Multi-modal Deep Learning Approach to Improve Sentence level Sinhala Sign Language Recognition

H.H.S.N. Haputhanthri, H.M.N. Tennakoon, M.A.S.M. Wijesekara, B.H.R. Pushpananda, H.N.D. Thilini

Abstract— **Sign language is used across the world for communication purposes within hearing-impaired communities. Hearing people are not well versed in sign language and most hearing-impaired are not good in general text, creating a communication barrier. Research on Sign Language Recognition (SLR) systems have shown admirable solutions for this issue. In Sri Lanka, machine learning along with neural networks has been the prominent domain of research in Sinhala SLR. All previous research is mainly focused on word-level SLR using hand gestures for translation. While this works for a certain vocabulary, there are many signs interpreted through other spatial cues like lip movements and facial expressions. Therefore, translation is limited and sometimes the interpretations can be misleading. In this research, we propose a multi-modal Deep Learning approach that can effectively recognize sentence-level sign gestures using hand and lip movements and translate to Sinhala text. The model consists of modules for visual feature extraction (ResNet), contextual relationship modeling (transformer encoder with multi-head attention), alignment (CTC) and decoding (Prefix beam search). A dataset consisting 22 of sentences used for evaluations was collected under controlled conditions for a specific day-to-day scenario (a conversation between a vendor and a customer in a shop). The proposed model achieves a best Word Error Rate (WER) of 12.70 on the testing split, improving over the single-stream model which shows a best WER of 17.41, suggesting a multi-modal approach improves overall SLR.**

Keywords— **Sign language, Sinhala Sign Language, Continuous Sign Language Recognition, Deep Learning, Multi-modal fusion.**

I. Introduction

Sign language is the main method used by the hearing-impaired communities for communication purposes. Different varieties of sign language can be found in many countries across the world , and even in a single country there may be a variety of regional variations developed within individual communities . According to the World Federation of the Deaf, there are more than 72 million people in the world who are deaf, distributed approximately among 300 different sign languages [1].

A. Sinhala Sign Language (SSL)

There are around 70,000 people who use sign language in Sri Lanka [2]. SSL, which is the official sign language in Sri Lanka , consists in excess of 2,000 sign -based words ,and they are interpreted through combinations of body pose , hand gestures , lip movements , facial expressions and eye gaze [2].

B. Sign Language Recognition (SLR)

A vast majority of the hearing communities are not well versed in sign language and most of the hearing-impaired are not literate with common languages, creating a communication barrier between the two communities. The most practiced solutions are to use interpreters or written material to exchange ideas which are quite impractical in common conversational scenarios. Hence a need for SLR had emerged with the purpose of translating expressed sign language to a more familiar form of communication for the hearing community. To address this issue, automatic SLR systems were developed in various methods to convert sign language into speech or text.

C. Continuous Sign Language Recognition (CSLR)

SLR can be undertaken in two main approaches [3] as, Static and Dynamic, where Dynamic SLR is further divided into two [3] as, Isolated (word level) Sign Language Recognition (ISLR) and Continuous (sentence level) Sign Language Recognition (CSLR).

CSLR is the most practically applicable approach even though it presents more challenges than ISLR, such as large vocabulary and scalability, movement epenthesis (ME) detection and elimination, person dependent variations, and sub-unit modeling resulting in low recognition accuracies for real world data [3]. To overcome these challenges, various approaches have been proposed under CSLR with the use of modern technical expertise, and hence is the focus of our research.

D. Multi-modality

Most of the previous researches have focused on individual modalities, such as the face ([4], [5]), head pose [6], mouth [7]-[9], eye-gaze [10] and body pose ([11], [12]), where the features can be classified as, manual features (intentional expressions when performing a sign such as hand gesture and body movement) and, non-manual features (un-intentional expressions when performing a sign such as lip movement and eye gaze). Researchers have mainly attended on using the manual features for SLR [13]-[15], and have ignored the important and rich information

Correspondence: Naveen Tennakoon (E-mail: naveentennakoon27@ gmail.com) Received:08.03.2023 Revised:22.03.2023 Accepted: 19.06. 2023

Naveen Tennakoon, Sachini Haputhanthri, Sachini Wijesekara, Randil Pushpananda and Dinusha Thilini are from University of Colombo School of Computing, Sri Lanka. (naveentennakoon27@gmail.com, haputhanthri.sn@gmail.com, sachinimadhumani3@gmail.com, rpn@ ucsc.cmb.ac.lk, hnd@ucsc.cmb.ac.lk)

DOI: https://doi.org/10.4038/icter.v16i2.7264

in the non-manual features. As a work-around, multi-stream Deep Learning frameworks [16], [17] show improvements in translating these complex spatial cues in sign language.

E. *Proposed Model*

Previous research work for SSL lacks the focus on multi-modality and Deep Learning concepts, and due to the novelty, there is no appropriate vision-based public dataset available to be used for CSLR purposes. To address these problems, we have proposed a novel model for CSLR of SSL with Deep Learning, integrating both hand and lip movements of a signer for sign recognition.

II. Related Work

A. *Sinhala Sign Language Recognition*

Many researchers in Sri Lanka have proposed several approaches to interpret signs performed by the hearing-impaired. However, the majority of such research for SSL is focused on recognizing static sign gestures and understanding dynamic sign gestures is not well investigated [18]. Additionally, they were conducted for ISLR using traditional (statistical model-based) non-vision-based approaches or vision-based approaches for a limited dataset, considering only hand gestures.

Fernando and Wimalaratne (2016) [3] have proposed a vision-based machine learning approach for communication with a chat application which includes a 3D avatar for imitating detected words and a software-based prototype which can translate a set of SSL into Sinhala words. The proposed approach is limited to 15 sign based static gestures.

Perera et al. (2017) [19] have developed a mobile application for both text and voice conversion in ISLR. The hearing-impaired person gets the text from the hearing person and then it gets converted into SSL in GIF format. They have proposed a sensor-based 2D model for finger joints in which the hearing-impaired can create the sign by stretching the fingers [20].

Madushanka et al. (2016) [21] have proposed a wearable armband which is composed with a combination of both gestural (using data from surface Electromyography) and spatial (using inertial measurement unit data) references of the hand and finger movements. They have used only one armband to make the study less complicated, hence only single-handed signs were selected, and simpler signs were chosen for the dataset.

Dilakshan and Priyadarshana (2020) [22] have proposed a novel vision-based approach to recognize Sinhala static sign gestures which incorporates only hand and finger movements using a Convolutional Neural Network (CNN). Their main intention was to translate SSL into Sinhala and Tamil texts with 12 basic Sinhala signs using 26,000 static images.

Dissanayake et al. (2020) [18] proposed a Deep Learning approach for a mobile application which can interpret both static and dynamic word-level signs. Although their experimental results show a significant performance, the used dataset sample size of 2,400 images is too small to evaluate the model’s effectiveness.

B. *Vision-based Sign Language Recognition*

SLR research work can be classified into two as sensor-based approaches and vision-based [16] approaches. Many research have been conducted related to sensor-based SLR but most of the commercial models are not much user-friendly and are expensive, that reduces the application and usability of such a model to a great extent.

Unlike sensor-based techniques, vision-based SLR is performed on image or video inputs captured through a camera device. This approach is more familiar for the users [23] than the sensor-based approaches because it is easier to use and has comparatively lower computational cost, and there are less problems created by limitations in users’ hand motion when performing signs [24]. The camera, computer, and software required to process data are the only costs in a vision-based recognition system.

C. *Vision-based Sign Language Recognition using Deep Learning*

Deep Learning is the state-of-the-art in vision-based SLR systems [3] and has proven to be more efficient than the traditional methods such as Dynamic Time Warping/Hidden Markov Model (DTW/HMM) based models [23], [25]. Deep Learning approaches may use data in a variety of formats in addition to numerical and textual input including images, videos and audio etc. Deep Learning evaluates data features and correlations entirely using neural networks, whereas statistical models involve the use of automated pre-existing algorithms to construct the desired model. Complex backgrounds, coarticulation, short-signs, finger spellings interleaving with dynamic signs etc. still are some of the challenges in Deep Learning models [3].

D. *Continuous Sign Language Recognition (CSLR)*

According to Aloysius and Geetha (2020) [3], CSLR is tackled in two ways.

1) Continuous sequence recognition with word boundary identification: Here, sequences of continuous videos are decomposed into isolated sign gestures. Tackling ME phases [25] is one of the main issues in this method. Also, both hand motion features and hand shape features are required to detect the sign word boundary in a sentence, because a noticeable number of sign gestures in sentences cannot be correctly detected by using only hand motion features [23]. In most of the existing research [23], [25] of this approach, classification of signs is based on traditional approaches.

2) Continuous sequence recognition without explicit word boundary identification: Most CSLR architectures contain a visual model that extracts the visual features from the input frame sequence, a contextual model to further find the correlation between the frames and an alignment model to investigate the correct mapping between the frame sequence and the gloss sequence. In the research ([5], [26], [27]), HMMs are used as the alignment model. But since frequent re-alignment is necessary for prior estimations and to reduce the exposure bias problem in Seq2Seq architectures, in more recent studies [28]-[31] researchers have adopted the Connectionist Temporal Classification (CTC). In contrast to HMM, CTC provides a soft full-sum alignment.

E. Lip Reading

Researchers have noticed the significance of lip movement when recognizing sign gestures. Therefore, it is important to understand how to model the lip movements in sign language. The following literature understands effective lip modeling for lip reading in sequential videos.

Stafylakis and Tzimiropoulos (2017) [32], have proposed a three-tier architecture lip reading showing promising results in the domain. A spatial-temporal frontend with 64 convolutional layers, a 34-layer residual network [33], and a 2-layer BLSTM backend comprises the final end-to-end model. The research focuses on establishing the fact that 3D convolutions outperform 2D convolutions in visual feature extraction and Long Short-Term Memory (LSTM) units are better for sequence learning than a temporal convolutional backend.

The research done by Lu and Li (2019) [34], incorporates a similar kind of approach as [32]. They have proposed a model with a CNN frontend (VGG19 network [35]) followed by an attention mechanism, ending with a unidirectional LSTM backend. CNN frontend is used for spatial information extraction and the LSTM backend is used for temporal feature learning. An attention layer is added to provide more attention to necessary parts in extracted features, giving the decoding LSTM network to learn more efficiently.

F. Multi-modal Fusion

More recently, multi-stream architectures have demonstrated excellent performance [26],36] over single stream models.

Instead of simple fusion, Zhou et al. (2020) [17] have proposed a Temporal Multi-Cue module which aims to combine spatial-temporal information from two aspects (intra-cue and inter-cue), with the goal of preserving the uniqueness of each cue while also exploring the interaction between cues. They have used full-frame, face, hands and body pose as the streams. From the experimental results, they have concluded that by considering the synergy of multi-cues, it shows a better performance and the lowest WER when compared to the single cues.

Camgoz et al. (2020) [37] have proposed a transformer architecture to incorporate both non-manual and manual features with hand shapes, lip movement and upper body pose as the three streams. In the above research, the data streams are fused using a deep fusion strategy called ‘late fusion’.

Zheng et al. (2021) [38] have proved the significance of non-manual features by considering facial expressions in sign language. They have used two streams, facial stream and mainstream where the mainstream has normal image frames or image frames with masked face areas or image frames with extracted human pose features. And for the fusion they have used 4 techniques: the concatenated method, the multi-level convolution method, the multi-head attention method and the non-local block method.

From the experimental results of the above research, the multi-modal approach shows a significant improvement in performance, and also that the multi-stream model integrating non-manual features is more powerful in CSLR than the single-stream model.

III. METHODOLOGY

The proposed model follows a multi-modal fusion Deep Learning architecture which combines the mechanisms of two identical single stream models (each for full frames and extracted lip frames). This section explains the individual model and the proposed fusion model separately.

A. Single Stream Models

The single stream model architecture (Fig. 1) consists of several distinct sections.

1) Visual Feature Extraction: Let the input frame sequence x = (x_1, x_2, …., x_T) where T is the total number of pre-processed frames. To overcome the overfitting of the model and to improve the time efficiency, 75% of the total frames are randomly dropped such that T’ = T x 0.25. First, using the ResNet CNN architecture pre-trained on ImageNet [39], the spatial visual feature frames (f) of the T’ number of frames are extracted independently, where f = (f_1, f_2, …., $f_{T'}$) are 512-dimensional feature vector representations of the input frames. There are 34-layer plain networks in the ResNet architecture that is inspired by VGG-19 in which the skip connections are added.

2) Contextual Relationship: An encoder with 2 transformer layers consisting of 4 attention heads each (for multi-head attention [40]) is used to train the model and an input shape of [T’, 512] dimensions is given as an input to the transformer. If the inputs are considered batch-wise, [bs, T’, 512] will be the input shape where ‘bs’ is the batch size, and all frames will be padded before going through the attention layers to the maximum time step length (T'_{max}) in the input batch.

Similarity scores between the signs of a sentence are computed using Self-attention layers with a scaled dot product scoring mechanism [40]. The input to the attention mechanism is comprised of queries (Q), keys (K) and values (V) with dimensions d_q, d_k, and d_v respectively, created from the input feature frame vectors. Dot products between the query and all the keys are calculated and each result is divided by $\sqrt{d_k}$, and the weights of the values are obtained with the use of a softmax function. The distance considered for attention calculation is Relative Positional

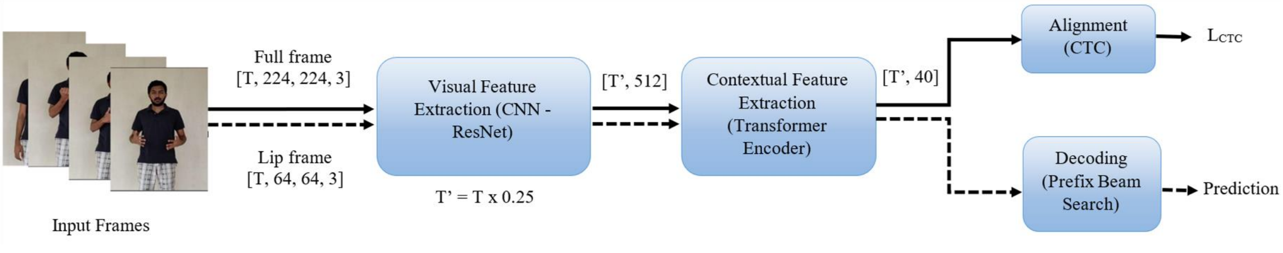

Fig. 1 - Single stream model architecture

Encoding [40]. The output of the matrix is calculated as,

$$\text{Attention}\ (Q, K, V) = \text{softmax}\ (QK^T/\sqrt{d_k})\ V \qquad (1)$$

Each of the layers in the transformer encoder has a fully connected position-wise feed-forward network, which is applied to each frame individually involving two linear transformations with a ReLU [40] activation function in between them. Both the attention layer and the feed forward network are wrapped with residual skip connections. The dimensionality of the output is [T', 512] same as the input, and the inner layers has a dimensionality of [T', 2048] where the attention weights are calculated from.

The output vectors then are passed through a linear classification layer and then through a softmax function to get the output probabilities for each T' time frame. The classification layer outputs a [n+1] dimensional vector for each T' time step where n is the vocabulary of the dataset (1 is used to denote the blank/space between each word occurrence).

3) Alignment and Decoding

The approach uses a CTC loss function for the training, given a matrix where columns correspond to timesteps (T') and each row corresponds to a word in the vocabulary. The CTC loss is used to maximize the sum of probabilities of all possible mappings between the frame sequence and the target label sequence. To represent the transition between two consecutive labels, The vocabulary (C) is extended with a blank label '-', such that the new vocabulary V = C∪{blank}. During the training of the model, the CTC loss is minimized such that p(y|x) is maximized. CTC loss is defined as,

$$L_{CTC}(x,y) = -\log p(y|x) \qquad (2)$$

Where x is the input frame sequence, and y is the target sequence. During inference, to decode the conditional probability sequence the prefix beam search algorithm [39] is used which is a breadth-first search algorithm that restricts the search space to reduce both the computation time and memory requirements.

B. Fusion Model

The best-found multi-modal architecture (Fig. 2) is based on fusing baseline and lip-based models, combining the separate transformer encoder outputs to create a fused Q, K, V representation. Each individual model acts according to the flow of the single stream models which calculates their individual losses (L_{ff}, L_{lf}). The output feature vectors of the separate transformer encoders are used in creating the fused input (Q from full body model, K and V from lip-based model). This combined representation is passed through the same transformer encoder to obtain the temporal dependencies extracted from the combined representation as shown in Fig. 3. This technique extracts the unique temporal correlations of the full body frames in relation with the lip frames context creating a unique feature representation that encapsulates the temporal features and attentions of both input streams. The output feature vector which has the shape [T', 512] then passes through a classification layer and a softmax function to obtain the final predictive probability distribution (x_{out}) for each frame in the T' time sequence. Then the model is followed by the CTC loss calculation (L_{com}) same as the individual models during training. The final loss for the multi-modular architecture training is taken as the average of all the losses.

$$\text{Final Loss}\ (loss_{avg}) = (L_{ff} + L_{lf} + L_{com})\ /\ 3 \qquad (3)$$

During inference of the model, only the output probability distribution of the combined transformer encoder (x_{out}) is considered since this is the probability sequence that incorporates both separate feature extractions and the combined representation dependency extractions. The prefix beam search algorithm is used in decoding the output probabilities to get the final predictions of the input sequence.

IV. Model Implementation

Implementation of the models were done using Python 3.6.13. The main libraries included were PyTorch (for model layers, utilities and networks) and TorchZQ (for model training, testing and evaluation helpers). All model functions were operated using the 'Legacy Runner' of TorchZQ utility classes, and the final quantitative metric calculations were obtained through the 'jiwer' library. Experimentation models were trained on a 11GB 'GeForce RTX 2080 Ti' GPU. Model testing and inference is detached from GPU usage, and therefore can be used within any device that has adequate CPU capabilities.

V. Dataset Preparation

A. Dataset Description

The dataset (SLSL-22) was collected on 22 SSL sentences with a vocabulary size of 40 words. At most 10 repetitions of one sentence were performed by 23 candidates (12 males and 11 females), including a mixture of signers and non-signers as it was very difficult to collect video clips only from hearing-impaired candidates mainly due to the prevailing Covid pandemic situation. When cleaning the data some videos had to be discarded since those were not up to the expected quality standards or simply had erroneous gestures. The final dataset includes a total of 3221 videos. An approximate 4:1 train-test split was made resulting in 2535 videos and 686 videos for the training and testing splits respectively. The training split features 13 candidates, and the testing split features all the 23 candidates (including the 13 candidates for training).

B. Data pre-processing

The initial captured frame size of videos was 1280x720p which were resized to 640x360p to improve lip frame extraction efficiency

1) Baseline Model: For the baseline model, the full frames after the preliminary resizing were again resized to 256×256p. For the training dataset, random cropping of size 224×224p was used and for the testing dataset, center cropping of size 224×224p was used. During both training and testing, a colour jitter of 0.1 was applied.

2) Lip Model: For the lip image-based model, lip frames of size 64×64p were extracted from each base image (640x360p full frame) using the MTCNN [41] network. No cropping was applied to frames since the lip images contained subtle information that could get lost if a crop of the image was to be taken. Additionally, during training and testing a colour jitter of 0.1 was applied.

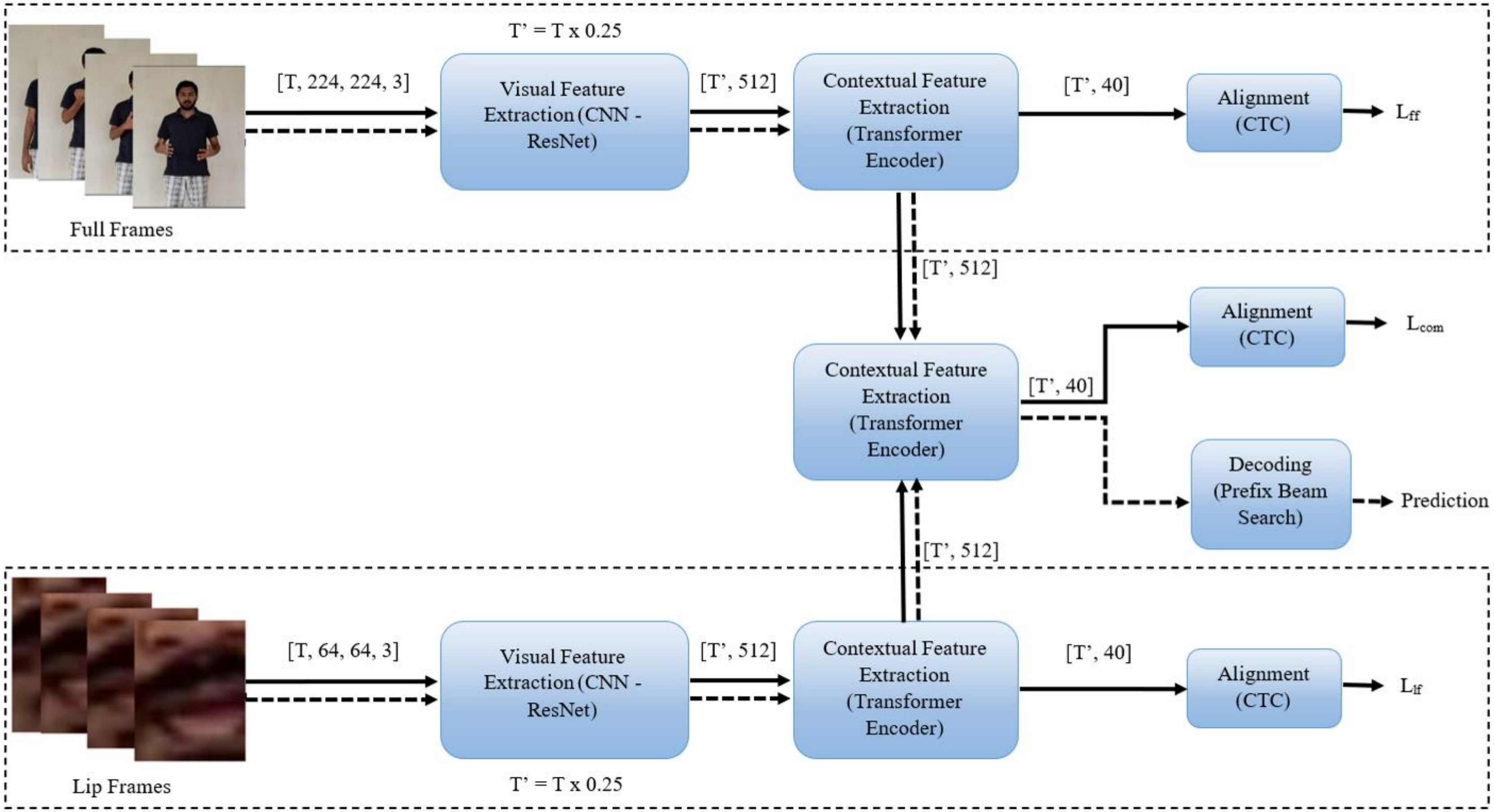
T' = T x 0.25
[T, 224, 224, 3]
Visual Feature Extraction (CNN - ResNet)
[T', 512]
Contextual Feature Extraction (Transformer Encoder)
[T', 40]
Alignment (CTC)
L_{ff}
Full Frames
[T', 512]
Alignment (CTC)
L_{com}
Contextual Feature Extraction (Transformer Encoder)
[T', 40]
Decoding (Prefix Beam Search)
Prediction
[T', 512]
[T, 64, 64, 3]
Visual Feature Extraction (CNN - ResNet)
[T', 512]
Contextual Feature Extraction (Transformer Encoder)
[T', 40]
Alignment (CTC)
L_{lf}
Lip Frames
T' = T x 0.25

Fig. 2 - Multi-modal fusion architecture (All three contextual models are the same transformer and is depicted separately for the ease of understanding)

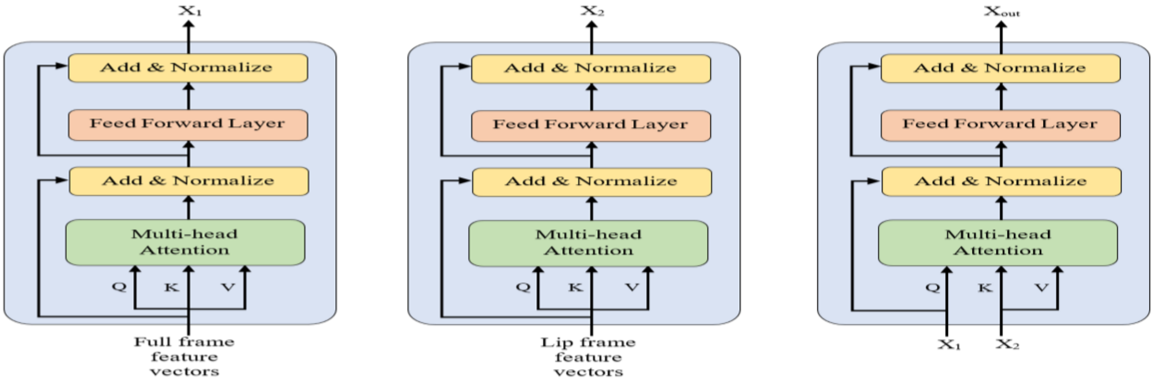

Fig. 3 - Passing of feature vectors through the transformer encoder. Extracting contextual relationships in full frame feature vectors *(left)*, Extracting contextual relationships in lip frame feature vectors *(middle)*, Extracting full frame temporal relationships with lip frame context

VI. Evaluation Metrics

The main quantitative evaluations needed are to substantiate the effectiveness of the proposed model. In a CSLR system mostly the metric used to calculate this is the Word Error Rate (WER) of the model. To get the WER, insertions, substitutions and deletions that occur in a sequence of identified words in the output label are added and are divided by the total word count of the truth label.

$$\text{WER} = \frac{(\#\text{Substitutions}) + (\#\text{Insertions}) + (\#\text{Deletions})}{(\#\text{Words in the Ground Truth Label})} \quad (4)$$

This metric is used in the baseline model, lip model and the final proposed model for a comparison between the performances of the models. Only the lowest WER observed during evaluations of each testing epoch is considered as the final WER of each specific model.

$$\text{SWA}= \frac{(\#\text{ Predictions including correct complex signs})}{(\#\text{ Predictions})} \quad (5)$$

To evaluate the proposed model on the recognition capability of the words that have similar manual features, we considered the accuracy of each such word (Specific Word Accuracy - SWA) being identified in each predicted sentence at any position within the predicted sentence. We do not consider the position-wise correctness as in WER calculation since these important words may deviate from their actual position due to the mispredictions of neighbouring words in a particular sentence. Therefore, if within a predicted sentence, that referred specific word exists then it is regarded as a correct prediction.

The implemented prototype application is quantified with regard to response time, in order to evaluate the usability of the application in a real world scenario. Average time taken for loading the model within the application, average time taken for lip extraction of a real time captured video and average time taken for video pre-processing and model prediction are calculated to obtain an overview of the performance of the application.

VII. Results and Discussion

A. Baseline Model

The ultimate multi-modal fusion is generated by starting with the baseline structure, which is basically a combination of a couple of single stream models (body and lip models).

The best baseline model architecture was based on the experimentation results (TABLE I and TABLE II) on several state-of-the-art visual feature extraction and contextual relationship extraction networks.

The models were trained for different epochs starting from 2 and with the increase in number of epochs the WER value decreased. Additionally, all the training curves with regards to the training loss displayed a similar pattern.

Therefore, by considering the time taken to train the model with the increase of the number of epochs, 30 was selected as the optimal number of epochs for training and testing all models. To find the optimal and the most effective hyperparameters for the baseline model, comprehensive experimentations were conducted as follows,

1) Model Dimensions: The relationship between the performance of the model and the dimensionality of the model is observed for different model dimensions. The obtained experimental results are presented in TABLE III, which shows that by increasing the dimensions, the WER is further reduced.

TABLE I
Results for Different CNN Architectures

CNN Architecture	WER
VGG-19	84.10
AlexNet [42]	23.69
ResNet 50	22.66
ResNet 34	20.96
ResNet 18	17.41

TABLE II
Best WER on Different Encoder Architectures

Encoder Architecture	WER
RNN	94.54
BLSTM	82.94
Transformer with Multi-head attention	17.41

TABLE III
BEST WER ON DIFFERENT MODEL DIMENSIONS

Model Dimension	WER
2048	31.06
1024	21.57
512	20.89

2) Number of Attention Heads: The model is trained and observed for different numbers of attention heads in the attention layer to identify the relationship of the model with attention heads. The obtained experimental results are presented in TABLE IV, which confirms that by increasing the number of attention head counts, the WER is further reduced.

3) Number of Transformer Layers: According to TABLE V, the WER value decreases with the increment of the number of layers. But it was also observed that the model parameter size exponentially increased with the number of transformer layers (1: 14618432, 2: 17775168, 3: 20931904), and therefore 2 layers were selected as the optimal number of layers for the transformer encoder.

4) Frame Dropping Percentage: During model training, to overcome overfitting and reduce training overhead, a fixed proportion of frames are randomly discarded from a continuous video sequence by uniform sampling. If the frame dropping rate is too low, as there were fewer variations within the frames and there was a lot of redundant information that was not necessary for sign recognition, the WER values showed lesser results as opposed to the WER values obtained for higher drop percentages. When 75% of the frames were dropped it showed the best result and the WER value was reduced to 17.41 (TABLE VI) and showed a better performance than other percentages.

TABLE IV
BEST WER ON DIFFERENT NUMBER OF ATTENTION HEADS

Number of Attention Heads	WER
1	24.71
2	24.23
4	20.89

TABLE V
BEST WER ON DIFFERENT NUMBER OF TRANSFORMER LAYERS

Number of Transformer Layers	WER
1	23.75
2	20.89
3	20.14

TABLE VI
BEST WER ON FRAME DROPPING PERCENTAGES

Frame Dropping Percentage	WER
0 %	21.09
25 %	19.59
50 %	20.89
75 %	17.41

By considering the obtained results, for both the single stream models (baseline and lip model) an 18-layer ResNet CNN architecture pre-trained on the ImageNet dataset and a transformer encoder consisting of 2 layers, 4 attention heads and a model dimensionality of d = 512 were adopted where a uniform frame dropping of 75% is used.

B. Proposed Fusion Model

The final proposed fusion model was tested alongside the single stream models. According to the results (TABLE VII) we can observe while the single stream lip model has failed to perform, the multi-modal approach shows significant improvements over both individual stream models.

Additionally, the Specific Word Accuracy was compared with the baseline and our proposed model (TABLE VIII).

The proposed model shows an increase in recognition accuracy of the specific words that have the same manual features but are differentiated with the non-manual features, which concludes that incorporating multi-modality has improved the recognition rate of the if not misclassified complex sign gestures.

Testing and the comparison of the results between the input streams were carried out for 3 other testing sets as well and the results are given below.

1) Test Set 01: The overall testing split includes videos of candidates in the training dataset (Different sample videos from the training split) and of candidates' exterior to the training dataset. This split was filtered to obtain the candidate videos exterior to the training split only, in order to check the generalization capability of the final proposed model. This testing set features 400 testing samples of videos recorded with 8 non-signer candidates and 2 signer candidates at 30 fps and 720p resolution similar to the video standards of the training dataset.

The results (TABLE IX) show that there is only a small deviation in the WER value compared with the final test split, and hence can conclude that the model performs just as well for people that do not feature in the training samples.

2) Test Set 02: Due to the Covid pandemic situation within the country, the number of actual signer candidates that were involved with the data collection procedure was limited. So, a separate limited (due to the lack of videos and the quality of videos) analysis was conducted with the available signer-only candidate sample videos. This helps to evaluate how well the model reacts to sign variations that are unique to actual hearing-impaired.

TABLE VII
Best WER Comparison on All Final Models

Model	WER
Baseline	17.41
Lip	29.69
Fusion (Baseline + Lip)	12.70

TABLE VIII
SWA comparison of Baseline and Fusion Models

Model	Specific Word Accuracy
Baseline	76.97 %
Fusion (Baseline + Lip)	84.40 %

TABLE IX
WER COMPARISONS FOR TEST SETS 01, 02 AND 03

Model	WER of Final Test Split	WER of Test Set 01	WER of Test Set 02	WER of Test Set 03
Baseline	17.41	21.59	66.77	82.38
Fusion	12.70	16.75	65.81	78.73

The videos were recorded with 2 signer candidates at 30 fps and 720p resolution. The testing set contains 142 video samples.

According to the results (TABLE IX), it can be observed that while the multi-modal approach outperforms the baseline model, the WER value difference has increased by a big margin, depicting that the model has not been trained enough to identify the signer only unique variations that hearing people were not able to mimic during dataset collection.

3) Test Set 03: Our dataset only consists of samples of individual variations of words (a word can have start and end movements based on the position of the word within a sentence) since it is our research focus. Optimistically, we also analyzed the capability of the final model to identify the different word variations, by testing the model on a new dataset of 20 sentences containing 34 interchanged words of the final dataset. The videos were recorded with 3 different candidates (non-signers) at 30 fps and 720p resolution. The testing set contains 300 videos in total.

As in the earlier study, we can observe (TABLE IX) that while the multi-modal approach outperforms the baseline model, the WER value is large compared with the final test. Even so, the best performing model (multi-modal fusion) shows encouraging results as the model was not at all trained to identify these word variations. The model can further be improved by incorporating word variation information in the training data.

C. Prototype Application

A prototype application was implemented upon Python as a Flask Application to evaluate the usage of the final proposed model in a real-world scenario. It consists of a single webpage composed of the camera recorder window and a single button to operate the camera recording. The video recording starts with the button press after a countdown, and it uses frame difference calculation thresholding [16] to stop the video recording detecting movement changes.

Both quantitative and qualitative analysis for the implemented prototype application were done through a structured questionnaire from a sample including selected candidates aged between 20 to 50 and a professional interpreter. Candidate sample consists of 11 signers and 9 non-signers including both males and females. Here, the two structured questionnaires were used separately for the candidates and the interpreter, and the obtained results were separately analyzed. For the quantitative analysis the frequency distribution of the responses was analyzed under three categories; usability, performance and overall idea of the application. And for the qualitative analysis, the user feedback including both candidates and the professional interpreter of the application was analyzed regarding their opinions, experiences and suggestions.

VIII. CONCLUSION

Our research proposes a novel fusion model for the task of CSLR in SSL based on a Deep Learning multi-modal architecture. Unlike traditional CSLR models that solely rely on visual inputs, our proposed model incorporates both visual and spatial-temporal features. By analyzing a frame sequence, the model can extract the spatial-temporal features by providing temporal attention to the frame sequence. This results in a more robust and accurate representation of the sign language gestures, which enhances the performance of the model in recognizing complex sign gestures. The proposed model outperforms the best baseline model on both; overall performance in sign language recognition (WER), and complex sign gesture recognition (SWA) for the created SLSL-22 dataset. The evaluation results demonstrate that the proposed methodology generalizes well to individual variations with the exception of signers due to the bias of the training split, which can be remedied by improving the dataset. Furthermore, the prototype application proves the effectiveness, performance, and the usability of the model within a practically applicable scenario. The application can help bridge the communication gap between the hearing and hearing-impaired communities in Sri Lanka by providing an efficient and accurate tool for real-time sign language interpretation. The proposed model has significant potential for improving communication accessibility and quality of life for the hearing-impaired in Sri Lanka and beyond.

IX. LIMITATIONS AND FUTURE WORK

The research is focused on the development of a model that is able to recognize SSL for the hearing-impaired community. However, the syntaxes between different sign languages may differ greatly from one another. As such, the model created here may not be generalizable for other sign languages like American Sign Language, Persian Sign Language, etc.

In order to create the dataset used in this study, sentence structure variations were ignored, and only sequence-to-sequence modeling was considered. The sample videos used in the dataset were collected under broad daylight in a uniform background, with the goal of preserving the scale of the candidate on the video. This study mainly focuses on single variations of words, and does not take into account the location of words within sentences, which can result in transitional movement variations for each sign. As such, the results obtained from this research are only applicable within the mentioned constraints.

As future work, there are several potential avenues that could be explored. For example, increasing the vocabulary used in the dataset and including different word variations could help to train the model on the most unique features in signs. Additionally, including more signers in the dataset could help to learn signer variations and improve the model's overall usability.

From the model's perspective, it could be interesting to investigate the effect of including other visual cues with the hand gestures for different interpretations in SSL. Furthermore, it may also be possible to investigate background elimination in video sequences as a way to build a more generalizable model. Overall, there are many potential avenues for future research in this area, and it will be exciting to see where this field goes in the coming years.

References

[1] D. Chatterjee, "International Day of Sign Languages: The Rights of Sign Language Users," *NDTV.com*, 23-Sep-2020. [Online]. Available: https://www.ndtv.com/india-news/international-day-of-sign-languages-2020-facts-and-human-rights-of-sign-language-users-2299705. [Accessed: 18-Apr-2022].

[2] P. Fernando and P. Wimalaratne, "Sign language translation approach to SINHALESE LANGUAGE," *GSTF Journal on Computing (JoC)*, vol. 5, no. 1, 2016.

[3] N. Aloysius and M. Geetha, "Understanding vision-based continuous sign language recognition," *Multimedia Tools and Applications*, vol. 79, no. 31-32, pp. 22177–22209, 2020.

[4] C. Vogler and S. Goldenstein, "Facial movement analysis in asl," *Universal Access in the Information Society*, vol. 6, no. 4, pp. 363–374, 2007.

[5] O. Koller, J. Forster, and H. Ney, "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers," *Computer Vision and Image Understanding*, vol. 141, pp. 108–125, 2015.

[6] M. Luzardo, M. Karppa, J. Laaksonen, and T. Jantunen, "Head pose estimation for sign language video," *Image Analysis*, pp. 349–360, 2013.

[7] E. Antonakos, V. Pitsikalis, I. Rodomagoulakis, and P. Maragos, "Unsupervised classification of extreme facial events using active appearance models tracking for sign language videos," *2012 19th IEEE International Conference on Image Processing*, 2012.

[8] O. Koller, H. Ney, and R. Bowden, "Read my Lips: Continuous SIGNER INDEPENDENT weakly Supervised Viseme Recognition," *Computer Vision – ECCV 2014*, pp. 281–296, 2014.

[9] O. Koller, H. Ney, and R. Bowden, "Deep learning of mouth shapes for sign language," *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2015.

[10] G. Caridakis, S. Asteriadis, and K. Karpouzis, "Non-manual cues in automatic sign language recognition," *Personal and Ubiquitous Computing*, vol. 18, no. 1, pp. 37–46, 2012.

[11] T. Pfister, J. Charles, M. Everingham, and A. Zisserman, "Automatic and efficient long term arm and hand tracking for continuous sign language tv broadcasts," *Procedings of the British Machine Vision Conference 2012*, 2012.

[12] J. Charles, T. Pfister, M. Everingham, and A. Zisserman, "Automatic and efficient human pose estimation for sign language videos," *International Journal of Computer Vision*, vol. 110, no. 1, pp. 70–90, 2013.

[13] B. S. Parton, "Sign language recognition and translation: A multidisciplined approach from the field of artificial intelligence," *Journal of Deaf Studies and Deaf Education*, vol. 11, no. 1, pp. 94–101, 2006.

[14] O. Koller, H. Ney, and R. Bowden, "Deep hand: How to train a CNN on 1 Million hand images when your data is continuous and WEAKLY Labelled," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[15] N. C. Camgoz, S. Hadfield, O. Koller, H. Ney, and R. Bowden, "Neural sign language translation," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[16] C. Gündüz and H. POLAT, "Turkish sign language recognition based on multistream data fusion.," Turkish Journal of Electrical Engineering & Computer Sciences, vol. 29, no. 2, 2021.

[17] H. Zhou, W. Zhou, Y. Zhou, and H. Li, "Spatial-Temporal Multi-Cue network for sign language recognition and translation," *IEEE Transactions on Multimedia*, pp. 1–1, 2021.

[18] I. S. M. Dissanayake, P. J. Wickramanayake, M. A. S. Mudunkotuwa, and P. W. N. Fernando, "Utalk: Sri lankan sign Language Converter mobile app using image processing and machine learning," *2020 2nd International Conference on Advancements in Computing (ICAC)*, 2020.

[19] Y. Perera, N. Jayalath, S. Tissera, O. Bandara, and S. Thelijjagoda, "Intelligent mobile assistant for Hearing IMPAIRERS to interact with the society In Sinhala language," *2017 11th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*, 2017.

[20] M. Ahmed, M. Idrees, Z. ul Abideen, R. Mumtaz, and S. Khalique, "Deaf talk using 3D animated sign language: A sign language interpreter using Microsoft's Kinect v2," *2016 SAI Computing Conference (SAI)*, 2016.

[21] A. L. P. Madushanka, R. G. D. C. Senevirathne, L. M. H. Wijesekara, S. M. K. D. Arunatilake, and K. D. Sandaruwan, "Framework for sinhala sign language recognition and translation using a wearable armband," *2016 Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer)*, 2016.

[22] S. Dilakshan and Y. H. P. P. Priyadarshana, "Convolutional neural Networks: A novel approach FOR Sinhala sign recognition system," *2020 11th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, 2020.

[23] M. Zadghorban and M. Nahvi, "An algorithm on sign words extraction and recognition of CONTINUOUS Persian sign language based on motion and shape features of hands," *Pattern Analysis and Applications*, vol. 21, no. 2, pp. 323–335, 2016.

[24] K. Murakami and H. Taguchi, "Gesture recognition using recurrent neural networks," *Proceedings of the SIGCHI conference on Human factors in computing systems Reaching through technology - CHI '91*, 1991.

[25] D. Kelly, J. McDonald, and C. Markham, "Recognizing spatiotemporal gestures and Movement epenthesis in sign language," *2009 13th International Machine Vision and Image Processing Conference*, 2009.

[26] O. Koller, N. C. Camgoz, H. Ney, and R. Bowden, "Weakly supervised learning with multi-stream cnn-lstm-hmms to discover sequential parallelism in sign language videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 9, pp. 2306–2320, 2020.

[27] O. Koller, S. Zargaran, and H. Ney, "Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMS," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[28] N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden, "Subunets: End-to-end hand shape and continuous sign language recognition," *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017.

[29] R. Cui, H. Liu, and C. Zhang, "Recurrent convolutional neural networks for continuous sign language recognition by staged optimization," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[30] H. Zhou, W. Zhou, and H. Li, "Dynamic pseudo label decoding for Continuous Sign Language recognition," *2019 IEEE International Conference on Multimedia and Expo (ICME)*, 2019.

[31] J. Pu, W. Zhou, and H. Li, "Iterative alignment network for Continuous Sign Language recognition," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[32] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with LSTMs FOR LIPREADING," *Interspeech 2017*, 2017.

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[34] Y. Lu and H. Li, "Automatic lip-reading system based on deep convolutional neural network and attention-based long short-term memory," *Applied Sciences*, vol. 9, no. 8, p. 1599, 2019.

[35] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).

[36] H. Zhou, W. Zhou, Y. Zhou, and H. Li, "Spatial-Temporal Multi-Cue network for Continuous sign language recognition," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 13009–13016, 2020.

[37] N. C. Camgoz, O. Koller, S. Hadfield, and R. Bowden, "Multi-channel transformers for multi-articulatory sign language translation," *Computer Vision – ECCV 2020 Workshops*, pp. 301–319, 2020.

[38] J. Zheng, Y. Chen, C. Wu, X. Shi, and S. M. Kamal, "Enhancing neural sign language translation by highlighting the facial expression information," *Neurocomputing*, vol. 464, pp. 462–472, 2021.

[39] Z. Niu and B. Mak, “Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition,” *Computer Vision – ECCV 2020*, pp. 172–186, 2020.

[40] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in neural information processing systems, vol. 30, 2017.

[41] J. Wen and Y. Lu, “Automatic lip reading system based on a fusion lightweight neural network with Raspberry Pi,” *Applied Sciences*, vol. 9, no. 24, p. 5432, 2019.

[42] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional Neural Networks,”*Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2