Sri Lankan Sign Language to Sinhala Text using Convolutional Neural Network Combined with Scale Invariant Feature Transform (SIFT)

L.L.D.K.Perera
Department of Industrial Management
University of Kelaniya
Dalugama, Sri Lanka
dileshakithmini@gmail.com

S.G.V.S.Jayalal
Department of Industrial Management
University of Kelaniya
Dalugama, Sri Lanka
shantha@kln.ac.lk

***Abstract*— Sri Lankan sign language (SSL) is a visual-gestural language used by the deaf community for communication. Hearing-impaired people cannot effectively communicate with a normal person due to the difficulty in understanding sign language. SSL to Sinhala text interpreting technology using gesture recognition helps to fill up this communication gap since Sinhala is the majority language used in Sri Lanka. Hand gesture recognition can be achieved by using vision-based or sensor-based approaches. Vision-based approaches are comparatively simple and less costly but sensor-based approaches are complex. Scale, rotation, occlusion affects the accuracy of gesture recognition, and keypoints act as better features to handle them. The research focuses on a combined approach of convolutional neural network (CNN) and Scale Invariant Feature Transform (SIFT) to develop a camera-based low-cost solution to interpret static gestures of SSL into Sinhala text. The SSL to Sinhala text translation model reached an accuracy of 86.5% when a dataset of images of 20 static SSL gestures was used. The classifier showed robustness to scale variations when the distance to the camera was varied and uniform color backgrounds were used. Further improvements can be done for the recognition of dynamic gestures and facial expressions of SSL.**

Keywords*— *Sign language, Keypoints, CNN, SIFT

I. Introduction

A hearing-impaired person is a person who is physically unable to speak or cannot communicate with another person using words and speech. Sign Languages have been introduced and are used nowadays as an effective way to help the hearing-impaired community to communicate with others.

Sri Lankan sign language (SSL) is the official sign language (SL) used among the hearing impaired community of Sri Lanka. Even though SL is understandable for a person who is familiar with it, the possibility of understanding the SL by a person who does not know the SL is very less. Hearing-impaired people face many problems when interacting with other people for their day-to-day activities since normal people cannot understand what they say unless the help of an interpreter from a translator service is used.

There are only a few qualified SSL interpreters in Sri Lanka who are not sufficient to facilitate all the translation needs of sign language to Sinhala or vice versa [1]. Available technologies like video calling may facilitate communication between deaf people but do not provide any translation services to support communication between deaf and normal people. Therefore these issues have created the need for SSL to Sinhala text interpreting technology which helps to fill up the communication gap between the normal and hearing-impaired community of Sri Lanka and allow the majority of Sri Lankans to understand the sign language.

This research study focuses on the development of a computer vision-based translation model to translate SSL into Sinhala text. Improvement on model accuracy with hand scale and variation of distance to camera are considered in the study.

II. Background

A. Sri Lankan deaf community

Sri Lankan deaf-dumb/hearing impaired community represents people who belong to the disability category of having problems in hearing and talking like a normal person. According to statistics by United Nations Economic and Social Commission for Asia and the Pacific in 2019, disability prevalence in Sri Lanka is stated as 8.7% [2] and out of the total disabled population, people with hearing impairments are reported as 389,077. This is approximately 24% of the total disabled population [3].

B. Sign language

Sign Language (SL) is a visual-gestural language that uses gestures of hands, upper body parts and facial expressions to convey meaningful messages to the listener. There is no universal sign language [4] and there are over 135 different sign languages around the world, including American Sign Language (ASL), British Sign Language (BSL), and Australian Sign Language (Auslan) [5]. Different sign languages are used in different countries or regions. Some countries adopt features of ASL in their sign languages [4]. Most countries that share a spoken language do not share the same sign language [5].

C. Sri Lankan sign language

SSL is used by the deaf-dumb community of Sri Lanka. The medium of communication for the provision of educational, social, cultural services as well as medical, health, and legal facilities to them is SSL [6]. Sri Lanka recognized SSL as an official language in September 2017 by the Conversational Sign Language Bill, becoming one of the 39 countries in the world at that time to give official status to sign language [7]. SSL dictionary is a collection of standard sign language symbols used in Sri Lanka by the deaf community and consists of about 350 different signs [8].

D. Sinhala language

Sinhala is a language that belongs to the Indo-European language family and is the majority language of Sri Lanka. Sinhala is spoken by about 16 million people as the first language and about 2 million people speak Sinhala as a second language [9]. The complete script consists of 60 letters, 18 for vowels, and 42 for consonants. The general sentence pattern in Sinhala follows the subject-object-verb format. There are distinct differences between literary and spoken Sinhala. Literary Sinhala is used for all forms of formal writing and spoken Sinhala which is distinguished between a formal and colloquial variety is used for informal day-to-day communication.

III. RELATED WORK

SL gesture recognition follows stages of data acquisition, data preprocessing, segmentation, feature extraction, and classification of hand gestures. Gesture recognition can be achieved by using either vision-based approaches (camera images/videos) or sensor-based approaches [10]. Compared to wearable device-based gesture recognition, vision-based gesture recognition systems enables low-cost gesture recognition [11]. Vision-based approaches have the challenges of handling the effects of backgrounds, scaling, rotation effects, and light intensity variations. Sensor-based approaches are better in tracking the motions but may have disadvantages considering usability and cost [12].

Several forms of color spaces are used for segmentation in SL recognition systems such as Red, Green, Blue (RGB), Hue, Saturation, Value (HSV), Hue, Saturation, Intensity (HIS) and Luma Chroma component-based color space (YCbCr). Hand segmentation based on RGB color space is not performing well with changing light conditions and HSV color space is found to handle illumination variations better than RGB [12]. Research on American SL recognition has mainly focused on using HSV and YCbCr color spaces. It is identified that YCbCr Color model is beneficial to optimize the performance of skin color clustering [13].

Factors like scale, rotation, occlusion affects the accuracy of gesture recognition in SL recognition systems, and Scale Invariant Feature Transform (SIFT), Speeded Up Robust Feature (SURF), Features from Accelerated Segment Test (FAST), Oriented FAST and rotated BRIEF (ORB) are commonly used algorithms for feature localization in images by minimizing the effect from the above factors.

SIFT algorithm was introduced as an invariant type of feature detector, and for improving the richness that feature descriptors can bring to feature matching [14]. SIFT describes an image using interest points(keypoints). Keypoint detection involves a multi-scale approach constructing a scale pyramid, convolving the upper and lower scales of the image using the ‘Difference of Gaussian’ (DoG) operator, and searching the local extreme in scale. This allows the localizing of regions on images that are invariant to scale and rotation variations [15].

SURF scales filter up instead of iteratively reducing the image size as in SIFT [10]. Comparative studies on these feature detection algorithms have shown that SIFT is better in feature detection, handling scaling and rotation effects [16], and SIFT based hand gesture detection has shown comparatively better results in performance in robustness to scale, illumination changes, and cluttered backgrounds [17]. Although SURF is faster than SIFT, detection is low in performance as well as not much invariant to rotation variations and illuminations [16].

K-means Clustering, Support vector machine (SVM), neural networks are commonly used classifiers in SL translator systems. SVM is relatively easier to build compared to neural networks but neural networks are performing well with multi-dimensional and multiclass classifications [12]. Deep learning and Convolutional Neural Network (CNN) based approaches have been using recently as an efficient classifier for gesture recognition with a higher rate of recognition accuracy [18]. Unlike traditional machine learning approaches where features are defined separately, CNN can learn to extract the features directly by itself from the training data [19].

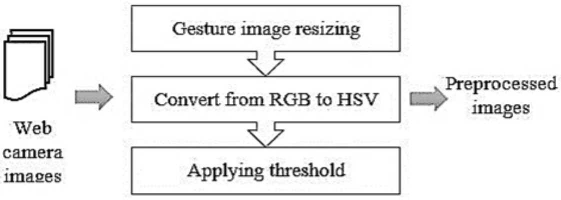

Figure 1 - Preprocessing of image data

Researches on SSL recognition and translation have been carried out focusing on appearance-based approaches, skeletal based approaches and 3D module based approaches. Research on image processing based SSL recognition using skin color filtering and centroid finding approach for the development of a still gesture mapping prototype has been conducted considering 15 gestures of the SSL alphabet [20]. This approach is low-cost and easy to implement but is less robust to multicolor backgrounds and needs to maintain a uniform distance between hand and camera. The impact of variation of scale and rotation effect on gesture recognition accuracy has not been considered in this research.

Another research has focused on developing a 3D avatar-based interpreter using visual basics and Blender animation software to display fingerspelled signs relevant to Sinhala text inputs in the form of a 3D avatar. Processing complexity of complicated gestures, challenges of texture detail handling of avatars are major problems in this approach. The research has used 61 SSL fingerspelling signs, 40 SSL number signs and shown rates above 70% for human-like animations of the signs [21].

A skeletal based approach for SSL to Sinhala translation has been carried out using a Kinect camera as the data acquisition device [22]. They have introduced a data normalization approach focusing on varying user heights to improve the gesture recognition accuracy with scale variations. Calculation of a central point based on ‘shoulder’ coordinates of the sign performer images has been used to ensure a user height or user distance independent recognition. Although this approach has given a recognition rate above 90% for a dictionary of fifteen signs in SSL, this is a complex and costly approach that requires additional devices like Kinect. This approach is difficult to use practically as the setup is not easily usable everywhere.

IV. PROPOSED METHODOLOGY

This research focuses on introducing a vision-based model to translate static gestures of SSL to Sinhala text. The major objective of developing the combined CNN-SIFT model is to achieve higher accuracy in SL gesture recognition using less training data and improve the robustness to scale and rotation variations of hand gestures. The proposed methodology consists of the major stages of data acquisition, image preprocessing, feature extraction, classification and displaying Sinhala text.

A. *Data acquisition*

The methodology follows a vision-based approach and a web camera is used to capture images of 20 selected static gestures of SSL. This supports the low-cost implementation as well as the enhancement of gesture recognition ability on low-quality images.

B. *Image preprocessing*

The images captured from a web camera which are RGB images are preprocessed to enhance features. Preprocessing is done by resizing images and color space conversion. HSV color space is more robust to illuminations compared to RGB so that images are converted from RGB to HSV. A mask is applied to HSV images to separate the hand region from the background based on skin color (Fig.1).

C. *Feature extraction*

Once the hand segmentation is completed features on the image needs to be extracted. The features that are in specific locations of the images are referred to as keypoints. During this stage, keypoints on preprocessed images are localized and SIFT feature descriptors are generated for each keypoint.

A SIFT descriptor includes details of coordinates of the relevant keypoint, orientation angle, response by which the strongest keypoint has been selected, and octave from which the keypoint was extracted. Since the size of each descriptor vary from image to image a uniform size feature vector is generated using K-means clustering. The result features from this stage are combined with the feature map from CNN to improve the robustness of the classifier to scale variations of hand.

D. *Classification*

A CNN classifier combined with features extracted from SIFT is used to classify gesture images to relevant gesture classes. The classification model consists of two input channels, one from CNN and the other from the SIFT as shown in Fig.2. The final fully connected layer will concatenate both feature vectors from SIFT and CNN layers to generate the final output of the gesture recognition model.

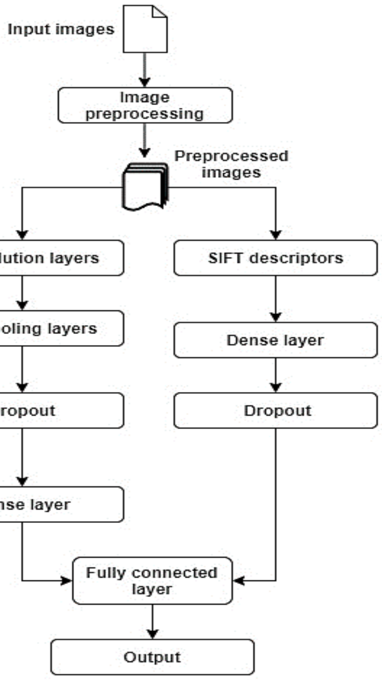

Figure 2 - Overview of proposed gesture classification model

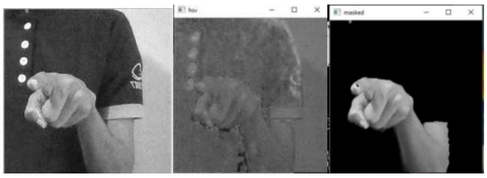

Figure 3 - Example of preprocessing a gesture image

E. *Displaying Sinhala text*

The classifier takes a gesture image as input and returns the class that the gesture image belongs to as a gesture ID. Displaying Sinhala text for the predicted gesture is done by gesture ID mapping based on a predefined gesture database. The database consists of gesture ID for each gesture and the corresponding meaning of the gesture as a Sinhala text. Once the mapping is successful the relevant Sinhala text for the input gesture is retrieved from the database and displayed as a text.

V. IMPLEMENTATION

The dataset of 20 selected static SSL gestures are collected using a web camera including 100 images per gesture. These image data are used to train and validate the classification model developed with the design as in Fig.2. The preprocessed dataset is separated into training (80%) and testing (20%) data sets.

Image preprocessing is done by resizing the images to 48x48 size, converting color space from RGB to HSV and applying a skin color mask to separate the hand region from the background (Fig.3). The HSV color ranges used are as HSV_min (0, 40, 30) and HSV_max (43, 255, 255).

Keypoints of input hand gesture images are located and SIFT descriptors are generated for each keypoint. A fixed-size feature vector is generated from the SIFT descriptors using K-means and bag of features so that the resulting vector of size K = 2048. These features are combined with a CNN feature map to improve CNN's robustness to scale variations of input hand gesture images.

The CNN consists of three convolution layers, three max-pooling layers followed by a dense layer. The combined feature map from both channels is used in the fully connected layer of CNN as an enhanced set of features for scale variations as in Fig.2. The concatenated output at the final dense layer is given as the output of the model. The training dataset is fed to the gesture classification model as input in the CNN channel and the SIFT feature vectors fed as the input for the SIFT channel to train the model.

The trained model is used for gesture classification. When an image is preprocessed and fed to the model it returns the classified gesture ID of the gesture. This gesture ID is mapped with a predefined gesture database of SSL, to retrieve and display the corresponding Sinhala text (Fig.4).

VI. RESULTS AND DISCUSSION

An image dataset with images of 20 selected static signs of SSL were used for training. Once the model is trained, it is tested and validated using the test dataset to check whether the model recognizes selected hand gestures correctly. The results were analyzed using the model's validation accuracy, precision, recall and F1 scores. Results were analyzed in two

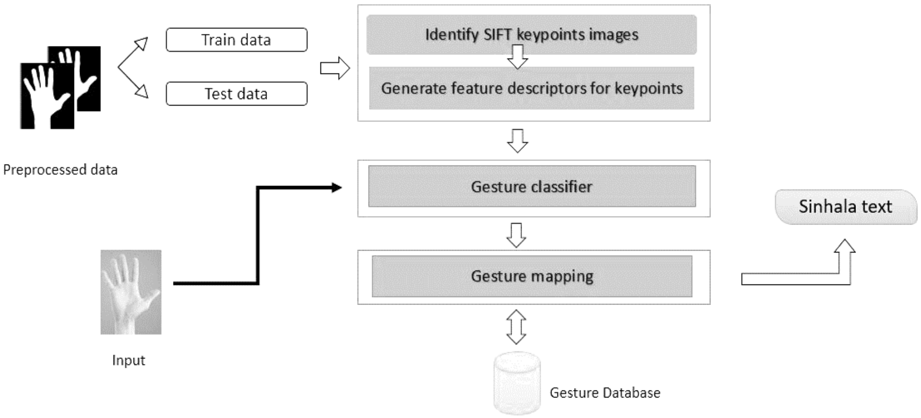

Figure 4 - Overview of the implementation of Sign language translator

ways as SIFT incorporated with CNN and without SIFT incorporated.

An average validation accuracy of 86.5% was achieved when the combined CNN-SIFT model was tested with images of 20 gestures of SSL. The majority of the gestures showed precision, recall and F1 scores above 0.8 but a few gesture classes like ‘මට’, ‘ළමයා’, and ‘මගේ’ showed low F1 values. Out of the 20 gesture classes, 13 showed more than 50% correct Sinhala text predictions when the distance between hand and camera were changed to 40cm and 60cm respectively. An average accuracy of only 68% was achieved when the SIFT features were not combined with the CNN classifier while keeping other parameters unchanged. The accuracy achieved was high in the combined CNN-SIFT approach than a single channel CNN classifier.

VII. CONCLUSION

Based on the analysis of different SL translation researches, vision-based approaches are identified as low-cost efficient solutions for SSL translation. Scale variations of hand and varying distance to camera affect the performance of vision-based SL gesture classifiers. The proposed SSL to Sinhala text translation model showed robustness and good performance for distance and scale variations of hand compared to some previous research. A combination of SIFT features with a CNN improved the robustness to scale variations in SL recognition and was implemented easily at a low cost. A combination of SIFT features improved the model accuracy compared to a single-channel CNN implementation. Further improvements can be carried out by extending the approach for dynamic SSL gesture recognition and combining facial expressions of SSL in the gesture recognition process.

REFERENCES

[1] H. Rupasinghe, Sri Lanka Terribly short of sign language interpreters, Article by Daily Mirror, http://www.dailymirror.lk/article/Sri-Lanka-Terribly-short-of-sign-language-interpreters-147833.html, 2018

[2] United Nations Economic and Social Commission for Asia and the Pacific (UNESCAP), Disability at a Glance 2019: Investing in Accessibility in Asia and the Pacific -Strategic Approaches to Achieving Disability-inclusive Sustainable Development, https://www.unescap.org/sites/default/files/publications/SDD-DAG-2019.pdf, 2019

[3] Department of Census & Statistics - Ministry of Policy Planning and Economic Affairs - Sri Lanka, Census of Population and Housing 2012, http://www.statistics.gov.lk/PopHouSat/CPH2011/Pages/Activities/Reports/FinalReport/FinalReportE.pdf, 2012

[4] https://www.nidcd.nih.gov/health/american-sign-language, accessed on 15/06/2020

[5] https://blog.ai-media.tv/blog/sign-language-alphabets-from-around-the-world, accessed on 18/07/2020

[6] https://socialservices.gov.lk/web/index.php, accessed on 20/06/2020

[7] https://newsin.asia/sri-lanka-recognize-sign-language-official-language, accessed on 15/06/2020

[8] https://www.rohanaspecialschool.org/sri-lankan-sign-language-dictionary, accessed on 12/05/2020

[9] Chathuranga P. D. T., Lorensuhewa S. A. S. and Kalyani M. A. L., Sinhala Sentiment Analysis using Corpus based Sentiment Lexicon, International Conference on Advances in ICT for Emerging Regions (ICTer)-2019

[10] Ming J.C, Omar Z and, A review of hand gesture and sign language recognition techniques, International Journal of Machine Learning and Cybernetics, 2017

[11] Tong Z, Huifeng L, Zhaojie J and Chenguang Y, Hand Gesture Recognition in Complex Background Based on Convolutional Pose Machine and Fuzzy Gaussian Mixture Models, International Journal of Fuzzy Systems, vol. 22, no. 4, 2020

[12] Munir O, Ali A.N and Javaan C, Hand Gesture Recognition Based on Computer Vision: A Review of Techniques Journal of imaging, vol.6,2020

[13] Shivashankara, S and Srinath, S, American Sign Language Recognition System: An Optimal Approach, International Journal of Image, Graphics and Signal Processing, 2018

[14] Davies, E.R, Computer and Machine Vision: Theory, Algorithms, Practicalities – Fourth Edition, Department of physics, University of London, 2012

[15] Lowe D.G, Distinctive Image Features from Scale-Invariant Keypoints, 2004

[16] El-gayar M.M, Meky N and Soliman H, A comparative study of image low level feature extraction algorithms, Egyptian Informatics Journal (2013), vol. 14, no.175, 2013

[17] Kaur J and Krishna C.R, An Efficient Indian Sign Language Recognition System using Sift Descriptor, International Journal of Engineering and Advanced Technology (IJEAT), vol.8, no.6, 2019

[18] Bowden R ,Camgoz N.C, Hadfield S, Koller O, Using Convolutional 3D Neural Networks for User-Independent Continuous Gesture Recognition, 23rd International Conference on Pattern Recognition (ICPR),2016

[19] Mundher Al-Shabi, Cheah W.P and Connie T, Facial Expression Recognition Using a Hybrid CNN–SIFT Aggregator,11th International Workshop, MIWAI 2017,vol.10607,no.139,2017

[20] Dissanayake, M.B, Herath, H.C.M, Kumari, W.A.L.V and Senevirathne, W.A.P.B, Image based sign language recognition system for sinhala sign language, SAITM Research Symposium on Engineering Advancements (SAITM – RSEA 2013),2013

[21] Punchimudiyanse M and Meegama R.G.N, Animation of Fingerspelled Words and Number Signs of the Sinhala Sign Language, International Conference on Engineering and Technology - Colombo, 2015

[22] Fernando P and Wimalaratne P, Sign Language Translation Approach to Sinhalese Language, GSTF Journal on Computing (JOC), vol. 5, no.1, 2016