**Literature Review**

Sinhala Sign Language is a visual language used by deaf people in Sri Lanka. According to Sri Lankan Federation of the Deaf, there are over three hundred thousand plus (300,000+) Deaf people in Sri Lanka. Moreover, the World Health Organization has revealed that approximately 9% of the population in Sri Lanka have speaking and hearing impairments. So Sign language is an extremely important communication tool for deaf, dumb and hard-of-hearing people. But not every single person in Sri Lanka can recognize the sign language. It simply takes time to learn sign language when compared to a natural language. S. P. More et al. [5] state that the hearing people never try to learn the sign language. And because of that deaf people cannot interact with the normal people without a sign language interpreter and eventually the deaf people get isolated in the society. Also having a personal translator for the communication can be costly. Therefore it is important to make a system that automatically recognize sign language.

**Sign Language**

As mentioned by P. Fernando el at. [3] currently there are nearly hundred sign languages can be identified around the world. American Sign Language (ASL), British Sign Language (BSL) Mexican Sign Language (LSM), French Sign Language (LSF), Italian Sign Language (LIS) and Spanish Sign Language (LSE) are just among few of them.

According to S. P. More et al. [5], Sign language is a method which uses manual communication and body language instead of acoustically conveyed sound patterns, to convey meaning.

According to V. Padmanabhan et al. [2], the dumb have their own manual-visual language referred to as sign language. Sign Language is also a non-verbal form of intercourse that's found among deaf communities

According to (11), one important means of communication method for the hearing impaired community is the use of sign language

**Sinhala Sign Language**

According to P. Fernando et al. [6] Sri Lankan Sign Language was fully built on the foundation of British Sign Language (BSL). However Sri Lankan Sign Language made lots of variations to the British Sign Language and currently consists of more than 2000 sign based words [8].

**Alphabet based sign language**

The systems that are translating Sinhala sign languages translate words and phrases into text only. There are no systems that are translating Sinhala sign language for Sinhala letters into Sinhala letter texts. Current systems are only capable of identifying few number of words and sentences among 2000+ Sinhala sign language. Alphabet of Sinhala language only consist of 60 letters. Since letters are the building blocks of any language, converting signs based on letters to text can increase the horizons to identify and translate almost any word.

M. R. Jagadish et al [6] stated that the system of communication called "finger spelling" involves spelling out words in an alphabetical language by using the letters of the manual alphabet -with hand shapes and positions corresponding to each letter of the written alphabet Conversations can be entirely finger spelled. But among deaf individuals, finger spelling is more often used in conjunction with sign language for proper names and terms for which there are no signs.

**Real time translators of the gesture recognition system**

Every researches are mainly focusing on building a real time translation of Sinhala Sign Language.

N. Kulaveerasingam et al. [7] mentioned that their research is to create a real time translator for SSL and the real time language translator leads to build effective communication between deaf people and general people.

According to P. Fernando et al. [3] their research focuses on an approach for a real-time translation from Sri Lankan sign language to Sinhalese language which will bridge the communication gap between deaf and ordinary communities.

**Categories of Gesture recognition system**

(9) say that there are basically two types of approaches for hand gesture recognition, vision-based approaches and data glove approaches. Also According to S. P. More el at. [5] the gesture recognition is mainly categorized into vision-based approach and Haptic-based approach. The vision based approach captures the movement of the signer’s hand using cameras. The haptic-based approach deals with instrumented gloves affixed with measurement devices which track hand movements. Furthermore they said that in vision-based approach, the cameras record the ever-changing image and position of the hand because the user signs and also the pictures are then processed to retrieve the hand form, position and orientation. Also as mentioned by them, the haptic-based approach methods require more power supply and it is very costly.

According to M. R. Jagadish et al [6], they specially deals with the haptic-based approach because owing to large data and complex computation involved in vision-based approach.

As N. Kulaveerasingam et al. [7] states, the existing systems are not affordable to use on behalf of impaired people. The economy is not flexible to implement such a system using the new technologies, especially in an Asian country like Sri Lanka. So their research is done based on vision-based approach.

**Glove methods (device based/ Haptic-based ) approach to convert hand gestures into required output**

The application model of M. R. Jagadish et al. [6] is a device that translates sign language of deaf-mute person to synthesized text and voice for communication. Their methodology provides a map for developing a digital wireless data glove which is fitted with Flex sensors. Flex sensors are analog resistors that function as analog voltage dividers which are sense the gestures of a person in the form of bend of fingers and tilt of the hand fist. They used flex because it is ideal for any application or device that requires the measurement of a repetitive bending, striking deflection, acceleration or range of motion. Additionally, Flex is proven for high speed impact measurements. In the proposed system they are implementing the FLEX sensor. The input data glove detects the hand gesture done by the deaf-dumb person wearing it. Next provides the analog input to the microcontroller for further interpretation according to the database. The final output is observed on the LCD display and the speaker.

V. Padmanabhan et al. [2] proposed a new technique called artificial speaking mouth for dumb people. In order to overcome the complexity of understanding the way of conveying the messages, the artificial mouth is introduced for the dumb peoples. According to dumb people, for every motion they have a meaning. That message is kept in a database. Likewise all templates are kept in the database. In the real time, the template database is fed into a microcontroller. This system is based on the motion sensor. The motion sensor is fixed in their hand. For every action the motion sensors get accelerated and give the signal to the microcontroller. The microcontroller matches the motion with the database and produces the speech signal. The system also includes a text to speech conversion (TTS) block that interprets the matched gestures. The output of the system is using the speaker.

The research paper uses database to keep meaning of the motion and fed into a microcontroller. Also fixed a motion sensor in hand and for relevant action give signals to microcontroller. Microcontroller compares those with database. They use a text to speech conversion and speaker to output the action.

**Vision-based approach to convert hand gestures into required output**

Because there are drawbacks in the haptic-based approach, researches of the gesture recognition of sign language has done by using vision-based approach.

According to S. P. More el at. [5], to overcome the drawbacks of the Haptic-based methods, they use the static hand gesture recognition system using digital image processing which is a vision based approach

H. C. M. Herath [4] presents a low cost approach to develop an image processing based Sinhala sign language recognition application for real time applications.

As mentioned in N. Kulaveerasingam et al.[7] while most of the projects use high technical features with high cost, “Nihanda Ridma” going to be an economical and affordable project, since it uses the vision based approach and the low cost technical equipment.

(9) says that The reason for choosing a system based on vision relates to the fact that it provides a simpler and more intuitive way of communication between a human and a computer.

**Methodologies used for image preprocessing**

The model of S. P. More et al. [5] will be used to recognized hand gesture captured from webcam. Before building the model, images will be captured for each hand gesture, which are the fist, index, palm, and little fingers, for different people, scales, and rotations and under different illuminations. The background has no texture or objects (white wall). Then it guarantee that all the key points extracted from training images using the SIFT algorithm will represent the hand gesture only.

In the prototype developed for the project H. C. M. Herath [4], a green background is used to capture the image for the simplicity of the implementation. And also used Matlab simulation package and a portable camera (Intex Model No IT-309WC, 16MP). First, the RGB image captured from the web camera is separated into the three matrices, red (R), green (G) and blue (B). Next G matrix is subtracted from the R matrix. This is done because it was experimentally found that red is the most dominant color of the skin as shown in the figure 1 and the background used is in green color as well. Shadows are removed in this process as in figure 2. In figure 2 (b) shadow effect remove by subtracting G matrix from R matrix and it is convert to binary image as in figure 2 (c Then again the resulted image is converted to binary image by defining a threshold. This is generated to facilitate faster mapping. The resulted binary image accuracy is depended on lighting condition at which the image is captured. If the lighting intensity is sufficient to capture the image with its natural colors or closer to natural colors then the binary image is noise free as illustrate in figure 4. This is how the image preprocessing done in H. C. M. Herath [4].



Figure 1:R, G, B value of a point in the hand



Figure 2:Shadow effect removal

**Methodologies used in image processing**

S. P. More et al [5] proposed that the feature extraction was done efficiently using SIFT computer vision algorithm. They design the SIFT algorithm for hand gesture feature vector. According to them SIFT features are distinctive and invariant features extracted from images that allow for efficient matching with various other viewpoints of the extracted features that may exist in the same or different gestures. The features are invariant to image translation, scaling, rotation and partially invariant to illumination changes. The SIFT features described in their implementation have been computed at the edges which are invariant to scaling, rotation, addition of noise. The computation of SIFT image features is in four basic steps. They are scale-space local extreme detection, key point localization, orientation assignment, key point descriptor. The advantage of using the algorithm is high processing speed which can produce results in real time.

After preprocessing the images H. C. M. Herath [4] proposed following method to process the images. After preprocessing images, the boundaries of the hand are identified by drawing smallest possible rectangle around the hand and the image is cropped to extract the region to interest. Then the cropped image is equally divided in to four parts as shown in Figure 5(b). Next centroid of each segment is calculated as shown in Figure 5 (c). The (Height/y) and (Width/x) ratios are calculated for each segment and then they are compared against precalculated values that are in the database. Errors of the ratios are also calculated by subtracting the ratio calculated for the real time image from the ratio calculated for the real time image. Finally, the image with minimum error is selected as the matched image.



Figure 3:Segmented hand and ratio calculation

As A.-A. Bhuiyan [1] proposed the ASL The ASL recognition system comprises with two segments. The feature extraction and the identification of signs. The feature extraction process is initiated with an image processing procedure, which involves an algorithm to detect and segment various desired segments of the sign.

First each color image is resized. Images of signs were resized to 30×24, by default uses nearest neighbor interpolation to determine the values of pixels in the output image. This research employs a lowpass filter before interpolation to reduce aliasing. Next step is that the hand images may be of poor contrast because of the limitations of the illumination conditions. the proposed fuzzy histogram equalization algorithm is used to reimburse for the illumination conditions and recover the contrast of the image, as shown in Figure 4. The ASL images are sometimes corrupted by numerous sources of noise. Therefore, Prewitt filtering is used to suppress the noise in the next step.



Figure 4:Fuzzy Histogram equalization

In the next step skin color segmentation is organized with visual information of the hand skin colors extracted from different images. This research uses HSV color space for skin color segmentation. In the HSV color space a color is described by three attributes: hue - the visual attribute of color sensation linked with the dominant colors, saturation -implies the relative purity of the color content and value- measures the brightness of a color. Since the human skin colors are clustered in color space and differ from person to person and of races, so in order to detect the hand parts in an image, the skin pixels are thresholded empirically. In this research, the hue values are chosen as h= [0, 40]. The detection of hand region boundaries by such an HSV segmentation process is illustrated in Figure 5. The exact location of the hand is then determined from the image with largest connected region of skin-colored pixels.



1. Original Color Image
2. Color Segmentation
3. Connected Component Analysis

Figure 5:Skin color segmentation

Finally the classification phase involves neural network training for the recognition of binary image patterns of the hand. In the classification stage, an Adaptive Resonance Theory (ART) neural network is employed. The ART contains 30×24 neurons in the input layer, 604 (70% of input) neurons in the hidden layer, and 26 neurons in the output layer.

**Methodologies used when using a Kinect camera to capture the gesture**

System of N. Kulaveerasingam et al. [7] had used Microsoft Kinect with AForge.NET Framework and the System has a multimedia database to store data using Microsoft SQL Server 2008. SQL server has the ability to store images, diagrams, graphical animations, sound and moving pictures as Binary Large Objects (BLOBs). Multimedia data have sorted in a suitable DBMS in a standardized and integrated manner.

The system of N. Kulaveerasingam et al. [7] is capable of capturing gestures one by one. It does not able to capture continuous gestures. Figure 6 enable user to record save gestures of deaf person one by one. User can compare the input gesture with default black screen by clicking the Compare button. Compare button navigates to Figure 6. Figure. 7 Show the comparison interface which compare hand gesture of deaf person with default black screen. Comparison done on the pixels of both images. The relevant word is displays if the comparison returns true. Error message is displays when the pixels are mismatched. Figure 8 Show the comparison interface which compare hand gesture of deaf person with default black screen. Comparison done by comparing pixels of both images. The relevant word is displays if the comparison returns true.



Figure 6:Gesture recognition Interface



Figure 8:Compare images



Figure 8:Comparison interface

Main objective of the proposed research of P. Fernando et al [3] is to develop software-based prototype, which can translate Sri Lankan Sign language into Sinhalese language. Microsoft Kinect SDK version 1.8, Microsoft Visual Studio 2010 IDE used to build the software using visual C# programming language. Furthermore, Windows Presentation Foundation (WPF) used since it was highly supported for representing visual information.

Each frame of the video that have when capturing the gesture will be retrieved by using kinect XBOX 360 device in order to track hand movements. As the initial point of implementation skeleton data stream and color data stream of Kinect sensor will be used. Once isolated frames are retrieved and analysed, information will be redirected to the gesture preprocessing module.

Main functionality of this module is to extract the feature points, which needs to be used for gesture detection. In gesture preprocessing by removing unnecessary details such as background information and unwanted skeleton points and using the useful skeleton points generate the feature frame. System implements a two-step normalization process done for extracted skeleton joint coordinates, before further processing. As the first step of normalization, a new center point will be calculated based on the shoulder coordinates of each user. Once the center point is repositioned, all observed coordinates will be aligned according to the new center point to make sure that the final output does not depend on the position of user while performing the gesture. As the second step of process, normalization for each coordinates will be performed based on length of two shoulder points, under the assumption of skeleton points which are symmetrical to ensure that the produced output does not depend on the physical size of the user. After the normalization process, candidate gesture identifier frame is generated. This process will be repeated until system receives minimum number of data frames which needs to detect the gesture.

They use the Training mode and store sufficient number of gesture identifier details within the gesture dictionary database. translation mode can be used and gesture identifier data will be sent to the dictionary database in order to perform a comparison and recognize the gesture. Each gesture will be performed under a window of 32 frames. From each frame, 3D coordinates of each skeleton point will be extracted and normalized. Each normalized coordinate data will be stored in the dictionary file, grouped according to sign word. Individual data files are generated for each gesture and combination of all data files form a single gesture dictionary. During the initial stage of development, five training samples have been used per gesture and dictionary of fifteen sign words has been used in the system. Therefore, final gesture dictionary consists of seventy five trained samples altogether. Once the dictionary is generated, system uses the classification module for comparing user performed signs and the training sample. Data which has the highest matching sequence will be selected as the gesture.

proposed research designed and implemented two step gesture Identification algorithm, where step 01 is based on Dynamic Time Warping algorithm and step 02 is based on Nearest Neighbor classification. Since gesture matching process follows the comparison between two sequences (Real time coordinate data and pre-trained sample data), Dynamic Time Warping algorithm (DTW) will be used with enhancements. In addition, nearest neighbor classifier will be used to choose the best matching gesture name based on the DTW classification results.

**References**

[1] A.-A. Bhuiyan, “Recognition of ASL for Human-robot Interaction,” p. 6, 2017.

[2] V. Padmanabhan and M. Sornalatha, “Hand gesture recognition and voice conversion system for dumb people,” vol. 5, no. 5, p. 5, 2014.

[3] P. Fernando and P. Wimalaratne, “Sign Language Translation Approach to Sinhalese Language,” *GSTF Journal on Computing (JoC)*, vol. 5, no. 1, 2016.

[4] H. C. M. Herath, “IMAGE BASED SIGN LANGUAGE RECOGNITION SYSTEM FOR SINHALA SIGN LANGUAGE,” p. 5, 2013.

[5] S. P. More and A. Sattar, “HAND GESTURE RECOGNITION SYSTEM FOR DUMB PEOPLE,” *International Journal Of Engineering*, vol. 3, no. 2, p. 4.

[6] M. R. Jagadish, R. Gayathri, R. Mohanapriya, R. Kalaivani, and S. Keerthana, “Hand Gesture Recognition System for Deaf and Dumb Persons,” vol. 2, no. 1, p. 8, 2018.

[7] N. Kulaveerasingam, S. Wellage, H. M. P. Samarawickrama, W. M. C. Perera, and J. Yasas, “‘The Rhythm of Silence’ - Gesture Based Intercommunication Platform for Hearing-impaired People (Nihanda Ridma),” Dec. 2014.

**(8) Sign Language Recognition using Convolutional Neural Networks**

**(9)** American Sign Language Character Recognition Using Convolution Neural Network

(10) **Real-time American Sign Language Recognition with Convolutional Neural Networks**

**(11) Real-time sign language recognition based on neural network architecture**

**Use CNN for image classification**

The system that (8) has implemented use the Microsoft Kinect, convolutional neural networks (CNNs) and GPU acceleration. The system of them able to recognize 20 Italian gestures cross-validation accuracy of 91.7%.

According to (9), their main focus of the work is to create a vision based system to identify Finger spelled letters of American Sign Language. They use CNN to implement their system.

(10) produced a robust model that consistently classifies letters a-e correctly with first-time users and another that correctly classifies letters a-k in a majority of cases. They use Convolutional Neural Networks in real time to translate a video of a user’s ASL signs into text.

(11) Their architecture is being proposed using the neural networks identification and tracking to translate the sign language to a voice/text format. They use CNN model to recognize letters in American Sign Language. First they will get a video sequence of the signer as the input from camera.

**Dataset**

(8) used the data set from the ChaLearn Looking at People 2014 [5] (CLAP14) challenge. They used 6600 gestures in the development set of CLAP14 for their experiments, 4600 for the training set and 2000 for the validation set. The test set of CLAP14 is also considered as the test set for this work and consists of 3543 samples.

(9) used an image dataset consists of ASL gestures from [1]. The dataset consists of 2524 depth images with 70 images per category. Each category represented a different character of ASL. This dataset was then augmented to create a dataset of 14781 images. Out of this dataset, 75% images were used for training and remaining 25% images were used for testing.

(10) Uses color images. They are close-ups of hands that span the majority of the image surface. They utilize ILSVRC2012 dataset, the Surrey University and Massey University ASL datasets in order to apply transfer learning to their task. Since there was little to no variation between the images for the same class of each signer, they separated the datasets into training and validation by volunteer. Four of the five volunteers from each dataset were used to train, and the remaining volunteer from each was used to validate.

**Preprocessing**

According to (8) first they have crop the highest hand and the upper body using the given joint information. The preprocessing results in four video samples (hand and body with depth and gray-scale) of resolution 64x64x32 (32 frames of size 64x64). Furthermore, the noise in the depth maps is reduced with thresholding, background removal using the user index, and median filltering.

According to (9), they have read and resize each of the image to the similar size of 224x224 pixel. As for them Only when all of the images in the dataset are of the same size can the images be fed into a neural network for training. The mean value of RGB over all pixels was subtracted from each pixel value. The mean is subracted because the model involves multiplying weights and adding biases to the initial inputs to cause activations then backpropagated [10] with the gradients to train the model. It is important that each feature has a similar range, in order to prevent the gradients from getting out of control.

(10) their Both datasets contain images with unequal heights and weights. Hence, they resize them to 256x256 and take random crops of 224x224 to match the expected input of the GoogLeNet. Also they have zero-center the data by subtracting the mean image from ILSRVC 2012. Since the possible values in the image tensors only span 0-255.

(11) states that, in order to satisfy the memory requirements and the environmental scene conditions, preprocessing of the raw video content is highly important [14]. They use a moving average or median filter as filteration preprocess. Next they did the Background subtraction. Because this is Neural network. They did feature extraction. Introduction of Point of Interest (POI) where The state of the hand gestures are given by the attributes called Point of Interest (POI) of the hands. The feature vector consists of 55 features.

**Augmentation**

According to (9) they augmented to produce several images from each image, thus increasing the size of the dataset and also tackling the problem of overfitting. Maximum ranges or degrees for shear, zoom, horizontal and vertical shifting were specified in the model

(10) They make horizontal flips of the images since signs can be performed with either the left of the right hand, and the datasets have examples of both cases. By this method they have augment their dataset

**Proposed architecture**

According to (8), For the pooling method, they use max-pooling in their ayatem. In max pooling only the maximum value in a local neighborhood of the feature map remains. They Use 2D convolutions, because it resulted in a better validation accuracy than 3D convolutions. The architecture of their model consists of two CNNs with Each CNN is three layers deep. Finally it have a classical ANN with one hidden layer. Also, local contrast normalization (LCN) as in [10] is applied in the first two layers and all artificial neurons are rectified linear units (ReLUs [14], [6])

According to (9), their Model that Used for training the dataset was inspired from VGG16 model. So VGG 16 model which is a deep convolutional neural network model used as the architecture of their CNN model. The max pooling layer used as the pooling layer. ReLU used as the activation function.

(10) They utilize a pre-trained GoogLeNet architecture trained on the ILSVRC2012 dataset, as well as the Surrey University and Massey University ASL datasets

(11) They used a new scheme called combinational neural networks (CNN). A three layer network called back propagation is used to build the CNN.

**Generalization and Training**

According to (8) dropout and data augmentation are used as main approaches to reduce overfitting during the training of the application.

**Result**

The system of (8) observed a validation accuracy of 91.70%. The accuracy on the test set is 95.68% and they observe a 4.13% false positive rate, caused by the noise movements.

The accuracy of the model of the system of (9) obtained using Convolution Neural Network was 96%.

(10) They evaluate two metrics in order to compare their results with those of other papers. The most popular criterion in the literature is accuracy in the validation set which is the percentage of correctly classified examples. One other popular metric is top-5 accuracy, which is the percentage of classifications where the correct label appears in the 5 classes with the highest scores. Additionally, we use a confusion matrix, which is a specific table layout that allows visualization of the performance of the classification model by class.

(11) This sign language recognition approach requires a computer with at least 1GHz processor and at least 256 MB of free RAM. The training set consists of all alphabets A to Z (26 patterns).