Sign Language Translator for Deaf and Speech Impaired People using Convolutional Neural Network

VJ Samarasinghe[1], SCM De S Sirisuriya[2], N Wedasinghe[3] and IA Wijethunga[4]
Faculty of Computing, General Sir John Kotelawala Defence University, Ratmalana, Sri Lanka
[1] *vinodya.jan@gmail.com*

Abstract— Sign Language is the main communication medium among deaf and speech impaired people. In order to express their thoughts and emotions, hand gestures are used. In Sri Lanka, the Sri Lankan Sign Language is considered as the native Sign Language. But unfortunately, in the Sri Lankan community, the deaf and speech impaired people are often ignored by society due to the language barrier. As a solution to the problem, this paper proposes a system to assist the deaf and speech impaired people in capturing their sign-based message via the camera and then convert it into Sinhala text and furthermore into audio form. So, the main aim of this research is to eliminate the communication gap and to improve interaction between them and the common people. Convolutional Neural Networks (CNN) has been used as the technology of this research. The proposed CNN model which consists of one convolution layer, one max pooling layer and two dense layers along with Relu and Softmax activation functions has the ability to automatically extract the features of the input static gesture and recognize it (out of 24 classes) and give it as the output in text form. And then a text to speech engine will eventually generate the audio output in the Sinhala language. The model was trained for more than 20 times and obtained an accuracy of 98.61%. The proposed model has been implemented through python and libraries like OpenCV, Keras, pickle, etc have been used in advance.
Keywords— **Convolutional Neural Networks, Static gestures, Gesture recognition, HSV**

I. INTRODUCTION

Machine learning (ML) which is a subfield comes under Artificial Intelligence (AI) since 1959, was first introduced by Arthur Samuel("Machine learning," 2019). For the last few decades, it has shown a tremendous contribution to most of the generic applications that are popular in the world. Neural Networks are a subdivision under ML and it consists of both Artificial Neural Networks (ANNs) and Convolutional Neural Networks (CNNs) which is also called Deep Learning (DL)(Alom et al., n.d.). Fig 1 represents the taxonomy of AI. DL is widely used currently in many AI applications including image recognition, speech recognition, medical diagnosis, classification and prediction("Convolutional neural network," 2019). This paper also discusses a system to recognize gestures which are a form of sign language performed by deaf and speech impaired people using DL.

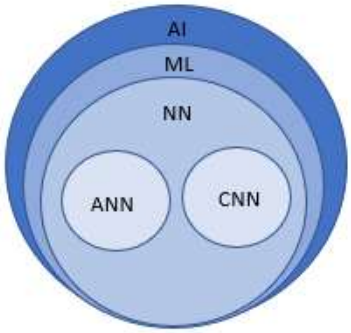
Sign Language is the main communication medium among deaf and speech impaired people. In order to express their thoughts and emotions, hand gestures are used which collectively form Sign Language. Sign Language varies from country to country, every nation (hearing and speaking impaired nations) of each country are adapted to their native Sign Language. Likewise, in Sri Lanka, the SSL (Sri Lankan Sign Language) is considered as the native Sign Language and unlike in the common Sinhala alphabet, its alphabet consists of 56 characters and collectively the whole language contains more than 350 signs(Stone, n.d.).

In the last two decades, a considerable number of gesture recognition systems were introduced under numerous technologies. Use of skin filtering technique, use of gloves, use of Microsoft Kinect sensor for tracking hand, Fingertip Search method, Hu Moments method, shape descriptors, Grid- Based Feature Extraction technique in feature extraction and Artificial Neural Network (ANN), Hausdorff distance, Eigenvalue weighted Euclidean distance technique, Hidden Markov Model (HMM)**,** Dynamic Time Warping (DTW), Extreme Learning Machine(ELM), k-Nearest Neighbors (k-NN) algorithm, Finite State Machine (FSM) in classification.

According to the Census of Population and Housing 2012, there were a total of 569,910 including both hearing and speech impaired people in Sri Lanka(Sri Lanka, 2015). But unfortunately, in the Sri Lankan community, the deaf and speech impaired people are often ignored by the common people due to the language barrier. As a result, they have been deprived of their right to live a normal life. Due to the language barrier, the deaf and speech impaired people find it hard to convey their messages correctly. Therefore, unlike common people, they fail to access their needs and wants easily. The facilities which support the disable people are very limited in Sri Lanka therefore, the service providers also face many difficulties in understanding their requirements and fulfilling them. As there's a language barrier between common people and deaf and speech-impaired people, they often feel isolated. Therefore, they are more likely to suffer from depression(Flexer, n.d.). Studies have revealed that they are twice liable to prone to psychological problems such as depression and anxiety than the common people. Consequently, this paper introduces a solution in order to break this unfairness and give them the same freedom to live like common people.

As a solution, the proposed system will act as the third party in mitigating the trouble of the language barrier. The

system will assist the deaf and speech impaired people in capturing their sign-based message via the mobile camera



and then convert it into Sinhala text and furthermore into audio form. Therefore, like common people, it will be very easy for them to communicate with any person at any place and get their work could be done very smoothly. As a result, the hardships they encounter during the engagement with common people will be curtailed.

Figure 1. Taxonomy of AI

The rest of the paper is designed as follows. Section 2 comparatively construe the overview of some existing gesture recognition systems. Section 3 gives the design of the proposed system. Section 4 discusses the results, issues and solutions for them. Finally, section 5 concludes the paper with further works.

II. 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), Hausdorff distance, Eigenvalue weighted Euclidean distance technique, Hidden Markov Model (HMM)**,** Dynamic Time Warping (DTW), Extreme Learning Machine(ELM), k-Nearest Neighbors (k-NN) algorithm, Finite State Machine (FSM) in classification.

For static gesture recognition, (Sánchez-nielsen et al., n.d.) have introduced a vision-based system for hand gesture recognition[. The system contains two major modules such as hand posture location and hand posture recognition. For the hand posture recognition part, the method of Hausdorff distance has been used and a success of 90% average rate has been achieved by implementing it for recognizing 26 hand postures. An OpenCV android based Sign Language translator application by (Triyono et al., 2018) has considered two approaches, Fingertip Search method and Hu Moments method. Fingertip Search method is more flexible compared to Hu Moments method and has shown a higher accuracy of 95%. Hu Moments method with the same restrictions has shown only an accuracy of 40. (Singha and Das, 2013)have proposed an Indian Sign Language (ISL) Recognition system which is comprised of 4 parts such as Filtering, Cropping, Extracting features and Classification. Classification is mainly done using Eigenvalue weighted Euclidean distance technique and has achieved success of 97%. Using CNN (Ahmed et al., 2019) have developed a model to recognize hand sign-based digits and to give the result in Bangla Language with an accuracy of 92%. The major drawback of all the above systems is the difficulty of implementation in complex backgrounds. A novel approach to the problems of removing complex background from gestures has been developed with the use of Extreme Learning Machine (ELM). A success of 84% is achieved for a simple background and 58% success for the complex background(Neiva and Zanchettin, 2016).

Dynamic gesture recognition is more difficult compared to static gesture recognition. (Pigou et al., 2015) have proposed a system using Microsoft Kinect, Convolution Neural Networks (CNNs) and GPU acceleration. The model is trained to recognize 20 Italian gestures with an accuracy of 91.7% on generalized users and surroundings. In this approach two steps are followed; the first step is the extraction of features using CNN from the frame sequences which will affect the computer to distinguish between the possible gestures. The second step is to classify gestures in order to figure out the different gestures. For the classification part, ANN is used and it is limited only to capture a single hand. With the use of a DataGlove (Rung-Huei Liang and Ming Ouhyoung, 1998)'s Real-time Continuous Gesture Recognition System could recognize 250 vocabularies in Taiwanese Sign Language (TSL). The Hidden Markov Model is applied in the system in order to continuously recognize vocabularies in real-time with an accuracy rate of 80.4%. Although the posture, orientation, motion models are easy to implement the position model is moderately complex to implement. Using recurrent neural network (Murakami and Taguchi, 1991), have developed a dynamic hand gesture recognition method for Japanese Sign Language of 42 symbols at an accuracy of 96%. (Amatya et al., 2018) have proposed a system to translate dynamic gestures in German Sign Language into text using Microsoft Kinect for Windows v2. In this paper, two approaches are considered. By using Dynamic Time Warping (DTW) algorithm, an accuracy of 65.45% is obtained but the computation time is increased with the increasing of the number of gestures in the dataset and by using Gesture Builder with DTW, only an accuracy of 20.42% is achieved but with constant time. The major drawback is the inability to capture start and

end points of a particular gesture precisely. (Mahanta et al., 2011)'s Dynamic Hand Gesture Recognition System has used MPEG-7ART based shape descriptors in extracting spatial information and particle filter in extracting trajectory features and finally Radial Basis Function neural network (RBF) to obtain success rate range of 80%-98%. Microsoft Kinect along with the DTW algorithm have shown good results(Putera et al., 2018). (Fernando and Wimalaratne, 2016) have proposed both DTW and nearest neighbor classification algorithm and have obtained an accuracy of 92.4%.

With the use of a range camera and the technique of chamfer distance matching for analysing hand shape together with Finite State Machine (FSM) for recognizing hand trajectory have obtained a high recognition rate(Li and Jarvis, 2009). The main drawback is the use of depth data which may result in erroneous hand positions. (Shenoy et al., 2018) have developed a system to recognize both hands pose and hand gestures from ISL using Grid- Based Feature Extraction technique and k-Nearest Neighbors algorithm for classification of hand poses and HMM for classification of gestures. An accuracy of 99.7% for hand poses and an accuracy of 97.23% for gestures have been acquired. Only limited to single handed gestures and the user needs to wear fully sleeved shirts.

III. DESIGN

A. Pre-processing

Data processing is an important phase in a machine learning project. If there are extraneous or unreliable or redundant or noisy data then it will not be easy to extract knowledge during the training phase.

1) Collection of Datasets: The dataset was created by using the hands of seven people. Created each image is a 50x50 HSV image. In order to make the model training more effective, three types of hand sizes including small, medium and large were included in the training set. For validation and test datasets, four different hand sizes taken from four persons were included as two different hand sizes for each dataset. The total of the collected image dataset was 57600 containing 24 classes (24 hand signs) and 2400 images for each hand sign.

2) Loading datasets: 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. In order to specify in terms of each hand sign, there were 1440 images in each training dataset, 480 images in each validate dataset and again 480 images in each test dataset.

B. Convolutional Neural Network Model

A CNN applies a set of filters to an image in order to extract the important features of that particular image. Therefore, the model can be used for better classification. CNN is composed of three components, convolution layer which applies an activation function (ReLU, sigmoid) and then gives an output passed through nonlinearities, pooling layer which performs down sampling also called reducing spatial dimension in order to reduce the processing time. The frequently used pooling is Max pooling. And finally, the dense layer which is also called Fully Connected layer performs the classification at the end using the features extracted by the previous layers. Therefore, in order to train and test, a deep learning model will pass each and every input image via these layers and at the end, the object will be classified as per the desired manner.

1) Proposed Model: In the proposed model, there is only one convolution layer and it has 32 kernels of size 5x5 to extract features from the input image. And for the subsampling process, this convolution layer is then followed by a 5x5 max pooling layer with a stride of 5. After pooling layer, 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 2.

2) Training and Validation Phase: The model was trained for 10 epochs with a batch size of 200. The model was converged well before reaching it to 10 epochs. Since the loss was needed for training and validation, the model was compiled using the adam optimizer and categorical_crossentropy loss function. The weights were saved with the model for future use.

3) Test Phase: After training the model, it was evaluated to receive the test accuracy based on the test dataset which is comprised of two different hand sizes of a male and a female. The accuracy obtained was in a better state with a less loss rate.

C. Recognized Sign into Speech

After developing the model, the next step was taken to convert the hand sign character into a speech in the Sinhala Language. Using the translator API; Google Translator("Cloud Translation documentation | Cloud Translation," n.d.), the recognized hand sign's text which was stored in the database as an English character was converted to a Sinhala character. Then the corresponding Sinhala character was converted into a Sinhala speech using Google Text to Speech API ("Cloud Text-to-Speech - Speech Synthesis | Cloud Text-to-Speech API | Google Cloud," n.d.).

The high-level architecture of the complete system is shown in Fig 3.

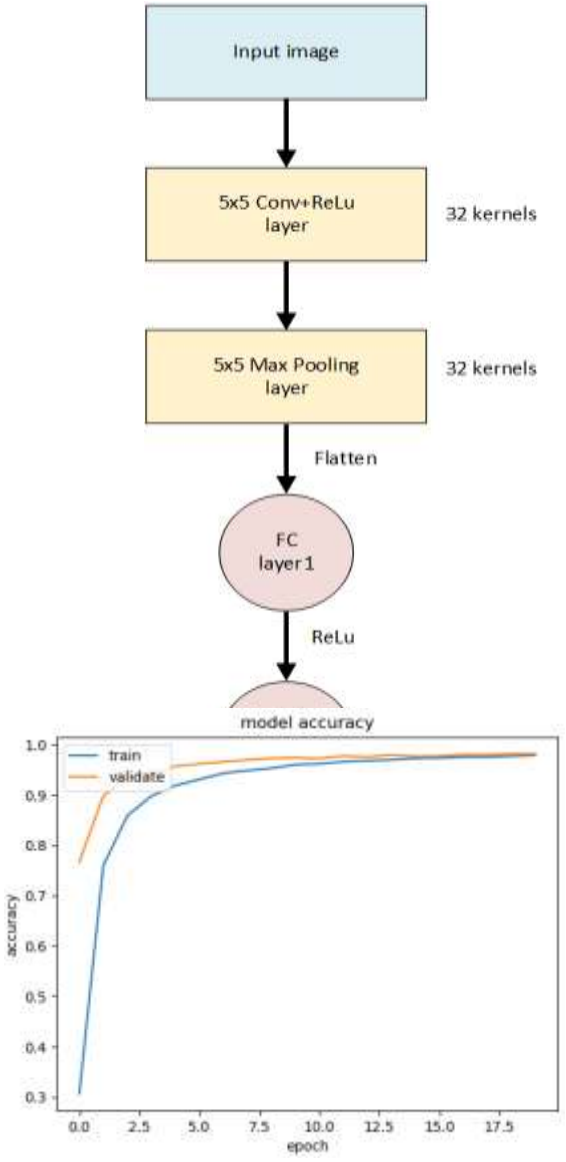


Figure 4. Training accuracy & Validation accuracy

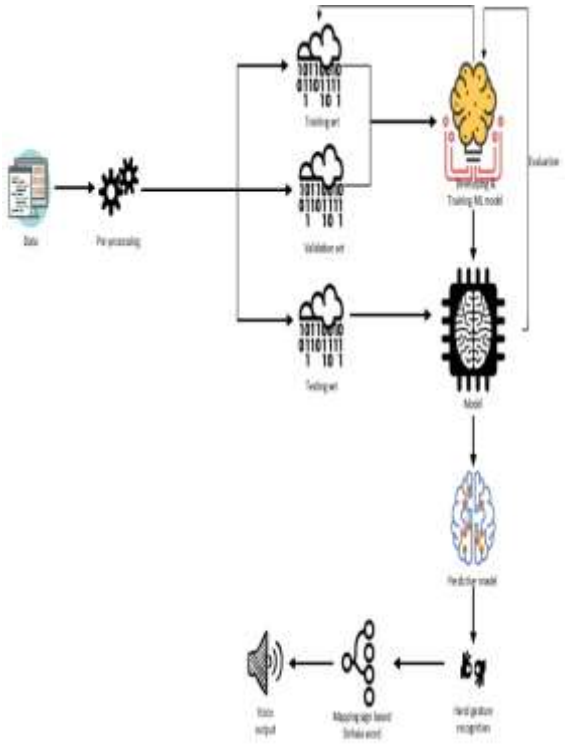


Figure 3. High level architecture of the proposed system

IV. RESULTS AND DISCUSSION

The model was trained for more than 20 times with different combinations of parameters to reach the best possible output. During the early training sessions, new layers were added with different combinations of parameters and the training and validation accuracies were compared. Validation accuracy was always around 0.4 even though the training accuracy was around 0.9. This was an overfitting condition and in order to solve this problem the layers were reduced and a dropout regularization was used. Finally, a test accuracy of 98.61% was given where the validation loss was 0.06 for only one convolution layer with 32 features and a dense layer with 64 features and a 0.2 dropout value. The training accuracy along with the validation accuracy is given in Fig 4. The training loss along with the validation loss is given in Fig 5.

The total picture of how the model performed correct/wrong classifications are graphed in the confusion matrix shown in Fig 6.

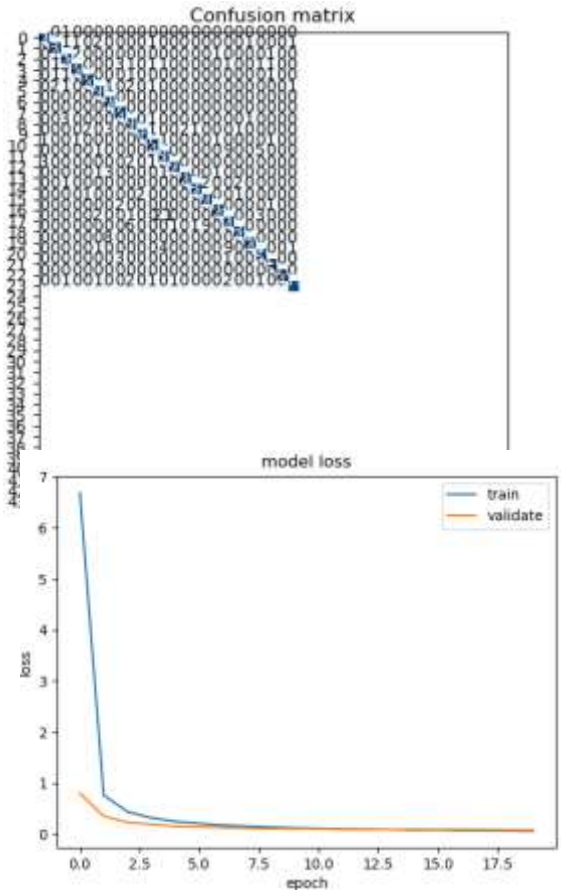


Figure 5. Training loss & Validation loss

V.CONCLUSION AND FURTHER WORKS

Our main aim is to develop a complete system for the deaf and speech impaired people in Sri Lanka and to make them freely behave as other common people in the society irrespective of the language barrier. So, this paper proposes a system to assist the deaf and speech impaired people in capturing their sign-based message via the camera and then convert it into Sinhala text and furthermore into audio form. On the way to solve that bigger problem, as the initial step, a system was built to recognize the static gestures of Sinhala alphabet with better accuracy and to give the output in audio form in Sinhala. The proposed CNN model has the ability to automatically extract the features of the input image and recognize its class (out of 24 classes) and give it as the output in text form. And then a text to speech engine will eventually generate the audio output in the Sinhala language. The proposed model has been implemented through python and libraries like OpenCV ("OpenCV," 2019), Keras ("Keras," 2019), pickle, etc. are used.

In the near future, the system will be deployed as a mobile application providing the user the easy access to the system. For further improvements, an additional option to form sentences based on the gestures recognized and to speak out the whole captured sentence in the Sinhala language will be introduced.

ACKNOWLEDGEMENT

I would like to extend my sincere thanks to all my supervisors and colleagues who supported me to make this project successful.

REFERENCES

Ahmed, S., Islam, M.R., Hassan, J., Ahmed, M.U., Ferdosi, B.J., Saha, S., Shopon, M., 2019. Hand Sign to Bangla Speech: A Deep Learning in Vision based system for Recognizing Hand Sign Digits and Generating Bangla Speech. ArXiv190105613 Cs.

Alom, Z., Taha, T.M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M.S., n.d. The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches 39.

Amatya, P., Kateryna Sergieieva, Meixner, G., 2018. Translation of Sign Language Into Text Using Kinect for Windows v2, in: ProceedingsThe Eleventh International Conference on Advances in Computer-Human Interactions. Presented at the The Eleventh International Conference on Advances in Computer-Human Interactions.

Cloud Text-to-Speech - Speech Synthesis | Cloud Text-to-Speech API | Google Cloud [WWW Document], n.d. URL https://cloud.google.com/text-to-speech/ (accessed 9.8.19).

Cloud Translation documentation | Cloud Translation [WWW Document], n.d. . Google Cloud. URL https://cloud.google.com/translate/docs/ (accessed 9.8.19).

Convolutional neural network, 2019. . Wikipedia.

Fernando, P., Wimalaratne, P., 2016. Sign Language Translation Approach to Sinhalese Language. GSTF J. Comput. JoC 5. https://doi.org/10.7603/s40601-016-0009-8

Flexer, S., n.d. Difficulties the Hearing Impaired Face Every Day [WWW Document]. URL https://www.disabilityexpertsfl.com/blog/difficulties-the-deaf-face-every-day (accessed 2.5.19).

Keras, 2019. . Wikipedia.

Li, Z., Jarvis, R., 2009. Real time Hand Gesture Recognition using a Range Camera 7.

Machine learning, 2019. . Wikipedia.

Mahanta, C., Yadav, T.S., Medhi, H., 2011. DYNAMIC HAND GESTURE RECOGNITION SYSTEM USING NEURAL NETWORK:, in: Proceedings of the 1st International Conference on Pervasive and Embedded Computing and Communication Systems. Presented at the International Conference on Pervasive and Embedded Computing and Communication Systems, SciTePress - Science and and Technology Publications, Vilamoura, Algarve, Portugal, pp. 253–256. https://doi.org/10.5220/0003299102530256

Murakami, K., Taguchi, H., 1991. Gesture Recognition using Recurrent Neural Networks, in: Proceedings: Conference on Human Factors in Computing Systems. pp. 237–242.

Neiva, D.H., Zanchettin, C., 2016. A Dynamic Gesture Recognition System to Translate between Sign Languages in Complex Backgrounds, in: 2016 5th Brazilian Conference on Intelligent Systems (BRACIS). Presented at the 2016 5th Brazilian Conference on Intelligent Systems (BRACIS), pp. 421–426. https://doi.org/10.1109/BRACIS.2016.082

OpenCV, 2019. . Wikipedia.

Pigou, L., Dieleman, S., Kindermans, P.-J., Schrauwen, B., 2015. Sign Language Recognition Using Convolutional Neural Networks, in: Agapito, L., Bronstein, M.M., Rother, C. (Eds.), Computer Vision - ECCV 2014 Workshops. Springer International Publishing, Cham, pp. 572–578. https://doi.org/10.1007/978-3-319-16178-5_40

Putera, Z.P., Anasanti, M.D., Priambodo, B., 2018. Designing Translation Tool: Between Sign Language to Spoken Text on Kinect TIME Series Data Using Dynamic TIME Warping. Sinergi J. Tek. Mercu Buana 22, 91–100.

Rung-Huei Liang, Ming Ouhyoung, 1998. A real-time continuous gesture recognition system for sign language, in: Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition. Presented at the Third IEEE International Conference on Automatic Face and Gesture Recognition, IEEE Comput. Soc, Nara, Japan, pp. 558–567. https://doi.org/10.1109/AFGR.1998.671007

Sánchez-nielsen, E., Antón-canalís, L., Hernández-tejera, M., n.d. Hand Gesture Recognition for Human-Machine Interaction.

Shenoy, K., Dastane, T., Rao, V., Vyavaharkar, D., 2018. Real-time Indian Sign Language (ISL) Recognition, in: 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT). Presented at the 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT), IEEE, Bangalore, pp. 1–9. https://doi.org/10.1109/ICCCNT.2018.8493808

Singha, J., Das, K., 2013. Indian Sign Language Recognition Using Eigen Value Weighted Euclidean Distance Based Classification Technique. Int. J. Adv. Comput. Sci. Appl. 4. https://doi.org/10.14569/IJACSA.2013.040228

Sri Lanka (Ed.), 2015. Jana hā nivāsa saṅgaṇanaya, 2012. Janalēkhana hā Saṅkhyālēkhana Depārtamentuva, Pratipatti Sampādana, Ārthika Kaṭayatu, ḷamā, Taruṇa hā Samskṛtika Kaṭayutu Amātyāṃśaya, Battaramulla, [Sri Lanka].

Stone, A., n.d. An Introduction to Sri Lankan Sign Language. Rohana Special School, Matara.

Triyono, L., Pratisto, E.H., Bawono, S.A.T., Purnomo, F.A., Yudhanto, Y., Raharjo, B., 2018. Sign Language Translator Application Using OpenCV. IOP Conf. Ser. Mater. Sci. Eng. 333, 012109. https://doi.org/10.1088/1757-899X/333/1/012109