Novel Cognitive Assisted Adaptive Frame Selection for Continuous Sign Language Recognition in Videos Using ConvLSTM

Priyanka Ganesan[1], Senthil Kumar Jagatheesaperumal[2], Matheshkumar P[3], Silvia Gaftandzhieva[4], Rositsa Doneva[5]
Department of Computer Science and Engineering, Mepco Schlenk Engineering College, Sivakasi, India[1, 3]
Department of Electronics and Communication Engineering, Mepco Schlenk Engineering College, Sivakasi, India[2]
Faculty of Mathematics and Informatics, University of Plovdiv Paisii Hilendarski, Plovdiv, Bulgaria[4]
Faculty of Physics and Technology, University of Plovdiv Paisii Hilendarski, Plovdiv, Bulgaria[5]

***Abstract*—People with a hearing impairment commonly use sign language for communication, however, they find it challenging to communicate with a normal person who does not recognise the sign language. They normally require an intermediary human to act as a translator for convenient means of expressing their thoughts. To address this issue, the work aims to enhance their communication capability by eliminating the need for an intermediary person by developing a sign language converter that uses a vision-based dynamic recognition strategy to convert continuous sign language into multimodal output. This work introduces a deep neural network based on convolutional long short-term memory (ConvLSTM) networks to determine the real-time dynamic gesture recognition of the actions of the impaired persons captured through cameras. The investigations of the continuous sign language recognition (CSLR) were deployed on the Chinese Sign Language Dataset, CSL-Daily, Phoenix-2014 and Phoenix-2014T datasets and the performance comparisons were done for conventional LSTM, Gated Recurrent Unit (GRU) and ConvLSTM. Experimental results have shown that the ConvLSTM network outperforms the other techniques, and they can detect the sign actions with a better accuracy of 90%, and a precision rate of 0.93, which ensures interpreting the meanings for each sign sequence with ease by integrating the proposed novel cognitive assisted adaptive keyframe selection. The proposed system could be easily implemented in the modern learning management system.**

Keywords—ConvLSTM; GRU; keyframes; LSTM; sequential learning; sign language recognition

I. Introduction

People with hearing disabilities use sign language for communication in day-to-day life. When spoken communication is impossible, sign language is used to communicate through body movements, particularly hands and arms. Because deaf-dumb people mainly use it and a normal person does not learn it, interpreters are required for deaf and hearing people to communicate. According to the WHO report, around 2.5 billion individuals will encounter hearing loss by 2050 [1]. So, it is necessary to develop an automated translation system for communication with them and reduce the gap between the hearing and deaf communities. Sign language recognition is meant to exact meaning for each sign in continuation with a sequence of signs i.e., mapping visual signs with words. On the other hand, creating meaningful sentences from the extracted signs is known as sign language translation.

Most sign language recognition systems focus only on isolated signs rather than a continuous sign sequence. Most sign language recognition systems use data gloves, and sensor gloves with sensors (depth sensors, optical sensors, thermal sensors, and leap motion sensors) for gesture recognition. In glove-based methods, the signer must wear a hardware glove, from which gestures are recognised [2]. These sign language recognition methods, which ensure higher accuracy, will be quite awkward to use in public places. In [3], the sign actions performed by the deaf were captured using cameras. Then the keyframes were identified adaptively and for those frames features like body pose, hand poses, and finger orientations were extracted using CNN. This method captures only spatial information and ignores other crucial features. Most of the sign language translation in the literature lacks accurate temporal data [4-5] and faces various linguistic challenges.

To address this, a vision-based dynamic recognition method for real-time gesture recognition with cameras is proposed in this article for sign language recognition and translation. Features are extracted automatically and adaptively by video streams of signs and gestures made by the impaired persons. First, the video sequences are segmented into frames and keyframes are extracted. Second, in the feature extraction stages, temporal information was sequentially learned using LSTM [6]. Third, a ConvLSTM cell is constructed by replacing an LSTM structure with weighted convolution operations at each cell gate. Further, the convolution operation in a ConvLSTM cell was assistive in extracting short-term spatial correlations process between successive measurements within a single time step. This striking feature of a ConvLSTM cell was useful for capturing the signs and gestures of hearing-impaired persons by identifying long-term temporal dependencies. Finally, the experimental outcomes were compared with the standard recurrent neural networks Gated Recurrent Unit (GRU). Here, encoder-decoder architecture is constructed using LSTM and used for sentence generation. An overview of the proposed work is shown in Fig. 1.

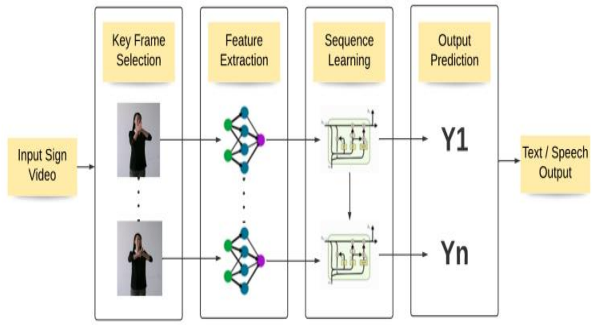

Fig. 1. System design for the proposed work.

The isolated output prediction description is as follows. (1) Key frame capturing based on similarity score for the input video. (2) Visual feature extraction for the identified keyframes. (3) Temporal information capturing through sequence learning using LSTM. (4) Isolated output prediction.

The proposed system will act as a machine-made interpreter that automatically converts continuous sign language into multimodal output, i.e., text or speech output.

This paper is organised as follows. Existing methods and related work are discussed in Section II. The proposed work with a detailed system design will be explained in Section III. Section IV will discuss experimental results. Conclusions for the entire work are given in Section V.

II. Related Work

In this section, state-of-the-art works that use glove and sensor-based approaches as well as vision-based approaches to recognise sign language were explored. Hand gesture recognition has been done in various ways, from which glove-based and vision-based are the most used. In glove-based methods, sensors attached to those gloves transfer electrical signals helping to determine the hand gesture. The signs made by the impaired will be identified from the acquired electrical signals. On the other hand, instead of a glove, the sensor camera could be used to capture the sign actions. Gestures will be classified from the captured sequences of images extracted from the video frames. Vision-based methods reduce the challenges and complexities compared with glove-based methods.

A. *Glove & Sensor based Approaches*

To create a hand sign recognition system, the authors in [7] used Electrical Impedance Tomography (EIT) imaging (gauss-newton image reconstruction algorithm) and robust CNN classification (support vector machine and softmax classifier). Mittal et al. [8] proposed a modified LSTM model for continuous sign language recognition. They captured hand gestures using a Leap Motion sensor and extracted 12 features. They fed 2D CNN feature maps into an LSTM model with a RESNET gate for output prediction. For sign language recognition, Deriche et al. [9] proposed a dual Leap Motion Controller (LMC), and to address the challenges of finger occlusions and missing data, they used both front and rear-side LMCs. Feature extractions were performed by selecting the set of the best geometric features from both controllers, such as finger length, width, hand roll, hand pitch, and hand yaw. They applied a Bayesian approach, a Gaussian mixture model and a simple linear discriminant analysis to do the final classification. An evidence-based fusion approach is used for combining data from two LMCs (Dempster-Shafer, theory of evidence). Marin et al. [10] proposed using LMC and Kinect devices to capture hand gestures. The LMC records finger distances, angles, and elevations, whereas the Kinect device records depth information. Theodorakis et al. [36] proposed lexicon-based sign language recognition. To improve recognition performance, leap motion data was combined with Kinect data, and the results were then classified using an SVM classifier. Further, for gesture recognition tasks, a one-against-one approach was delayed. The gesture with the most number votes was chosen as the desired output.

To distinguish American Sign Language (ASL) alphabets, Lee et al. [11] created a smart wearable with five flex sensors, two pressure sensors, and a three-axis inertial motion sensor. The device's embedded SVM classifier recognises alphabets. For gesture recognition, Huang et al. [12] designed a wearable glove with less graphene oxide fibre. It keeps track of the movement of ten joints in one hand. For British Sign Language (BSL) recognition, Dias et al. [13] used an instrumented glove with five flex sensors and two contact sensors. The information gathered is divided into three categories: construction, alphabet gesture, and relaxation period. For recognition, these data are fed into MLP-NN, KNN, SVM, RF, and NB classifiers. Li et al. [14] developed a sign language recognition system based on ultra-wideband radar. The Micro Doppler spectrogram input is used to calculate cumulative energy distribution, which divides the density bands for each cumulative energy distribution image. Gurbuz et al. [15] use a multi-frequency RF sensor network to measure ASL in a non-invasive, non-contact manner regardless of lighting conditions. Further, the authors used SVM, KNN, and random forest to classify the signs and compare their performances.

B. *Vision-based Approaches*

Guo et al. [16] used a combination of 3D CNN and LSTM to capture spatiotemporal representations in a video. A stacked decoding network is also used to predict gloss and query adaptive fusion is used to generate sentences. Zhou et al. [17] proposed a multi-cue framework (spatial multi-cue and temporal multi-cue) for sign language recognition and translation to learn spatial-temporal correlations of visual cues. They used 2D CNN to generate multi-cue features, CTC-Decoder for sign language recognition, and SA-LSTM for sign language translation in this study. Breland et al. [18] proposed a Deep CNN model for gesture recognition, that is light-independent and is based on high-resolution thermal imaging. Passos et al. [19] used a two-step method with feature mapping and classification for gesture recognition in videos. They used deep neural network architecture to segment each body part, then used gait energy images to encode body part motion.

For CSLR, Huang et al. [20] proposed a sequence-to-sequence learning method using keyframe-centered clips

(KCCs) split out from the input video. The CNN features of RGB keyframes, HOG of depth motion maps, and trajectory features of skeleton joints are fused by the feature fusion layer for feature extraction. Finally, all multimodal features are combined and fed into an LSTM model for sub-word and then word construction. For accurate CSLR classification, Wei et al. [21] proposed a semantic boundary detection method based on reinforcement learning. Initially, a discriminative representation for sign video with multiscale perception loss is learned using a spatial-temporal CNN and bidirectional LSTM. Each segment's clip-level features were refined between adjacent boundaries to form a single feature vector. The sentence is then decoded from the refined video representation. Huang et al. [22] proposed a 3DConv neural network based on attention for SLR. Wu et al. [23] proposed a semi-supervised hierarchical dynamic framework based on a Hidden Markov Model for simultaneous gesture segmentation. The high-level spatiotemporal representation is learned using this method. The skeletal dynamics were handled by a Gaussian-Bernoulli Deep Belief Network (DBN), and batches of depth and RGB images were managed and fused by 3DCNN. Yu et al. [24] deal with the segmentation of old queries.

III. Proposed System

The proposed work's general framework is depicted in Fig. 1, and a detailed flow diagram is depicted in Fig. 2. Keyframes extracted from the sign video were considered for input to the proposed model. Further, after spatial and temporal learning through CNN and LSTM, isolated words will be identified. Notations of the parameters used in this proposed work are given in Table I.

TABLE I. Parameter Notation

Symbol	Description
S_v	Similarity value between images
N	Number of signs
$F_{n'}$	Total number of frames in the video. Different for different inputs.
F_{L1}	Total key frames selected in level 1 from $F_{n'}$.
F_{L2}	Total key frames selected in level 2 from F_{L1}.
P_{F1}, P_{F2}	Pixel in frames
F_w, F_h	Frame width and height
P_a	Probability for each action

Fig. 2 presents a flowchart of the proposed model:

- key frame capturing based on similarity score for the input video]
- Spatiotemporal learning through CNN and LSTM;
- Isolated Output Prediction]
- Sentence output using the encoder-decoder network.

In our work, the vision-based dynamic recognition approach is used for real-time gesture recognition. Two subsequent phases of this method are sign language recognition and translation. Sign language recognition is performed through the one-to-one mapping between signs and isolated words. Sign language translation is performed for generating sentences from the mapping of sign and isolated words. The following sections discuss in detail the modules in the proposed work.

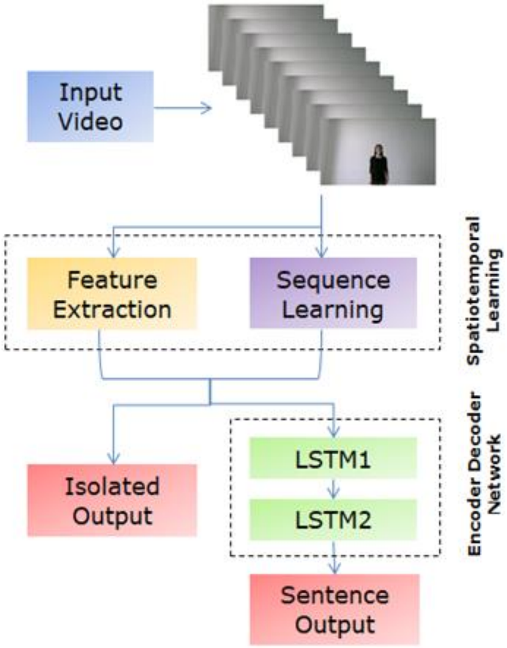

Fig. 2. Flowchart for the proposed model.

A. Key Frame Selection

As discussed in Section I, keyframes from the input video will be extracted by comparing similarity scores between images adaptively. The steps for similarity score generation for the two inputted images were detailed in Algorithm 1. The diff_ratio is computed by analyzing each pixel in the two images which plays a vital role in similarity score identification. As there will be slight changes in pixel values in the key coordinates of the subsequent frames in the sign language video, each pixel change contributes to the accurate prediction of sign gestures. Also, from the literature, it is predominantly found and concluded that different sign gestures have slight changes in the pixel value. Thus, a novel pixel-wise similarity score generation algorithm was coined with a time complexity of O(n log(n)) where n denotes the total number of pixels in the image. Based on the similarity value of one frame with another frame in the video, the number of keyframes suitable for sign language recognition will vary.

Fig. 3 depicts the adapted two-level key frame selection strategy. At level 1, frames with Sv greater than 1.5 will be selected and at level 2, Sv with values greater than 2 will be selected as keyframes.

B. Feature Extraction

Visual features are essential while working with images. Hence for the adaptively identified FL_2 keyframes, visual features will be extracted. Since CNN works best with image data, each identified keyframe will be passed into it to extract spatial information at a given time step in the input video. Features like body pose, hand position, and finger orientation will be identified for each frame.

```
Algorithm 1 : Generate_Similarity_Score
Input        : Images, $I_1$ & $I_2$
Output       : Similarity value, $Sv$
$Sv \leftarrow 0$   #Initially two images were not similar
$totDiff \leftarrow 0,\ pixelDiff \leftarrow 0, imgSize \leftarrow 0$
$H \leftarrow height(I_1)$     //finds height of image
$W \leftarrow width(I_1)$     //finds width of image
$I_2 \leftarrow resize(H, W)$
$imgSize \leftarrow H + W$
for each Pixel $px_1 \in I_1$, Pixel $px_2 \in I_2$
    a. $pixelDiff \leftarrow px_1 - px_2$
    b. $totDiff \leftarrow totDiff + pixelDiff$
$diff\_ratio \leftarrow \frac{totDiff}{imgSize}$
$start\_threshold \leftarrow 1$
$end\_threshold \leftarrow diff\_ratio\ /2$
while ($start\_threshold <= end\_threshold$)
    a. $Sv \leftarrow \frac{start\_threshold + end\_threshold}{2}$
    b. if ($diff\_ratio > Sv^2$)
        i. $start\_threshold \leftarrow Sv + 1$
        ii. $Sv\_update \leftarrow Sv$
    c. else if ($diff\_ratio < Sv^2$)
        $end\_threshold \leftarrow Sv - 1$
    d. else
            return $Sv$
return $Sv\_update$
```

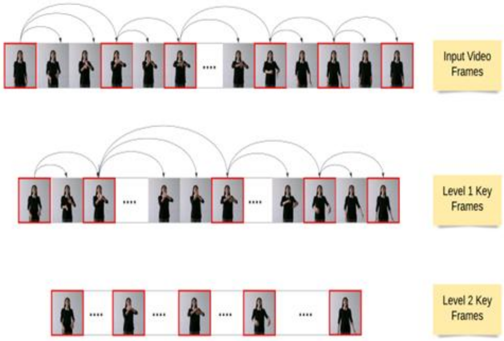

Fig. 3. Adaptive key frame selection using similarity score.

In our model kernel size of 3 X 3 is used with a 2 X 2 pool size. To reduce computational complexity and extract low-level features like body edges, a max-pooling strategy is adopted. Max pooling mathematical notation as shown in (1) where H_{wh}^{l} denotes the activation layer of l.

$$H_{wh}^{l} = max_{x=0,\dots,s,y=0,\dots,s} H_{(w+x)(h+y)}^{l-1} \tag{1}$$

C. Sequence Learning

To work with a continuous range of inputs across different time steps, temporal information through the entire input must be captured. Since RNNs are very effective in solving complex sequence-related problems, extracted visual features from CNN will be fed into a sequence of the RNN layer. In our proposed system, LSTM is the basic RNN cell. Temporal relations between subsequent frames will be captured throughout the entire video.

In our proposed work to handle spatiotemporal information as depicted in Fig. 4, all the inputs $F_1, F_2, \dots, F_t$, cell outputs C_1, hidden states Y, and gates i_t, f_t, o_t ConvLSTM are 3D tensors whose last 2 dimensions are spatial. Extracted features passed through CNN for spatial & visual feature learning. CNN followed by LSTM used for temporal learning across the video. The equations of ConvLSTM are represented in (2), where the convolution operation was denoted by '*' and the Hadamard product was denoted by '⊙'. Internal matrix multiplications are performed with convolution operations in ConvLSTM, a recurrent layer like LSTM. The data passes through the ConvLSTM cell, which maintains the input dimension of 3D until the end.

$$i_t = \sigma\,(W_{xi} * F_t + W_{hi} * Y_{t-1} + W_{ci} \odot C_{t-1} + b_i)$$

$$f_t = \sigma\,(W_{xf} * F_t + W_{hf} * Y_{t-1} + W_{cf} \odot C_{t-1} + b_f)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot tanh\,(W_{xc} * F_t + W_{hc} * Y_{t-1} + b_c)$$

$$o_t = \sigma\,(W_{xo} * F_t + W_{ho} * Y_{t-1} + W_{co} \odot C_{t-1} + b_o)$$

$$Y_t = o_t \odot tanh\,(C_t) \tag{2}$$

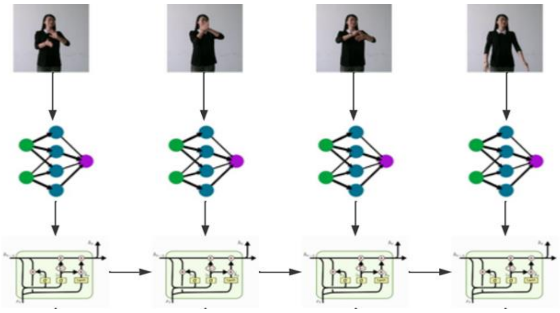

Fig. 4. Spatiotemporal learning through CNN and LSTM.

D. Isolated Output Prediction

For isolated SLR, our ConvLSTM model (see Table II) returns multiple sets of words as depicted in Fig. 5. From the obtained sequence of words, the class with the highest probability will be selected as output.

$$P_a = \sum_{i=0}^{N} \max(P_a) \tag{3}$$

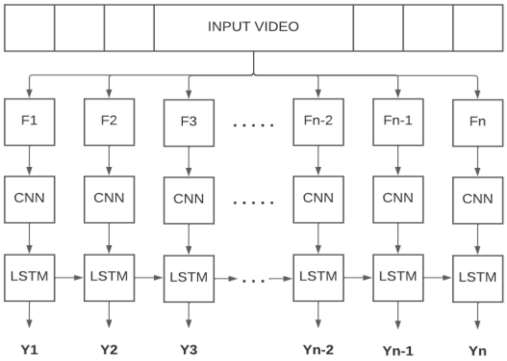

Fig. 5. Isolated SLR output prediction.

TABLE II. THE ARCHITECTURE FOR ISOLATED WORD PREDICTION USING CONVLSTM.

Layer	Kernel	Output Size
Keyframes	-	F_{L2} x 128 x 128
ConvLSTM2D	3, 3	F_{L2} x 126 x 126
MaxPooling3D	1, 2, 2	F_{L2} x 63 x 63
ConvLSTM2D	3, 3	F_{L2} x 61 x 61
MaxPooling3D	1, 2, 2	F_{L2} x 31 x 31
ConvLSTM2D	3, 3	F_{L2} x 29 x 29
MaxPooling3D	1, 2, 2	F_{L2} x15 x 15
ConvLSTM2D	3, 3	F_{L2} x 13 x 13
MaxPooling3D	1, 2, 2	F_{L2} x 7 x 7
Flatten	-	1 x 24304

E. Sentence Generation

For Continuous SLR, an encoder-decoder network is used for sentence generation. In our proposed work both encoder and decoder were developed using LSTM models. Instead, different RNNs could also be used for sentence generation. The encoder reads the input sequence and stores information in internal state vectors. The decoder's initial states are initialised from the output of the encoder. This vector triggers the decoder to start generating the output. Fig. 6 depicts the encoder-decoder architecture using LSTM (X_1 to Xn are identified keyframes) and Fig. 7 represents the sentence generation with natural language processing.

Hidden states in the encoder state are computed using (4) and in the decoder, states are computed using (5). Instead of the output layer in Isolated SLR, an encoder-decoder using LSTM is added for sentence generation.

$$h_t = f\,(W_{hh} * Y_{t-1} + W_{hx} * X_t) \tag{4}$$

$$h_t = f\,(W_{hh} * Y_{t-1}) \tag{5}$$

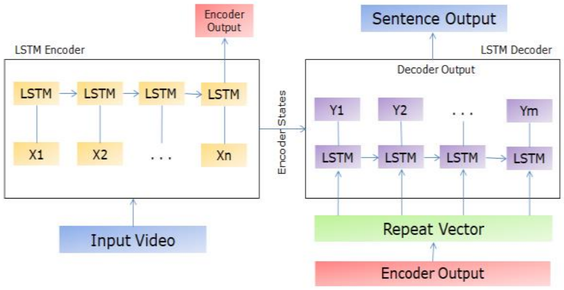

Fig. 6. Encoder decoder architecture built with LSTM for sentence generation.

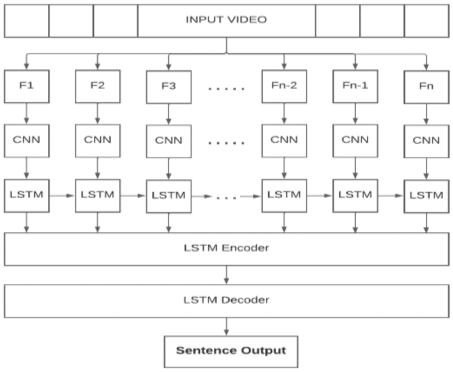

Fig. 7. Continuous SLR output prediction.

IV. EXPERIMENT

A. Dataset

The Chinese Sign Language Dataset [25-30] is used in our experiments for analysis and evaluation of the proposed approach. Isolated SLR and Continuous SLR are two types of SLR. Isolated SLR consists of 500 words, each spoken five times in sign language by 50 people. Continuous SLR consists of 100 sentences, each spoken five times by 50 signers in sign language. Each sentence contains 4 to 5 words on average. In addition, the experimental evaluation was also carried out in another notable dataset, CSL-Daily [31], which focuses on Chinese sign language, featuring 18401 training, 1077 development, and 1176 testing video samples, showcasing performances from ten signers across diverse topics like family life, medical care, and school life. CSL-Daily encompasses a gloss vocabulary of size 2000. The datasets Phoenix-2014 and Phoenix-2014T, as discussed by [32] and [33] respectively, are prominent in the field of Sign Language Recognition (SLR) in Germany. Phoenix-2014 includes 5672 training, 540 development, and 629 testing samples, with a gloss vocabulary of 1295. Conversely, Phoenix-2014T is an extension of Phoenix-2014, offering 7096 training, 519 development, and 642 testing samples, with a gloss vocabulary of 1085.

B. Model Setting for Isolated & Continuous SLR

Initially, the given input video will be split into F_n' frames i.e., $F_1 F_2, \dots F_n'$. By default, frame 1 (F_1) will be selected as one of the keyframes. Then a similarity score will be calculated between F_1 and its subsequent frames. If the Sv value between two frames is greater than 1.5 then it will be selected as a keyframe. Further, F_1 will be replaced by the chosen keyframe and the similarity value will be calculated from the immediate next frame. The loop continues till the end of the last frame. At level 1, FL_1 frames are selected as keyframes. The following Fig. 4 shows sample key frame selection in level 1.

For the selected keyframes from level 1, similarity scores will be calculated again. Sv values greater than 2 will be selected as keyframes in level 2. Finally, FL_2 frames were selected as key frames from the input video. Fig. 5 shows frames selected from level 1 with $Sv > 2$.

As shown in Fig. 3, the given input video keyframes will be identified in two levels. At level 1, frames with an Sv value less than 1.5 are not considered keyframes. Level 1 frame selection is shown in Fig. 8. At level 2 frames with an Sv value, less than 2 are not considered keyframes. Level 2 key frame selection is shown in Fig. 9.

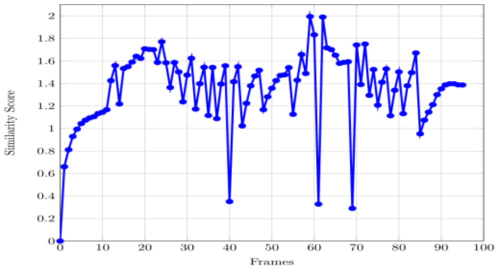

Fig. 8. Level 1 key frame selection with Sv> 1.5.

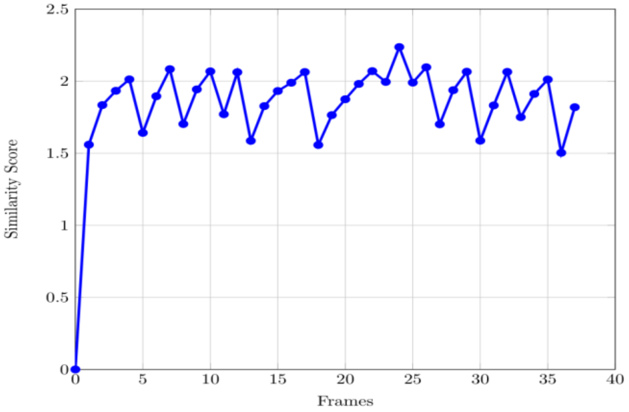

Fig. 9. Level 2 key frame selection with Sv> 2.

Followed by the adaptive key frame extraction from the input video, those frames are resized to a size of 64x64 and normalized to reduce the computational complexity. Fig. 10 shows body pose features extracted using our custom-designed lightweight CNN architecture to cope with the overall less complex design. At layers 1 and 2 outer body pose features are learned. At levels 3 & 4 features are learned vertically for effective capturing of hand and finger orientation.

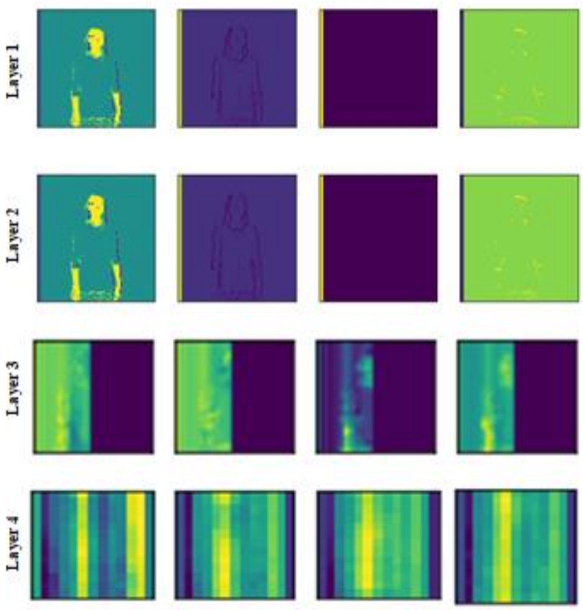

Fig. 10. Features extracted using CNN.

The proposed model for Isolated SLR consists of four ConvLSTM layers with a kernel size of 3x3 and four pooling layers with a size of 2x2. Softmax activation function was used with optimiser as Adam, loss as categorical cross-entropy, learning rate as 0.001, and batch size as 4. For Continuous SLR, an encoder-decoder network has been added with the previously trained model for isolated SLR. Both encoder-decoder networks have a kernel of size 3x3 and tanh as an activation function.

A. Model Comparison

Since our proposed model was constructed with the encoder-decoder framework, it is compared with other similar models: S2VT [34] (standard 2-layer encoder-decoder architecture), LSTM-E [35] (deep 2DCNN and 3D CNN features with mean pooling for high semantic embedding), LSTM-Attention [36] (attention mechanism to capture temporal relations), LSTM-Global-Attention [37] (global attention mechanism is explored for NMT), and HRF [16] (hierarchical recurrent deep fusion).

B. Evaluation of LSTM

In the simple LSTM approach, the extracted keyframes were directly fed into LSTM for sign language recognition. The LSTM model achieves a training accuracy of 0.54 and a testing accuracy of 0.53. Fig. 11 and Fig. 12 show the comparison of training and validation loss and training and validation accuracy for LSTM (Table III).

TABLE III. COMPARISON OF SLT (OURS) WITH ENCODER-DECODER ARCHITECTURE

Model	PRECISION
S2VT [34]	0.897
LSTM-E [35]	0.882
LSTM-Attention [36]	0.851
LSTM-global-Attention [37]	0.858
HRF-S [16]	0.924
HRF-S-att [16]	0.929
Ours	**0.932**

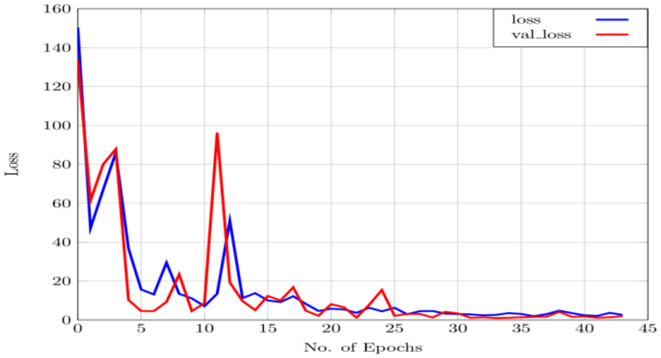

Fig. 11. Simple LSTM: Training vs. testing loss comparison.

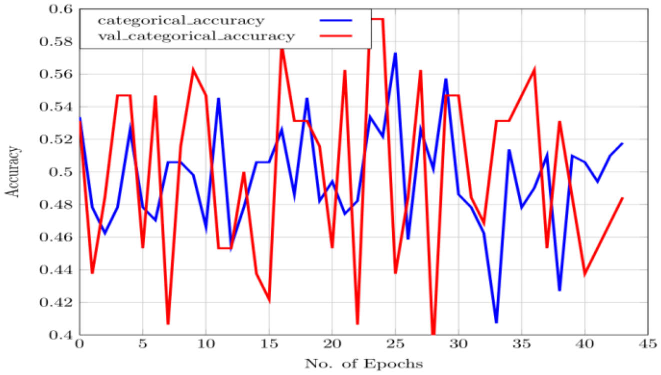

Fig. 12. Simple LSTM: Training vs. testing accuracy comparison.

C. Evaluation of GRU

Instead of LSTM, the extracted keyframes will be directly fed into GRU for sign language recognition. GRU model achieves a training accuracy of 0.56 and a testing accuracy of 0.53. Fig. 13 and Fig. 14 show the comparison of training and validation loss and training and validation accuracy for GRU.

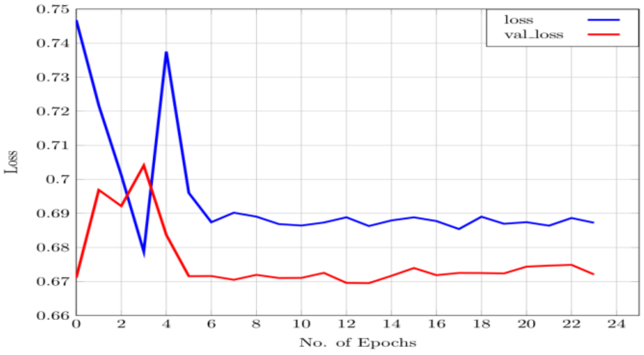

Fig. 13. Simple GRU: Training vs. testing loss comparison.

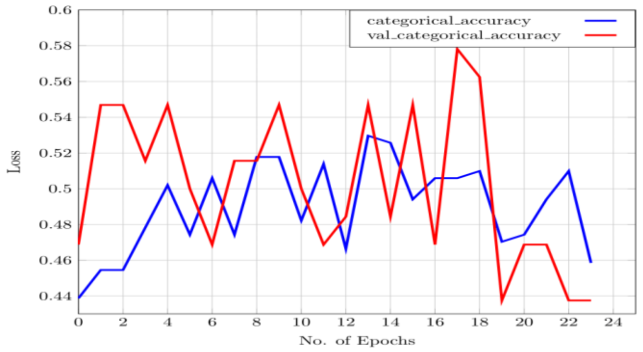

Fig. 14. Simple GRU: Training vs. testing accuracy comparison.

D. Evaluation of ConvLSTM

In ConvLSTM (Ours), extracted keyframes will be directly fed to ConvLSTM for spatial and temporal learning to recognise sign language. Compared to the previous two models ConvLSTM model achieves a training accuracy of 0.90 and a testing accuracy of 0.88. Fig. 15 and Fig. 16 show the comparison of training and validation loss and training and validation accuracy for the ConvLSTM model.

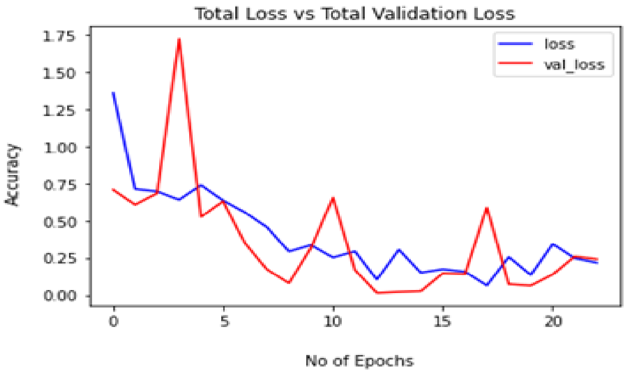

Fig. 15. ConvLSTM Model: Training vs. testing loss comparison.

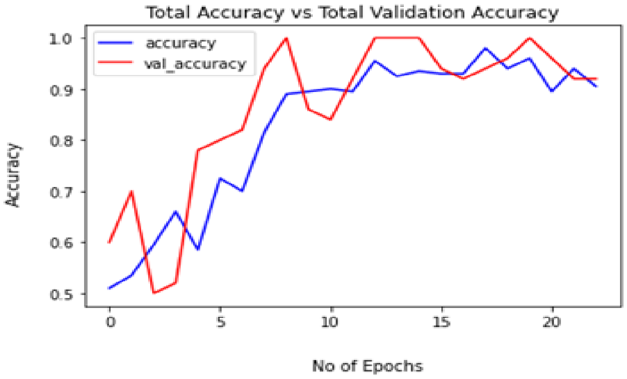

Fig. 16. ConvLSTM Model: Training vs. testing accuracy comparison.

E. Comparison of LSTM Vs. GRU Vs. Convlstm

Comparing all the three models ConvLSTM gives better performance since it captures both spatial and temporal information. The ConvLSTM model achieves an accuracy of 0.90 while the other two approaches produce 0.54 and 0.56 for LSTM and GRU respectively and it is illustrated in Table IV.

TABLE IV. COMPARISON OF SLT (OURS) WITH ENCODER-DECODER ARCHITECTURE ON DIFFERENT EVALUATION METRICS

Model	Training		Testing	
	Accuracy	Loss	Accuracy	Loss
LSTM	0.54	0.62	0.53	0.41
GRU	0.56	0.64	0.53	0.48
Ours	**0.90**	**0.21**	**0.88**	**0.26**

F. Comparison with State-of-the-ART Works

Table V shows the evaluation of different metrics in the CSL dataset.

TABLE V. THE EVALUATION OF DIFFERENT METRICS IN THE CSL DATASET

Model	PRECISION	BLEU	METEOR	ROUGE-L
S2VT [34]	0.897	0.902	0.642	0.904
HRF-S [16]	0.924	0.942	0.699	0.944
HRF-S-att [16]	0.929	0.948	0.703	0.951
Ours	**0.932**	**0.949**	**0.710**	**0.951**

In this section, the detailed analysis is carried out with other datasets and it is given in Table VI, Table VII and Table VIII.

TABLE VI. SOTA FOR BLEU-4 AND ROUGE ON PHOENIX-2014T BENCHMARK

Method	Dev		Test	
	BLEU-4	ROUGE	BLEU-4	ROUGE
PT [38]	11.82	33.18	10.51	32.46
AT [39]	12.65	33.68	10.81	32.74
MDN [40]	11.54	33.40	11.68	33.19
MoMP [41]	14.03	37.76	13.30	36.77
FS-NET [42]	16.92	35.74	21.10	42.57
SignDiff [43]	18.26	39.62	22.15	46.82
Ours	**27.93**	**52.81**	**29.25**	**54.58**

TABLE VII. SOTA FOR WER ON PHOENIX-2014 AND PHOENIX-2014T

Method	**Phoenix-2014**		**Phoenix-2014T**	
	Dev (%) WER	**Test (%) WER**	**Dev (%) WER**	**Test (%) WER**
SubUNets [44]	40.8	40.7	-	-
CNN-LSTM-HMMs [45]	26.0	26.0	24.1	26.1
FCN [46]	23.7	23.9	23.3	25.1
Joint-SLRT [47]	-	-	24.6	24.5
SignBT [31]	-	-	22.7	23.9
Two-Stream-SLR [48]	18.4	18.8	17.7	19.3
CorrNet [49]	18.8	19.4	18.9	20.5
CVT-SLR [50]	19.8	20.1	19.4	20.3
SEN [51]	19.5	21.0	19.3	20.7
Ours	**16.8**	**16.2**	**15.9**	**15.7**

From the experimental evaluation it is found that by incorporating a novel adaptive key frame extraction technique, there is a significant improvement in the BLEU-4 and ROUGE score on the diversified datasets of consideration. Also, the noticeable decrease in the Word Error Rate by around 2% indicates that the proposed system to extract keyframes for continuous sign language in video outperforms SOTA systems.

TABLE VIII. SOTA FOR WER ON THE CSL-DAILY DATASET

Method	**WER**	
	Dev (%)	**Test (%)**
SubUNets [44]	41.1	41.0
FCN [46]	39.0	39/4
Joint-SLRT [47]	33.1	32.0
SignBT [31]	33.2	33.2
Two-Stream-SLR [48]	25.4	25.3
CorrNet [49]	30.6	30.1
SEN [51]	31.1	30.7
Ours	**19.3**	**19.1**

V. CONCLUSION

The paper has proposed multimodal output from continuous sign language using ConvLSTM with adaptive frame selection that achieves an accuracy of 90% with **a** precision rate of 0.93. The proposed network provides better performance by capturing spatial and temporal information in the video through CNN and LSTM. The experimental discussion shows that our proposed model performs well for Isolated and Continuous SLR. However, the Continuous SLR still has some issues due to word order mapping with signs, and sentences with an average of 4-5 words are only taken. As a future work, it is planned to enhance the performance of sign-word mapping in the sentence, particularly augmenting the sign language-based education system with the outcomes presented in this study. The analysis and evaluation will be experimented with over bigger sentences with long sign actions to test the robustness of the proposed model.

ACKNOWLEDGMENT

The paper is financed by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project № BG-RRP-2.004-0001-C01.

REFERENCES

[1] L. Humes, “The World Health Organization’s hearing-impairment grading system: an evaluation for unaided communication in age-related hearing loss”. International Journal of Audiology, 58(1), pp.12-20, 2019.

[2] G. Haidar, H. Reefat, “Glove-Based American Sign Language Interpretation Using Convolutional Neural Network and Data Glass”. In 2020 IEEE Region 10 Symposium (TENSYMP) (pp. 370-373). IEEE, 2020.

[3] A. Jalali, M. Lee. “High cursive traditional Asian character recognition using integrated adaptive constraints in the ensemble of DenseNet and Inception models”. Pattern Recognition Letters, 131, pp. 172-177, 2020.

[4] P. Kumar, H. Gauba, P. Roy, D. Dogra. “Coupled HMM-based multi-sensor data fusion for sign language recognition”. Pattern Recognition Letters, 86, pp. 1-8, 2017.

[5] R. Nihal, S. Rahman, N. Broti, S. Deowan. “Bangla Sign alphabet recognition with zero-shot and transfer learning”. Pattern Recognition Letters, 150, pp. 84-93, 2021.

[6] M. Zbakh, Z. Haddad, J. Krahe. “An online reversed French Sign Language dictionary based on a learning approach for signs classification”. Pattern Recognition Letters, 67, pp. 28-38, 2015.

[7] B. Atitallah, Z. Hu, D. Bouchaala, M. Hussain, A. Ismail, N. Derbel, and O. Kanoun. “Hand sign recognition system based on EIT imaging and robust CNN classification”. IEEE Sensors Journal, 22(2), pp.1729-1737, 2022.

[8] A. Mittal, P. Kumar, P. P. Roy, R. Balasubramanian and B. B. Chaudhuri. “A Modified LSTM Model for Continuous Sign Language Recognition Using Leap Motion”. IEEE Sensors Journal, 19(16), pp. 7056-7063, 2021.

[9] M. Deriche, S. O. Aliyu and M. Mohandes. “An Intelligent Arabic Sign Language Recognition System Using a Pair of LMCs With GMM-Based Classification”. IEEE Sensors Journal, 19(18), pp. 8067-8078, 2019.

[10] G. Marin, F. Dominio and P. Zanuttigh. “Hand gesture recognition with leap motion and Kinect devices”. In 2014 IEEE International Conference on Image Processing (ICIP) (pp. 1565-1569). IEEE, 2014.

[11] B. G. Lee and S. M. Lee. “Smart Wearable Hand Device for Sign Language Interpretation System With Sensors Fusion”. IEEE Sensors Journal, 18(3), pp. 1224-1232, 2018.

[12] X. Huang, Q. Wang, S. Zang, J. Wan, G. Yang, Y. Huang and X. Ren. “Tracing the Motion of Finger Joints for Gesture Recognition via Sewing RGO-Coated Fibers Onto a Textile Glove”. IEEE Sensors Journal, 19(20), pp. 9504-9511, 2019.

[13] T. Dias, J. Júnior and S. Pichorim. “An Instrumented Glove for Recognition of Brazilian Sign Language Alphabet”. IEEE Sensors Journal, 22(3), pp. 2518-2529, 2022.

[14] B. Li, J. Yang, Y. Yang, C. Li and Y. Zhang. “Sign Language/Gesture Recognition Based on Cumulative Distribution Density Features Using UWB Radar”. IEEE Transactions on Instrumentation and Measurement, 70, pp. 1-13, 2021.

[15] S. Gurbuz, A. Gurbuz, E. Malaia, D. Griffin, C. Crawford, M. Rahman, E. Kurtoglu, R. Aksu, T. Macks, R. Mdrafi. “American sign language recognition using RF sensing”. IEEE Sensors Journal, 21(3), pp.3763-3775, 2020.

[16] D. Guo, W. Zhou, A. Li, H. Li and M. Wang. “Hierarchical Recurrent Deep Fusion Using Adaptive Clip Summarization for Sign Language Translation”. IEEE Transactions on Image Processing, 29, pp. 1575-1590, 2019.

[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, 24, pp. 768-779, 2021.

[18] D. S. Breland, A. Dayal, A. Jha, P. K. Yalavarthy, O. J. Pandey and L. R. Cenkeramaddi. “Robust Hand Gestures Recognition Using a Deep CNN and Thermal Images”. IEEE Sensors Journal, 21(23), pp. 26602-26614, 2021

[19] W. L. Passos, G. M. Araujo, J. N. Gois and A. A. de Lima. “A Gait Energy Image-Based System for Brazilian Sign Language Recognition”. IEEE Transactions on Circuits and Systems I: Regular Papers, 68(11), pp. 4761-4771, 2021

[20] S. Huang, C. Mao, J. Tao and Z. Ye. “A Novel Chinese Sign Language Recognition Method Based on Keyframe-Centered Clips”. IEEE Signal Processing Letters, 25(3), pp. 442-446, 2018.

[21] C. Wei, J. Zhao, W. Zhou and H. Li. “Semantic Boundary Detection With Reinforcement Learning for Continuous Sign Language Recognition”. IEEE Transactions on Circuits and Systems for Video Technology, 31(3), pp. 1138-1149, 2020.

[22] J. Huang, W. Zhou, H. Li and W. Li. “Attention-Based 3D-CNNs for Large-Vocabulary Sign Language Recognition”. IEEE Transactions on Circuits and Systems for Video Technology, 29(9), pp. 2822-2832, 2018.

[23] D. Wu, L. Pigou, P. Kindermans, N. Le, L. Shao, J. Dambre and J. Odobez. “Deep Dynamic Neural Networks for Multimodal Gesture Segmentation and Recognition”. IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(8), pp. 1583-1597, 2016.

[24] R. Yu, C. Tian, W. Xia, X. Zhao, L. Wang, Y. Yang. “Real-time human-centric segmentation for complex video scenes”. Image and Vision Computing. 126. p.104552, 2022.

[25] J. Pu, W. Zhou, and H. Li. “Iterative Alignment Network for Continuous Sign Language Recognition”. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4165-4174), 2019.

[26] H. Zhou, W. Zhou, and H. Li. “Dynamic Pseudo Label Decoding for Continuous Sign Language Recognition”. In 2019 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1282-1287), 2019.

[27] J. Huang, W. Zhou, Q. Zhang, H. Li and W. Li. “Video-based Sign Language Recognition without Temporal Segmentation”. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 32, No. 1), 2018.

[28] J. Pu, W. Zhou, H. Hu, and H. Li. “Boosting Continuous Sign Language Recognition via Cross Modality Augmentation”. In Proceedings of the 28th ACM International Conference on Multimedia (pp. 1497-1505), 2020.

[29] J. Pu, W. Zhou, and H. Li. “Dilated Convolutional Network with Iterative Optimization for Continuous Sign Language Recognition”. In IJCAI (Vol. 3, p. 7), 2018.

[30] D. Guo, W. Zhou, M. Wang, and H. Li. “Hierarchical LSTM for Sign Language Translation”. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1), 2018.

[31] H. Zhou, W. Hou, W. Qi, J. Pu, and H. Li. “Improving sign language translation with monolingual data by sign back-translation”. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1316–1325), 2021.

[32] 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, 141, pp. 108–125, 2015

[33] N. Camgoz, S. Hadfield, O. Koller, H. Ney, R. Bowden. “Neural sign language translation”. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 7784-7793), 2018.

[34] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko. “Sequence to sequence—Video to text”. In Proceedings of the IEEE International Conference on Computer Vision (pp. 4534–4542), 2015.

[35] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui. “Jointly modelling embedding and translation to bridge video and language”. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4594-4602), 2016.

[36] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. “Describing videos by exploiting temporal structure”. In Proceedings of the IEEE International Conference on Computer Vision (pp.4507-4515), 2015.

[37] T. Luong, H. Pham, and C. D. Manning. “Effective approaches to attention-based neural machine translation”. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1412–1421, 2015.

[38] B. Saunders, N. Camgoz, and R. Bowden. “Progressive transformers for end-to-end sign language production”. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16 (pp. 687-705). Springer International Publishing, 2020.

[39] B. Saunders, N. Camgoz, and R. Bowden. “Adversarial training for multi-channel sign language production”. arXiv preprint arXiv:2008.12405, 2020.

[40] B. Saunders, N. Camgoz, and R. Bowden. “Continuous 3d multi-channel sign language production via progressive transformers and mixture density networks”. International Journal of Computer Vision, 129(7), pp. 2113-2135, 2021.

[41] B. Saunders, N. Camgoz, and R. Bowden. “Mixed signals: Sign language production via a mixture of motion primitives”. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1919-1929), 2021.

[42] B. Saunders, N. Camgoz, and R. Bowden. “Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production”. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5141-5151), 2022.

[43] S. Fang, C. Sui, X. Zhang, Y. Tian. “SignDiff: Learning Diffusion Models for American Sign Language Production”. arXiv preprint arXiv:2308.16082, 2023.

[44] O. Koller, S. Zargaran, H. Ney. “Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs”. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 4297-4305), 2017.

[45] O. Koller, N. 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, 42(9), pp.2306-2320, 2019.

[46] K. Cheng, Z. Yang, Q. Chen, and Y. Tai. “Fully convolutional networks for continuous sign language recognition”. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16 (pp. 697–714), Springer, 2020

[47] N. Camgoz, O. Koller, S. Hadfield, and R. Bowden. “Sign language transformers: Joint end-to-end sign language recognition and translation”. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10023–10033), 2020.

[48] Y. Chen, R. Zuo, F. Wei, Y. Wu, S. Liu, and B. Mak. “Two-stream network for sign language recognition and translation”. Advances in Neural Information Processing Systems, 35, pp. 17043–17056, 2022

[49] L. Hu, L. Gao, Z. Liu, and W. Feng. “Continuous sign language recognition with correlation network”. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp.2529–2539), 2023

[50] J. Zheng, Y. Wang, C. Tan, S. Li, G. Wang, J. Xia, Y. Chen, and S. Li. “Cvt-SLR: Contrastive visual textual transformation for sign language recognition with variational alignment”. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 23141–23150), 2023

[51] L. Hu, L. Gao, Z. Liu, and W. Feng. “Self-emphasizing network for continuous sign language recognition”. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, pp. 854–862, 2023