Paper ID: 566

Finger spelled Sign Language Translator for Deaf and Speech Impaired People in Srilanka using Convolutional Neural Network

HKK Perera#, DMR Kulasekara, Asela Gunasekara

Department of Computer Science, General Sir John Kotelawala Defense University, Sri Lanka.

#adorekasun@gmail.com

Abstract: Sign language is a visual language used by people with speech and hearing disabilities for communication in their daily conversation activities. It is completely an optical communication language through its native grammar. In this paper, hoping to present an optimal approach, whose major objective is to accomplish the transliteration of 24 static sign language alphabet words and numbers of Srilankan Sign Language into humanoid or machine decipherable English manuscript in the real-time environment. Since Srilanka has a native sign language deaf/Signers become uncomfortable when expressing their ideas to a normal person which is why this system is proposed. Artificial Neural Networks (ANN) and Support Vector machines (SVM) have been used as the technologies of this proposed system. Pre-processing operations of the signed input gestures are done in the first phase. In the next phase, the various region properties of the pre-processed gesture images are computed. In the final phase, based on the properties calculated of the earlier phase, the transliteration of signed gesture into text and voice is carried out. The proposed model is developed using Python and Python libraries like OpenCV, Keras, and Pickle.

Keywords: Artificial Neural Networks, Static gestures, Gesture recognition, Support Vector Machines, Gesture Classification

Introduction

Now a day's technology has advanced to help people with any kind of disability. It almost makes IoT devices and there are robots to help people in every situation. People who are suffering from hearing impairments, struggle daily in communicating their ideas to other people. Essentially to the oral language, sign language has a lexical, a "phonetic". (Saldaña González et al., 2018) (rather than verbalized sounds it has enunciated signs), a "phonology" (rather than phonemes, it has components from various natures that achieve a similar differential capacity from the words visual structure), punctuation, a semantic, and it's very own practice. Being a trademark from every nation and culture, and not widespread permits portraying all the truth that includes us, what users see, feel, or think. In order to fill that gap in Srilanka, Video Relay Service (VRS) (Gebre et al., 2014) is used in Srilanka for now. In VRS what happens is, a human interpreter translates the hand signs to Voice and the same is the inverse. The problem that lies there is that the communication speed is kind of slow and there is no privacy at all. Many researchers have been carried out to solve this problem in Srilanka, but the problem is that Srilankan sign language is not static. Meaning that the same word can be interpreted in many ways. Just as speakers have different voices, signers have different signs. (Gebre et al., 2014) And also the camera induces scale, translation, and rotation errors which may make noises. (Saldaña González et al., 2018)

This system is carried out using the camera, so the environmental conditions also affects but some systems don't care about the environmental conditions because they are carried out with a wearable device.

Sri Lankan Sign Language (SSL) consists of 56 characters and overall, it has about 1000+ signs for now and its increasing day by day because people tend to form new signs for the words they use usually. According to the census results of housing 2012, there are a total of 569910 people including people with both hearing and speech impairments in Srilanka. Since the inability to convey the messages from the normal people and them they have lost the right to live a normal life. This paper suggests a method to solve that barrier and break the unfairness between the people. (Garcia and Viesca, n.d.) This system is proposed to find solutions for the following facts about the previous systems made.

- Environmental concerns.
- Occlusion (Detecting signs which are made from the reference box).
- Coarticulation (Signs are affected by the sign which is made before or after).

This proposed system only identifies the manual signs which are made by the hands (fingers). Taking ASL (American sign language) as a reference, where the dictionary contains 7154 sign words that visually look the same, which can lead to miscommunication. For example, signs for 'Chocolate' & 'Cleveland' are similar, they are not the same at least they cannot get even close. It is very hard to catch when a conversation is messed up (Mahesh Kumar et al.., 2018). Basically, Communication via gesture recognition is a significant utilization of motion recognition. Gesture-based communication acknowledgment has two unique approaches.

1. Glove Based Approach
2. Vision-Based Approach

Mainly there are 2 main gesture types in Sinhala sign language. (Punchimudiyanse and Meegama, 2017) One is Conversational type, that has a set of sign gestures for common words and phrases which are made dynamically for the ease of use. And the second one uses fingerspelling alphabet to decode Sinhala words that are not there in the alphabet using letter by letter. For example, the word "Janaki/ජානකී" is spelled as J + A + N + A + K + I when converted to SSL its, ජ් + අ + න් + අ + ක් + ඉ.

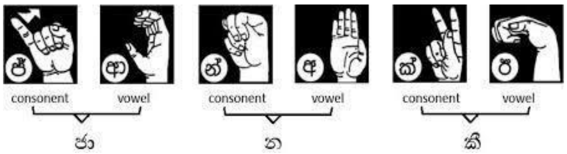

Figure 1: Sinhala Finger Spelling Word

Talking about the numbers in SSL, the basic numbers which normal people use is valid as it is and special number patters are available for,

- Numbers from 0 to 19.
- Signs from 20 to 90.
- Signs from 100 to 900.
- Signs for 1000, 100 000, Million and Billion.

The proposed system will help the deaf and speech impaired individuals in catching their sign-based message by means of the portable camera and afterward convert it into Sinhala content. Consequently, like average talking, it will be simple for them to speak with any individual ~~at~~ wherever and complete their work easily. Subsequently, the hardships they experience during the commitment to talk will be abridged.

This article is organized as follows: Section II presents the Related work on sign language recognition, Section III

Methods, and Approach, Section IV offers Design, Section V contains Results and

Evaluation and finally, section VI discusses the Conclusion and Future works.

Related Works

Many types of research have been conducted in the field of Sign Language Recognition using various novel approaches. Such as skin filtering technique, use of gloves, use of Microsoft Kinect sensor for tracking hand, Fingertip Search method, Hu Moments method, Convolutional Neural Network (CNN), shape descriptors, Grid-Based Feature Extraction technique in feature extraction, and Artificial Neural Network (ANN), etc. Sign language recognition can be mainly categorized into 2 main parts as,

A. Manual Signs

Manual signs are the signs which are made from the hands. It can again be categorized into 2 main parts,

1) Vision-Based Recognition: Image processing algorithms are utilized in a Vision-based system to recognize and track hand signs and outward appearances of the signer. This system is simpler to the signer since there is no compelling reason to wear any additional equipment. Notwithstanding, there are exactness issues identified with image processing algorithms, and these issues are yet to be adjusted. There are two distinct methodologies in vision-based sign language acknowledgment:

1. 3D model based
2. Appearance-based

3D model-based techniques utilize 3D data of key components of the body parts. Utilizing this data, a few significant parameters, like palm position, joint edges and so forth., can be taken. This methodology utilizes volumetric or skeletal models or a mix of the two. The volumetric technique is more qualified for the PC animation field and PC Computer vision. This methodology is exceptionally computational concentrated and furthermore, frameworks for live identification are still to be created (Escudeiro et al., 2014).

Appearance-based methods use images as information sources. A few formats are the deformable 2D layouts of the human pieces of the body, especially hands. The arrangements focus on the blueprint of an article called deformable layouts. This proposed system is an Appearance-based system. Some algorithms based on vision-based recognition are,

Support Vector Machine (SVM): American Sign Language recognition system which is proposed by Shivashankara (Department of Computer Science & Engineering, Sri Jayachamarajendra College of

Engineering, Mysuru, India et al., 2018) uses SVM for both numbers and letters and letters performed with an accuracy of 97.5% and numbers performed with an accuracy of 100%.Data was collected from the ASL which is a freely available dataset.

Artificial Neural Networks (ANN): Recognition and classification of sign language systems (Saldaña González et al., 2018) proposed for Spanish was made using a glove while also carrying the same project on artificial neural networks giving an accuracy for the Ann 90%. And giving an error rate of 2.61% which is independent of the user. The same paper presented a system using a glove and ANN which was 97% independent of the user.

Convolutional neural networks. (CNN): Real-time ASL recognition system with CNN (Garcia and Viesca, n.d.) proposes a system that compromises over 90% of accuracy without using any motion-tracking gloves. This also uses heuristics and the letter classification was done in ConvNet. Dataset was consisting of 65000, 150*150px images each image with 2 of the same images which are color image and depth image.

SURF and Hu- Moment: Paper (Rekha Jayaprakash et al.., 2011) combines Speeded Up Robust Features (SURF) and Hu-moment invariant methods to get the maximum features extracted from the image. The system has taken a dataset of 3 different backgrounds and environmental conditions and with 50 videos and a Total of 600 images,400 negatives, and 200 positive images. Many Backgrounds were used here. Hand segmentation is done using the K Means clustering and skin color pixels are separated. To recognize the letters SURF and Hu moment is used while minimum Euclidean distance identifies the gesture. SIFT with KNN gave 68%, SURF with KNN gave 84.6% with 3 times faster. Word recognition accuracy was 96%. So overall SURF with KNN performed well.

2) Sensor-Based Recognition: Sensor-based recognition systems are those that were proposed to be made from the sensors which were made by a company like Microsoft Kinect. In this section, the signer wants to wear a glove before making his gestures. The disadvantage of this system is the signer always wants to wear the sensor hardware and the glove during the operation of the system.

Research and Implementation of a sign language recognition system using Kinect paper (Yuqian Chen and Wenhui Zhang, 2016) proposes a method that gives 89.6% accuracy by using relative distance, angle, and motion vector theories. The gesture was identified by the golden section search algorithm. Each word is signed by 7 signers each with 5 times color depth and skeleton info was taken by Kinect. The dataset was created using 25200 images and obtained the highest accuracy of 98.6% with a single hand and 70.3% with double hands.

From the above data, it can be concluded that the most accuracy is obtained through the vision-based approach and overall accuracy is good in convolution neural networks and that’s over 90%. And due to the hardships faced in using the glove-based method Srilankan culture won’t fit in that. Here is a summary of the most relevant and effective steps that recent studies have conducted, on the Sign language recognition systems so far made using the visionbased method.

- Image Process: Manipulating the image to machine reprehensible format
- Edge Detection: To spot the dark side of the hand.
- Thresholding: Converting resourceful image to binary format
- Grey Scale: to acquire the full image without color variations.
- Training the Model: appending the images to the dataset to increase the accuracy.
- Making Predictions: Making predictions from the trained data on the live feed

Methods and Approach

A. Classifier Development

1) Algorithm Overview: Proposed SSL classification is finalizing utilizing a convolutional neural system (ConvNet / CNN). CNNs are AI algorithms that have seen amazing achievement in taking care of an assortment of assignments identified with preparing recordings and pictures. Recently, the field has encountered a blast of development and applications in picture classification and Object detection. An essential bit of leeway of using such procedures stems from CNNs capacities to learn includes just as the weights compared to each feature. Like other machine learning algorithms, CNNs look to enhance a few target functions, explicitly the loss function. Here, a SoftMax-based loss function is used:

$$Loss = \frac{1}{N}\sum_{i=1}^{N} -\log\left(\frac{e^{f_{i,y_i}}}{\sum_{j=1}^{C} e^{f_{i,j}}}\right) \qquad (1)$$

$$f_j(z) = \frac{e^{z_j}}{\sum_{k=1}^{C} e^{z_k}} \qquad (2)$$

N = total number of training examples
C = total number of classes

Figure 2: SoftMax Equation

Above shown is the SoftMax equation.To produce total SoftMax loss, the equation takes feature vector z for the given training data and separates its values to a format of [0,1] and that values go to the 1st equation. Then equation 01 gives mean a loss for each training data. Utilizing a SoftMax-based classification head permits us to yield esteems like probabilities for every SSL letter. This contrasts from another famous decision: the SVM loss. Utilizing an SVM classification would bring about scores for every ASL letter that would not straightforwardly guide to probabilities. These probabilities stood to us by the SoftMax loss permit us to more naturally decipher our results and demonstrate value when running our classifications through a language model.

2) Transfer Learning: It is the reusing of a model that is previously built for a problem that is related to the current problem. In deep learning, transfer learning is a procedure whereby a neural system model is first prepared on an issue like the issue that was already solved. At least one layer from the prepared model is then utilized in another model prepared on the issue of interest (at least one layer should be updated by changing parameters). The most essential case of this would be a completely prepared network whose last order layer weights have been changed to have the option to characterize a few new arrangements of data. The essential advantage of these methods is they require less time and fewer data processing. Notwithstanding, the test in transfer learning comes from the contrasts between the first data used to prepare and the new data being characterized (new training dataset). Bigger deviation in these data sets regularly requires re-introducing or expanding learning rates for deeper layers in the net. In simpler words, it's using a pre-trained model. Perhaps the most popular pre-trained models are,

01.VGGnet which is also known as VGG16 or VGG19

02.Google Net (ex.Inception v3)

03.Residual Network. (Resnet50)

From the above mentioned pre-trained models, the proposed system uses the VGG net.

3) VGG (VGG16 or VGG19): VGG model was developed by a team called Visual Graphics group in oxford which is described in the paper (Garcia and Viesca, n.d.). mainly VGG has consistent and repeating steps that can integrate into the proposed model. By default, images should be scaled to 224*224 squares but in this project, have modified and inserted images of 96*96 px to increase the accuracy of the system.

B. General Techniques.

The main target of this phase is to finetune the VGG model according to the images which have got in the prepared SSL dataset. The data which is collected for this project is on 20 signs (since this is an ongoing project) which is completely different from one another in one class of dataset, it has got the same sign which is differentiated in backgrounds and different angles and orientations. Then to test the effectiveness of the system, changed the weights, depths, and stepped-up the learning rates of the neural network algorithm by a factor of 10 (roughly).

C. Developing Pipeline

To take the Realtime signs done by the signer, made a desktop application and a mobile application where the desktop application uses OpenCV for capturing images and the mobile application uses java(android). when capturing images, the system captures images in frames and then sends the frame to the model and then takes the output of the model which is the prediction of what the sign is. When developing the model to fit the mobile application, first made the dataset and the model using the computer and then converted the model file (format:h5) to the '.tflite' format which can use it to the mobile application without any delay due to performance errors.

1) Model Design: In the proposed model, there is only one convolution layer and it has 32 kernels of size 3*3 to extract features from the input image. And for the subsampling process, this convolution layer is then followed by a 3*3 max pooling layer with a stride of 5. After pooling, the feature map obtained will be flattened into a vector and fed into the fully connected layer (FC layer). In order to classify the outputs, ReLU and SoftMax activation functions have been used in the FC layer. A dropout regularization of 20% has been used to cut off the images temporarily in each update cycle to reduce overfitting in the model. A diagram of the proposed model is given in Fig 3.

Design

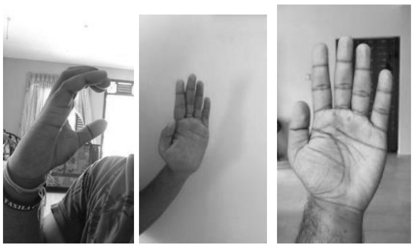

Figure 3: Left-ආ Middle-ඇ Right-ඇ

A. Dataset Description

Datasets of this sign language can be mainly separated into 2 categories.

01.Color images

02.Depth images

To use depth images, it needs to have another camera that has depth sensors that are not available in most of the cameras as well as web cameras. So, used color images to develop the dataset. The dataset comprises 20 signs which are taken from 5 people and in many backgrounds. Altogether it contains about 50,000 RGB images. Some are closeups of the hand sign while some are a bit distanced. Since data of 5 people have been collected, have used a mixed combination to train, validate, and to test the data. To take the signs of the user in real-time made a different application that accesses the camera of the laptop and takes the video and then splits it into frames. it was made using python and OpenCV (cv2) and stores captured images in the local storage. Then again also took some photos using my mobile camera (12MP) and in there, one picture was around 5mbs which is too much for a dataset which took many gigs when completing the dataset. Therefore, it was decided to take a video the same as in the desktop version and split them to the frames in which one image was around 80kb. Mainly the dataset contains the static images of the Sinhala sign language.

1) Preprocessing Data: Since the images were taken by the video splitting have many errors like images are in different sizes and so on. Then changed the image size to 270*480 px to match the input image size of the VGG net which has overwritten. furthermore, then made a horizontal flip to all images to take the right hand and left-hand effect. And, applied some zooming effect to the images to differentiate between the distances. Since the dataset needs to have

variety in it, also changed the qualities of the images, some were increased, and some were decreased. Most of the images are in the quality range of 65.5%.

2) Loading Dataset: The whole dataset was divided into three categories as train, validate, and test datasets. A ratio of 60:20:20 has been selected to separate out the train, validate and test datasets, respectively. To specify in terms of each hand sign, there were 1200 images in each training dataset, 400 images in each validate the dataset, and again 400 images in each test dataset.

B. Cnn Model

1) Proposed Model:

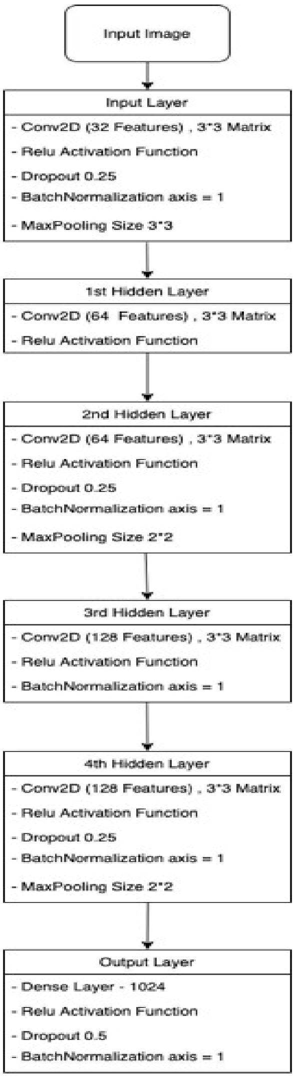
Input Image
Input Layer
- Conv2D (32 Features) , 3*3 Matrix
- Relu Activation Function
- Dropout 0.25
- BatchNormalization axis = 1
- MaxPooling Size 3*3
1st Hidden Layer
- Conv2D (64 Features) , 3*3 Matrix
- Relu Activation Function
2nd Hidden Layer
- Conv2D (64 Features) , 3*3 Matrix
- Relu Activation Function
- Dropout 0.25
- BatchNormalization axis = 1
- MaxPooling Size 2*2
3rd Hidden Layer
- Conv2D (128 Features) , 3*3 Matrix
- Relu Activation Function
- BatchNormalization axis = 1
4th Hidden Layer
- Conv2D (128 Features) , 3*3 Matrix
- Relu Activation Function
- Dropout 0.25
- BatchNormalization axis = 1
- MaxPooling Size 2*2
Output Layer
- Dense Layer - 1024
- Relu Activation Function
- Dropout 0.5
- BatchNormalization axis = 1

Figure 4: Architecture of the Proposed Model

2) Training and Validation Phase: Mainly the model was trained and tested many times by changing parameters and variables which the results are included in the results in the Results part. Main cases where I changed the parameters are,

- Number of features to be extracted
- Activation function
- Matrix Size in which the image is being classified.
- Dropouts size
- Number of Hidden Layers etc.

The main activation function which I used in the project is "Relu" and the matrix size is 3*3.

3) Test Phase: The model I got after the training process freely as other common people in the society irrespective was tested with the images of the testing dataset which are of the language barrier. The technique pursues a visioncompletely different from the training dataset and gave based sign recognition framework to perceive static, good results.

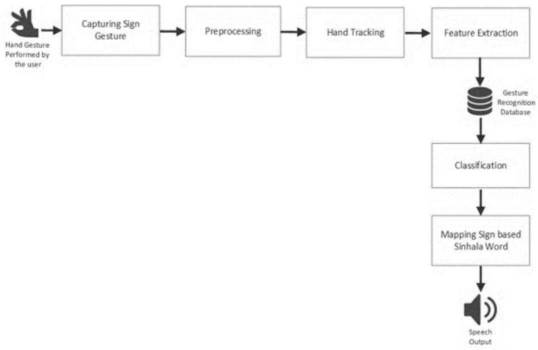
Hand Gesture Performed by the user
Capturing Sign Gesture
Preprocessing
Hand Tracking
Feature Extraction
Gesture Recognition Database
Classification
Mapping Sign based Sinhala Word
Speech Output

Figure 5 High-Level Architecture of Proposed System.

Results and Evaluation

After checking the graph received for the first training using the separated frame images the validation accuracy was always around 3.54 while the training accuracy was 99.34% always. This was the overfitting condition faced in the dataset and then reduced some images of the database to overcome the challenge. Initial Configs are as below,

- Epochs = 3, Batch Size = 32, Image dimensions = (96, 96, 3)
- Result: loss: 0.3269 - acc: 0.8870 - val_loss: 0.0521 - val_acc: 0.9827

To overcome the overfitting drop out of the final layer was increased from 0.25 to 0.5. And, the neurons of the final layer were also increased to 1024. Final configs are,

- Epochs = 8, Batch Size = 64, Image dimensions = (96, 96, 3)
- Result: loss: 0.0731 - acc: 0.9743 - val_loss: 0.2114 - val_acc: 0.9322

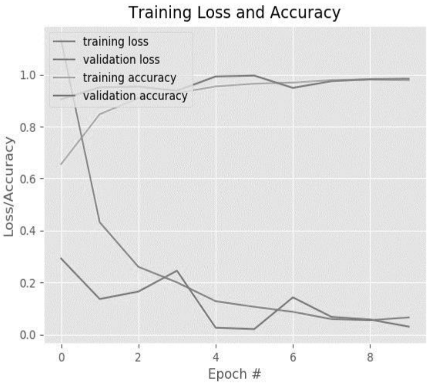

Figure 6: Final Model Details

Conclusion and Future Works

The aim of this research is to develop a system for the deaf and speech impaired Sri Lankans and to help them behave dynamic, and fingerspelling expressions of SSL. This methodology is conservative and can be executed even with a portable camera which makes it very easy to use to be utilized by a typical man. The presentation of sign-based strategy created exceptional outcomes in unique motion acknowledgment. Because of the absence of accessibility of the dataset in SSL, another dataset is made which incorporates static gestures, dynamic gestures, and fingerspelling letters in order. Examination results show great acknowledgment precision for static and dynamic gestures, fingerspelling words, and co-explanation discovery and disposal. On the way to solve that bigger problem, as the initial step, a system was built to recognize the static gestures of the Sinhala alphabet with the future of information procurement of hand gestures indicating NUI (Natural User Interfaces) utilizing profundity sensors.

Acknowledgement

The authors would like to sincerely thank all who rendered their support in this study.

References

Athira, P.K., Sruthi, C.J., Lijiya, A., 2019. A Signer Independent Sign Language Recognition with Co-articulation Elimination from Live Videos: An Indian Scenario. J. King Saud Univ. - Comput. Inf. Sci.

S131915781831228X. https://doi.org/10.1016/j.jksuci.2019.05.002

Boulares, M., Jemni, M., 2012. Mobile sign language translation system for deaf community, in: Proceedings of the International Cross-Disciplinary Conference on Web

Accessibility - W4A '12. Presented at the International Cross-Disciplinary Conference, ACM Press, Lyon, France, p. 1. https://doi.org/10.1145/2207016.2207049

Department of Computer Science & Engineering, Sri Jayachamarajendra College of Engineering, Mysuru, India, S, Shivashankara, S, Srinath, 2018. American Sign Language Recognition System: An Optimal

Approach. Int. J. Image Graph. Signal Process. 10, 18–30. https://doi.org/10.5815/ijigsp.2018.08.03

Escudeiro, P., Escudeiro, N., Reis, R., Barbosa, M., Bidarra, J., Gouveia, B., 2014. Automatic Sign Language Translator Model. Adv. Sci. Lett. 20, 531–533. https://doi.org/10.1166/asl.2014.5344

Garcia, B., Viesca, S.A., n.d. Real-time American Sign Language Recognition with Convolutional Neural Networks 8.

Gebre, B.G., Crasborn, O., Wittenburg, P., Drude, S., Heskes, T., 2014. Unsupervised Feature Learning for Visual Sign Language Identification, in Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Presented at the Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Baltimore, Maryland, pp. 370–376. https://doi.org/10.3115/v1/P14-2061

Gebre, B.G., Wittenburg, P., Heskes, T., 2013. Automatic sign language identification, in 2013 IEEE International Conference on Image Processing. Presented at the 2013 20th IEEE International Conference on Image Processing (ICIP), IEEE, Melbourne, Australia, pp.

2626–2630. https://doi.org/10.1109/ICIP.2013.6738541

Joslin, C., El-Sawah, A., Qing Chen, Georganas, N., 2005. Dynamic Gesture Recognition, in 2005 IEEE Instrumentation and Measurement Technology Conference Proceedings. Presented at the 2005 IEEE Instrumentation and Measurement Technology, IEEE, Ottawa, ON, Canada, pp. 1706–1711. https://doi.org/10.1109/IMTC.2005.1604461

Kishore, P.V.V., Rajesh Kumar, P., 2012. A Video-Based Indian Sign Language Recognition System (INSLR) Using Wavelet Transform and Fuzzy Logic. Int. J. Eng. Technol. 4, 537–542. https://doi.org/10.7763/IJET.2012.V4.427

[No title found], n.d. . Int. J. Adv. Res. Comput. Sci. Softw. Eng.

Parcheta, Z., Martinez Hinarejos, C.D., 2018. Sign Language Gesture Classification using Neural Networks, in: IberSPEECH 2018. Presented at the IberSPEECH 2018, ISCA, pp. 127–131. https://doi.org/10.21437/IberSPEECH.2018-27

Saldaña González, G., Cerezo Sánchez, J., Bustillo Díaz, M.M., Ata Pérez, A., 2018. Recognition and Classification of Sign Language for Spanish. Comput. Sist. 22. https://doi.org/10.13053/cys-22-1-2780

Yuqian Chen, Wenhui Zhang, 2016. Research and implementation of sign language recognition method based on Kinect, in: 2016 2nd IEEE International Conference on Computer and Communications (ICCC). Presented at the 2016 2nd IEEE International Conference on Computer and Communications (ICCC), IEEE, Chengdu, China, pp. 1947–1951. https://doi.org/10.1109/CompComm.2016.7925041

Punchimudiyanse, M., Meegama, R.G.N., 2017. Animation of Fingerspelled Words and Number Signs of the Sinhala Sign Language. ACM Trans. Asian LowResour. Lang. Inf. Process. 16, 1–26. https://doi.org/10.1145/3092743

Author Biographies

Kalhara Perera is a final year undergraduate of Kotelawala Defense university who is a passionate researcher of image processing and machine learning technologies. And, he is a Software Engineering intern at Epic Technology Group. His Final Year individual Project is also based on a Sinhala Sign Language translator.