using convolutional neural networks. In 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI) (pp. 40-43). IEEE. 2016.
[18] Khder, Moaiad Ahmad, Mohammad Adnan Sayfi, and Samah Wael Fujo. "Analysis of World Happiness Report Dataset Using Machine Learning Approaches." Int. J. Advance Soft Compu. Appl 14, no. 1 (2022).
[19] Jason Brownlee.Difference Between a Batch and an Epoch in a Neural-Network. Machine Learning Mastery. https://machinelearningmastery.com/difference-between-a-batch-and-an-epoch/.2021.
[20] Almomani, Omar, Mohammed Amin Almaiah, Adeeb Alsaaidah, Sami Smadi, Adel Hamdan Mohammad, and Ahmad Althunibat. "Machine learning classifiers

for network intrusion detection system: comparative study." In 2021 International Conference on Information Technology (ICIT), pp. 440-445. IEEE, 2021.
[21] AlKishri, Wasin, Abdallah Abualkishik, and Mahmood Al-Bahri. "Enhanced Image Processing and Fuzzy Logic Approach for Optimizing Driver Drowsiness Detection." Applied Computational Intelligence and Soft Computing 2022 (2022).
[22] Adewale, Victoria A., and Adejoke O. Olamiti. "Conversion Of Sign Language To Text And Speech Using Machine Learning Techniques." (2018).
[23] Mumtaz, Rusul, Venus Samawi, Aysh Alhroob, Wael Alzyadat, and Ikhlas Almukahel. "PDIS: A Service Layer for Privacy and Detecting Intrusions in Cloud Computing." Int. J. Advance Soft Compu. Appl 14, no. 2 (2022).

BIOGRAPHIES OF AUTHORS

Abdallah Mohd Abualkishik is currently an assistant professor at the Faculty of Computing and Information Technology, Sohar University. He received his his Ph.D. in Artificial Intelligence & Software Engineering in 2011 from Malaysia National University (UKM), Malaysia. He has published dozens of research papers in different fields of computer science. His research interests include artificial intelligence, information technologies, internet of things, big data and smart systems to solve real-life, and community problems.

Wael Jumah Alzyadat is currently an assistant professor at Department of Software Engineering/Faculty of Science and IT, Al-Zaytoonah University, Amman, Jordan. He received his Ph.D. in Software Engineering in 2014 University Putra Malaysia (UPM), Malaysia. He has published research papers in different fields of computer science. His research interests include Big Data, Machin learning, Genatic Algorithm, information technologies, internet of things, and Cloud computing.

Marwan Al Share is currently an assistant professor at the Faculty of Computing and Information Technology, Sohar University. He received his Ph.D. in Cyber Security 2016 from Malaysia National University (UKM), Malaysia. He has published different research papers in the fields of computer security. His research-interests include Cyber Security, Information system, Modelling and visualization, and 4th Industry technologies studies.