Transforming Sign Language Recognition: Leveraging Deep Learning for Improved Accuracy and Real-Time Translation

by

Eshtiak Alam Shihab
21301502
Md Shamsur Shafi Nur E Aziz
21301432
Kazi Israrul Karim
21301509
Tashfia Saad
21301320
Neelavro Shafin Qais
21301501

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
June 2024

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Shafi

Eshtiak Alam Shihab
21301502

Md Shamsur Shafi Nur E Aziz
21301432

ISRAR

Saad.

Kazi Israrul Karim
21301509

Tashfia Saad
21301320

Neelavro Shafin Qais
21301501

Approval

The thesis/project titled "Transforming Sign Language Recognition: Leveraging Deep Learning for Improved Accuracy and Real-Time Translation" submitted by

1. Eshtiak Alam Shihab(21301502)
2. Md Shamsur Shafi Nur E Aziz(21301432)
3. Kazi Israrul Karim(21301509)
4. Tashfia Saad(21301320)
5. Neelavro Shafin Qais(21301501)

Of Summer, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on August 23, 2015.

Examining Committee:

Supervisor:
(Member)

Dr. Farig Yousuf Sadeque
Assistant Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

The need for effective sign language recognition and translation has become more critical to create a more inclusive society that addresses the communication concerns of the Deaf community. In recent years, the field has seen a revolutionary progress arc, spearheaded by the development of transformative Deep Learning based approaches such as reinforcement learning, spatio-temporal residual networks, temporal convolution modules, iterative alignment networks, and attention mechanisms. Yet, vision-based real time continuous sign language recognition (CSLR) continues to face several application challenges, encompassing its visual, sequential, and alignment modules. As such, we propose an end-to-end training model inspired by the recent successes of transfer learning and attention-based mechanisms in particular to achieve new state-of-the-art performance on current benchmarks. Our paper also includes a comprehensive comparison of our model with existing ones to portray its effectiveness based on standard evaluation metrics.

Keywords: Sign Language Recognition; Continuous Sign Language Recognition; Deep Learning; Connectionist Temporal Classification; Spatio-temporal Residual Networks; Attention-based Mechanisms; Transformers

Table of Contents

List of Figures

Chapter 1

Introduction

Sign Language is the fundamental mode of communication for around 6.1% of the entire population of the earth, the Deaf community, according to the World Health Organization. There exists regional variations in dialect and certain linguistic aspects within the sign language communities owing to the unique characteristics of American, Bengali, German, Chinese, Russian and other sign languages. These languages not only exist as a linguistic tool, but also as a medium which allows people with hearing and speech disabilities to carry out regular everyday activities such as going to work, providing service to others, expressing their needs in required times, etc. From individuals who are born with hearing loss to those who lost their hearing midway through their lives, Sign Language provides the opportunity to remain an active and interactive individual of the community.

However, unlike spoken language, sign language expressions incorporate a manual component such as hand shape, pose etc alongside non-manual ones pertaining to facial expressions. It also comes with its own set of vocabulary, grammar and linguistic nuances. As such, there exists a communication gap between a signer and a speaking-hearing person which can be bridged effectively by Automatic Sign Language Recognition (Hao et al., 2021). The sensor-based approach to achieve this has been increasingly replaced by vision-based ones in recent years. Subcategories of the latter involve Isolated Sign Language Recognition (ISLR) which aims at identifying the gloss illustrated by a specific video segment. In contrast, Continuous Sign Language Recognition (CSLR) takes on the more complex challenge of predicting an appropriate gloss sequence from a sign language video.

Despite remarkable outcomes, CSLR is still undergoing progressive improvements. The traditional approaches involved hand-crafted feature extraction followed by naive Machine Learning models such as Support Vector Machines. The standardised approach today involves more robust frameworks that utilise Deep Learning for automatic feature extraction. The functionality of these frameworks is distributed among 3 modules: visual, sequential and alignment. The purpose of the visual module is to extract visual cues from video frames. The sequential model then uses the extracted features to mine correlation between glosses (Niu & Mak, 2020). Eventually, the alignment module maps the correct video frames and gloss sequences. An objective function of Connectionist Temporal Loss (CTC) is usually associated with these three modules to leverage its ability of automatic alignment to allow for end-to-end training of models. The initial models relied on Hidden Markov Model (HMM) but have been replaced by Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) architectures lately. Long Short-Term Memory (LSTM) is also integrated into most frameworks. The resultant models have achieved great success on frequently used large benchmark datasets such as the PHOENIX-2014, PHOENIX-2014-T, and the CSL datasets. Our goal is to contribute to the same cause by improving a Sign Language recognition and translation system with high accuracy and addressing some barriers that continue to pose challenges.

The following sections of our paper encompass our research statement and objectives. Figure 1.1 presents an overview of the different approaches to Sign Language Recognition (SLR), categorizing them into isolated sign language recognition (ISLR) and continuous sign language recognition (CSLR). The diagram further breaks down CSLR into sub-methods, showcasing the evolution from traditional methods to deep learning models. Then, Figure 1.2 illustrates the pipeline for continuous sign language recognition and translation, detailing the process from input to output. The visual module captures and processes the sign language input, followed by the alignment module, and culminating in the translation module that converts the signs into English sentences. These are followed by a deep-dive into existing works in this field, particularly in the past 3 to 5 years. We explore the prominent breakthroughs which influence the research landscapes of isolated and continuous sign language recognition, focusing more on the latter. We categorically look into recent studies trying to resolve the more challenging sub-domains within CSLR: feature extraction and gloss segmentation, alignment, translation, multilinguality and dataset. Finally, we end with our tentative work plan and a concluding remark on the work conducted to achieve accurate and representative CSLR so far.

1.1 Aims and Objectives

The research addresses the current challenges in the evolving landscape of Continuous Sign Language Recognition by analysing the latest developments and progress in the field, ensuring a comprehensive understanding. As such, the research objective aims to enhance Real-Time Sign Language Recognition by leveraging advancements in deep learning models and transformers, thereby facilitating seamless communication between the deaf and hard of hearing communities.

1.2 Research Objectives

1.2.1 Data Collection and Preprocessing

- Collect datasets which cover the sentence-level video structure of Sign Language Recognition (SLR), as they contain continuous data suitable for both isolated and continuous sign language recognition.

- Apply standard video and image preprocessing techniques such as noise reduction, segmentation, and data augmentation for accurate video fragmentation and feature extraction.

1.2.2 Model Development and Implementation

- Implement state-of-the-art machine learning techniques and deep learning models to extract relevant frames and glosses from video input.

- Employ an advanced sequence-to-sequence (seq2seq) modeling technique for translation.

- Enhance an existing framework to serve as a real-time sign language interpreter.

1.2.3 Performance Evaluation

- Evaluate the performance and effectiveness of the proposed models and methods using appropriate metrics and methodology.

- Compare the performance and results of our model with other current state-of-the-art models.

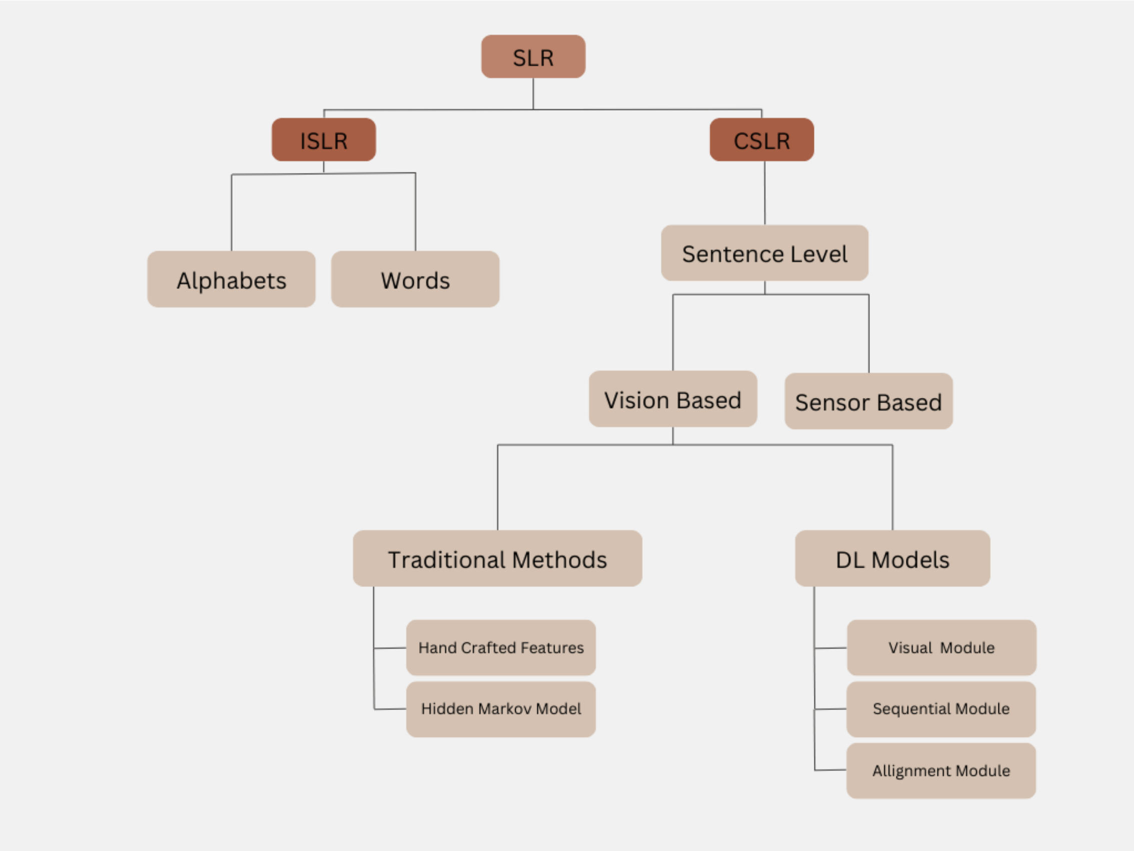

Figure 1.1: An overview of different approaches to Sign Language Recognition (SLR)

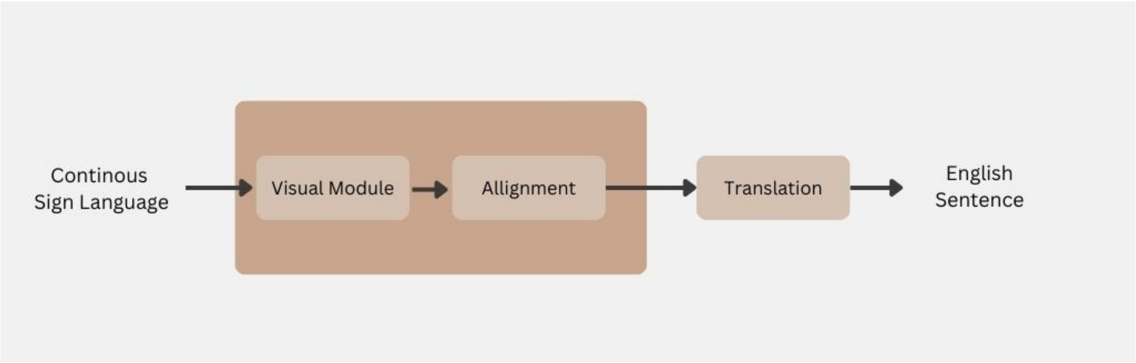

Figure 1.2: Pipeline for Continuous Sign Language Recognition

Chapter 2

Literature Review

2.1 Survey Methodology

We thoroughly read many technical research articles related to Sign Language Recognition. We began our research by first understanding the complex nature of sign language and how they differ from our spoken languages. Then we moved on to the latest review papers to understand the challenges faced in this field. Finally, we narrowed down our research to understand how different state-of-the-art models are working on these challenges and developing more effective and accurate Sign Language Recognition models. Furthermore, our paper selection method was to consider relatable papers with large numbers of citations excluding very similar research. We highly focused on recent publications as well. By systematically categorising these selected papers according to the stages of CSLR, we can construct a coherent and comprehensive body of knowledge that will underpin our future research endeavours.

2.2 Analysis of Sign Language

To begin with, Sign Language recognition and translation is an extensive research field, with the primary goal - to accurately curate translations of Sign Language, which has been approached in various ways by different researchers. Achieving this goal comprehensively by addressing every single aspect has proved to be a significant challenge, even with modern cutting-edge technology. Hence, several different studies have each tackled one or numerous facets of the broader problem.

The review paper by Manoharan and Roy (2022) mentions the branches into which SLR can be divided and summarises the advancements done in each sub-field between 2009 and 2021. It also discusses the major limitations of these models and approaches. In almost every research there are six major features with respect to SLR - firstly, SLR Translation can either be isolated alphabet/word recognisable for which an isolated dataset has to be used, or it can be continuous SLR Translation for which continuous dataset has to be used. Secondly, the user input can be obtained either via camera or likewise components, i.e vision related hardware or via sensory gloves, i.e sensor based hardware. Next, there are two kinds of features that can be processed for the recognition and translation, manual markers (i.e hand gestures) and non-manual markers (i.e eyebrows, nose, chin, mouth, nose, cheek, etc.). Usually, manual markers are used for the translation itself, while non-manual markers are used for a more comprehensive and contextual analysis behind the translation. To research on these aspects, many different models like deep learning models, traditional methods, hybrid methods can be implemented. However, their application comes with challenges. Other than the gesture recognition and translation process being complex, there is the pressing issue of context. Comprehensive understanding can only be achieved by considering the several non-manual markers simultaneously with the gestures. Some other significant barriers include scaling and orienting the image properly in case of different signers with variable physical attributes, light intensity, and the level of interference by non-uniform and dynamic backgrounds. These issues can be found in both vision and sensor-based input methods. From this paper, we also get an outline of the Isolated Manual SLR research during the period. Chronologically, the challenges faced by earlier papers have been addressed and resolved over the years. As such, in the last few years, with model accuracy reaching up to 99.93%, the research demands have shifted away from Isolated SLR. Instead, researchers have been intrigued by the fact that accuracy obtained by recent papers have not met the highest standards when it comes to Isolated Non-Manual SLR work, reaching a maximum of 78%. Even when trained by Machine Learning (ML) or Deep Learning (DL) models, there remain significant barriers concerning visual sensitivity, and insufficient dataset and real-world experimental testing. Inadequate real-time accurate classification and translation, and heavy computational time and resource requirements also present challenges. Research overview on continuous SLR shows that even accuracy involving manual works are as high as 98% in recent papers. Hence, there remains very limited opportunities to contribute any further improvements here too. On the contrary, for Continuous Non-Manual SLR, the accuracy obtained using recent technology are comparatively much lower, reaching a highest of only 28%. Older methodologies have obtained 92% accuracy, but failed to address some crucial classification issues that have surfaced later. Analysing the statistics have revealed that even with DL models, sufficient continuous datasets are extremely limited. As such, several models perform well with the training dataset but do not perform similarly on other ones (overfitting). Other barriers include high training time and complexity, sensitivity to overlapping during classification and unalignment problems.

It is evident that there has already been extensive research on both isolated and continuous manual marker dependent research for SLR (Manoharan & Roy, 2022). In their study, Wali et al. (2023) delves deeper into some of the aforementioned branches within this field. They examine the primary obstacles in Sign language recognition (SLR) that are associated with written text units, including letters, words, and sentences. Every unit poses individual challenges and necessitates specific strategies. At the letter level, SLR addresses challenges involving the recognition of characters, shapes, or objects. Pre-trained models like VGG-16, have demonstrated commendable outcomes by utilising transfer learning. Ultimately, the sole prevailing issue in letter-level pertains to the identification of shapes in dynamic environments. Word-level spoken language recognition can be categorised into two main types: static recognition and sequential recognition. Static word recognition is identifying individual signs that are captured in a single frame. In contrast, sequential word recognition involves temporal dimension, adding complexity to the task. Wali et al.

(2023) recognize that identifying words in continuous signing sequences is particularly strenuous as there are no noticeable breaks between signs. Consequently, this level presents various obstacles, such as the identification of hands, classification of sequences with multiple inputs, and the presence of changing backgrounds. Finally, for sentence-level or Continuous Sign Language Recognition (CSLR), the most complex amongst the challenges, a general three-step dissection procedure is presented: converting signed videos into gloss alignment, recognizing individual glosses (the basic lexical units in sign language), and translating these glosses into coherent sentences. This mechanism faces several barriers such as consecutive signs having no visible pauses, making boundary detection difficult, challenges of sequence-to-sequence (seq2seq) modelling, and gloss segmentation. Then, the study suggests that while Transformers show great potential, they typically require extensive datasets to be effective. Therefore, further research is recommended to identify the most suitable pre-trained Transformer models for SLR and to explore how these models can be adapted to achieve improved results. Reiterating the premise established in the previous paper about the significance of non-manual markers in SLR, the need for incorporating facial expressions into the segmentation module has been highlighted for future research here as well. Ultimately, the fundamental problem in current CSLR research has been the lack of cohesive studies building upon one another to refine methods and approaches. However, as we will be seeing in our paper later this landscape is changing rapidly in the last three years. In conclusion, while significant progress has been made in SLR, the field faces ongoing challenges related to segmentation, alignment, and the integration of non-manual signals. Cohesive research efforts and the exploration of advanced models like transformers are essential for future advancements.

2.2.1 Isolated Sign Language Recognition (ISLR)

One of the most highlighted works in the field of ISLR was published by Joshi et al. (2017). Their paper consists of some of the most fundamentally relevant and simple techniques to translate ASL gestures taken via user video input to produce both English characters and English words using manual markers entirely. The authors address the problem in the massive gap between the communication of individuals from the Deaf community and the rest of the world by creating a medium that accepts inputs from generally used devices such as computers and other generic webcams. After ensuring that the model is user friendly by not requiring complicated input hardware, the authors create a software which analyses the video input frame by frame and compares each frame's images to their dataset, the characters or words are formulated. Finally, they formulate the whole word by applying a gesture recognition method. This involves calculating the cross-correlation between the frames to eventually deduce the full word. For their research, the authors create their own dataset which consists of two parts. Firstly, the Alphabet Database (consisting of ASL A-Z images) which is created with the help of their aforementioned sign image acquisition software. The software takes image inputs of the signs for each alphabet, then uses image segmentation to convert the image inputs from RGB to binary (black and white). This helps eliminate any interference from the surrounding background. To further remove noise, they carry out morphological filtering. The original image input is also passed through edge detection algorithm and the two images (image with edge detection and morphologically filtered image) are combined to be finally stored as a comparison model to be used for a particular alphabet. They create 10 such samples for each alphabet. Secondly, to curate the ASL Word Database, they record video input, on which the same steps as Alphabet Database is carried out and then every frame is matched with the alphabets and a cross-correlation matrix is used to find the similarity between the two images to come to a conclusion. Instead of using any particular Machine Learning model, the authors carry out the translation procedure using their software. For Alphabet translation, again image segmentation, morphological filtering and edge detection is carried out on the input image and finally correlation matrix used to come to an identifiable conclusion. For word/sentence translation, the video input is divided into frames and for each frame, the same steps as before are carried out to give an output of the sentence. For alphabet translation and word translation, the accuracy obtained is 94.23% and 92%, respectively.

Amongst later attempts at recognising signs as independent entities, Matlani et al. (2022) achieved some interesting results. Their paper acknowledges two methods for sign language recognition: from images or by using sensors to track the data. It proposes a method based on machine learning and neural networks for real-time sign language recognition followed by converting it into speech or text and vice versa.Matlani et al. (2022) observes that the performance of RNN, CNN, ANN and HMM models is compromised in the absence of large datasets. Hence, they opt for a different approach by using a google open source framework, Mediapipe, to recognise human body parts. For the dataset, they use Indian Sign Language (ISL) and ASL, consisting of signed images of each character in the English alphabet and numbers from 0-9. The input video is captured using either a webcam or via real time video capturing which contains hand signs. The hand trajectories are later tracked using google's library mediapipe where mediapipe found out coordinates associated with the sign captured. These coordinates, referred to as key points, then undergo a 2-stage pre-processing for training purposes. Consequently, data files are separated into training and test sets. Finally, DT, KNN, Random Forest, SVM and XGBoost were used for training. A variety of regression models such as Linear Regression, SVR and XGBoost regressor were also implemented and the paper recommends using a light NN for eventually matching the signs with appropriate letters. It also states that this system is resilient and cost effective because it achieves its goal of ISLR without using any smart sensors or expensive computational power. Moreover, the research reveals that Mediapipe can be effectively used to recognise complex hand signs exactly with very high accuracy, ranging between 78% to 100%.

More recently, Kezar, Thomason, et al. (2023) have provided a transformative novel approach of improving the accuracy of ISR by leveraging the use of phonological features of sign language such as identifying the minor locations (chin), the shape of the hand, and path movement of the hand through space (circular) etc. For the dataset, they combined the phonological annotation in ASL-LEX 2.0 with the signs in WLASL 2000 ISLR benchmark, in total adding 16 new columns each indicating different phonemes. A graph convolutional network based on SL-GCN and a Transformer based model was used.The SL-GCN uses spatial and temporal attention to learn the different pose- estimation. Then it is passed on to the transformer which uses 5-layers to learn the sequences of the coordinates with respect to time. Using these models helps to capture the phoneme even if it is for a short amount of time. They modified the decoder to classifier by implementing n fully-connected layers as it not only apprehends the targeted gloss but also the phonemes. Furthermore, the model is trained until it improves the validation accuracy in the last 30 epochs.Training the 16 labelled phonemes was not cost efficient so to further improve the model they used an utility function:

$$P^*(n) = \left\{ \arg\max_P U(P) : |P| = n \right\}$$

The models were trained with their auxiliary loss for $P^*(n)$, $n \in \{2, 5, 9, 16\}$. It was concluded that handshapes and minor location were the most significant phonemes as they improved the result the most. Evaluation metrics used were top-1, top-3 and MRR (mean reciprocal rank). The model shows a 3-9% gain in top-1, 5-11% gain in top-3 and 0.04-0.09 gain on MRR. However there were limitations as the existing dataset had incorrect labels, information of the signers like age, race, fluency is also missing. To conclude, the use of phonemes during training enables the models to learn a more holistic approach to ISLR rather than just learning the target gloss, overall showing a significant improvement in the results.

On the other hand, Jamwal et al. (2022) explores a new avenue of ISLR by incorporating emotional analysis using ML models into their research on sign language recognition and translation. Previously, as evident from our investigations, the key association of non-manual markers with the emotions expressed through signs had been well established. Hence, this paper uses non-manual markers namely eyes, eyebrows, nose tip, lips to classify the sentiment. For training the models, Jamwal et al. (2022) use a static hand sign model dataset whose details have not been included. For the gesture translation itself, the authors used the MediaPipe Regression model that was trained with over 30,000 real world images. Eventually, the gesture has been split into five categories - angry, happy, sad, surprise and neutral. They use the Support Vector Machine (SVM) model and a custom Convolutional Neural Network (CNN) model alongside the MediaPipe to obtain a 0.77 recall and F1-score, and a 0.81 average score. The paper reflects upon a heavy dependency on the CNN model for both translation and emotion classification. Additionally, it should be noted that a major limitation of this approach was that the emotional recognition was solely independent of the translation. So for instance, if a positive statement was signed with a sad facial expression, the system would give the translation with a sad emotion, which is contradictory. Regardless, Jamwal et al. (2022) achieves significant strides in integrating emotion analysis with sign language recognition. In another research, Jalaja and Shekar (2022) build a two-way communication platform for the Deaf community aimed at sentiment analysis based on the text context. Their work is of significance since it includes a class of Deaf people who may not be literate to understand the English language. The system involves character level sign recognition and translation. No detailed information is given about the dataset used, other than the fact that it is composed of all 26 alphabets of the English Language. For the model, they use a custom CNN model along with RNN models in order to display predictive text for faster typing in the user interface. Firstly by implementing CNN and RNN architecture, they create a sign language recognition and translation system. Consequently, they propose the design for a robotic arm, which would perform the gesture of the text inputted by the user into the software interface. The robotic arm is 3D, portable and printable, and can be further optimised using advanced controllers such as RaspberryPi, etc. Eventually, using the DenseNet architecture, Jalaja and Shekar (2022) reach an accuracy of 98.07%. Their sentiment analysis model achieves a remarkable accuracy of 93%. Although there exists future scope for implementing other architectures such as ResNet to see the impact on accuracy, this model sets a high benchmark in sentiment analysis from ISLR with its impressive accuracy rates.

2.2.2 Continuous Sign Language Recognition

Driven by the potential for contribution in continuous sign language recognition (CSLR) and upon careful consideration of its potential challenges, we are motivated to work in this field. Majority of our following research concerns the recent breakthroughs in this branch of sign language recognition.

In the paper titled "A Deep Neural Framework for Continuous Sign Language Recognition by Iterative Training", Cui et al. (2019) introduce a sophisticated framework designed to enhance continuous sign language recognition (CSLR) through the application of deep neural networks. This framework directly transcribes videos of SL sentences into ordered sequences of gloss labels, addressing the limitations of traditional methods that rely heavily on hidden Markov models (HMMs) and handcrafted features. These traditional approaches often fail to capture the complex temporal dynamics of SL and require costly temporal boundary annotations for each gesture. The proposed architecture is composed of two primary modules: a spatiotemporal feature extraction module and a sequence learning module. The spatiotemporal feature extraction module employs deep convolutional neural networks (CNNs) with stacked temporal fusion layers to effectively capture both spatial and temporal features from video data. Meanwhile, the sequence learning module utilises bidirectional recurrent neural networks (RNNs) to learn temporal dependencies and align sequences of gloss labels with video frames, ensuring coherent and accurate transcription. A notable innovation in this research is the iterative optimization process designed to maximise the representation capabilities of deep neural networks, particularly when data is limited. This process begins with an initial end-to-end training phase to generate alignment proposals. These proposals then serve as strong supervisory signals for fine-tuning the feature extraction module. By iteratively refining this process, the framework progressively enhances recognition performance. Specifically, the iterative training involves repeating the alignment generation and fine-tuning steps multiple times, each iteration improving upon the last. Furthermore, the authors explore the fusion of RGB images and optical flow data to enhance recognition accuracy. Optical flow provides motion information that complements the spatial cues captured by RGB images, leading to a more comprehensive understanding of dynamic gestures in SL videos. This multimodal approach significantly boosts the system's ability to capture and interpret the nuances of SL. The framework is rigorously evaluated on two challenging SL recognition benchmarks: the RWTH-PHOENIX-Weather 2014 and CSL (Chinese Sign Language) datasets. The results demonstrated substantial improvements over state-of-the-art techniques, with a relative performance increase of over 15% on both datasets. On the RWTH-PHOENIX-Weather 2014 dataset, the proposed method achieved a word error rate (WER) reduction from 27.1% to 23.0%, while on the CSL dataset, the WER was reduced from 38.3% to 32.2%. These significant performance gains underscore the effectiveness of the deep neural network architecture and the iterative training strategy. In comparison with previous methods, the proposed framework shows clear superiority. Traditional HMM-based methods and non-iterative deep learning approaches often fail to capture the intricate temporal dependencies and dynamic variations in SL videos. The iterative process and multimodal fusion approach adopted in this study address these shortcomings, leading to more accurate and reliable SL recognition. In terms of future scope, the framework's reliance on deep neural networks opens numerous possibilities for further research and improvement. Future work could explore the integration of more advanced neural network architectures, such as transformers, which have shown remarkable performance in various sequence-to-sequence tasks. Additionally, refining the multimodal fusion process by incorporating additional data modalities, such as depth information or skeletal data, could further enhance recognition accuracy. Expanding the dataset to include more diverse SL videos from different sign languages and contexts would also improve the model's generalizability and robustness. Overall, the proposed framework presents a robust and efficient approach for CSLR, providing a strong foundation for related applications in multimedia and computer vision in the future as well.

On the other hand, research done in 2020 emphasises the contribution of non-manual markers in CSLR and translation (Mukushev et al., 2020). The motivation behind the paper was to find out experimentally whether or not considering non-manual markers during CSLR improves the translation accuracy or not. The authors created their own Russian/Kazakh Sign Language dataset which is relatively smaller than popular benchmark datasets used. They recorded full signed sentences from five native Kazakh signers and extracted 5200 isolated frequently used sign samples. The recorded signs were chosen such that although the gestures were similar, involvement of facial expressions, eyebrow height, mouth and head orientation altered the meaning of the phrase or sentence in K-RSL. The authors worked on K-RSL because in this language, non-manual markers can alter the context of a sentence entirely. On the lexical level, signs which are manually identical can be distinguished by facial expression or specifically by mouthing (Mukushev et al., 2020). Whereas, on a morphological level, facial expressions and mouth patterns are used to convey adjectival and adverbial information (Mukushev et al., 2020). For example, universally, the negation in many sign languages is expressed by head movements while questions are distinguished from statements by eyebrow and head position. The authors used the classic ML model Logistic Regression to compare the accuracy obtained when the translation was done using manual markers only and when the translation was done using both manual and non-manual markers. Their results showed that by adding non-manual features it was possible to correctly identify 8 more samples as questions and 5 more samples as statements, which were classified wrongly when using only manual features. Testing mean accuracy scores are 73.4% and 77% on manual-only and both manual and non-manual features respectively

(Mukushev et al., 2020). There was a 3.6% improvement by the inclusion of non-manual markers (face. eyebrows, mouth) during translation.

2.3 Challenges of Sign Language Recognition

SLR continues to present multiple challenges that researchers are yet to fully resolve. The four most prominent hindrances, alongside their corresponding most promising progresses, have been addressed below.

2.3.1 Feature Extraction and Gloss Segmentation

Fang et al. (2023), introduce an interesting hypothesis that sign sequences conveying the same semantic information in sign language videos will exhibit greater visual similarity, while those expressing different semantics will be less similar. Such a claim is based on real-world observations. They design an end-to-end framework where at first, input of video streams apass on the 2D CNN to extract the frame-level features from each frame. Then it passes onto the KPBR structure for short-term temporal feature learning. To break down the KPBR structure, K stands for 1D convolution on the time dimension, P stands for the 1D-max pooling layer, lastly B and R stands for batch normalization and ReLU activation respectively. The clip level features pass onto the BiLSTM for long-term feature learning. Lastly, CTC loss aligns the prediction matrix. After that comes the most important part of the architecture, the Reinforcement Learning (RL) model which segments the clip-level features into small groups, giving them pseudo labels. The minimum cosine distance among the features allows the RL agent to create the groups of similar results.Then using gradient ascent for RL and backward propagation an optimized and fully working end-to-end model is created. The dataset used to evaluate the model is once again the popularly used RWTH-PHOENIX-Weather 2014 and CSL dataset. A total of four experiments are carried out to compare the results, first an end-to-end model without auxiliary task, second experiment uses auxiliary task but only decodes a prediction sequence and uses Cross-entropy. The third experiment is the designed VFS-RL method without the CTC auxiliary loss. Finally in the fourth experiment the intact VFS-RL model is used. For the evaluation metric, the commonly used WER is used to compare the results, and the VFS-RL method shows the best WER result of 24.4 in Dev and 24.9 in test. Apart from the experimented results, the VFS-RL model is also compared to the other respective state-of-the-art end-to-end models which carried out their work on the CSL dataset and obtains the best result of 1.5 WER(%) while other methods like STMC, FCN has a WER(%) of 2.1,3 respectively. To conclude, the use of RL allows the feature extractor to abstract more discriminative information from the video and also helps to reasonably divide the features of similar gestures into one group.

Meanwhile,Moryossef et al. (2023) utilizes linguistic cues observed in sign language corpora to introduce a novel approach to segmentation problems. Although spoken language text can be divided into a linear sequence of words with the help of punctuation marks, sign language's simultaneity obstructs it to such linear equations.

Previous segmentation work typically used binary classification to predict if each frame or pixel is part of a segment, neglecting the seamless transitions in continuous signing where boundaries are not clearly defined. This paper replaces this predominant Inside-Outside (IO) tagging scheme with Beginning-Inside-Outside (BIO) tagging leveraging linguistic prosodic cues. These cues are also known as non manual markers that have repeatedly come up in our discussions so far. This approach unifies individual sign segments and phrase segments (larger units comprising several signs) using a mechanism where in addition to predicting a frame to be in or out of a segment, it also classifies the beginning of the segments. This paper confronts phrase boundaries using optical flow features as a proxy for prosodic processes and sign boundaries utilizing linguistic information such as limited number of hand shapes in a sign and dominant hand's signal. This paper uses The Public DGS Corpus dataset which is a comprehensive linguistic dataset that includes both accurate sign-level annotation from continuous signing, and well-aligned phrase-level translation in spoken language. The proposed sign language segmentation method involves adjusting video to 25 fps, estimating and normalizing poses with MediaPipe Holistic, applying optical flow and 3D hand normalization, encoding sequences with an LSTM, using BIO tagging for classification, and decoding segments using a greedy algorithm. The experimental results show significant improvements in segmentation performance. The model achieves an F1 score of 63% and an IoU score of 69% for sign segmentation, and an F1 score of 65% and an IoU score of 85% for phrase segmentation. These results outperform previous IO tagging models. Moreover, the model demonstrates robustness in out-of-domain settings, maintaining competitive segmentation scores in zero-shot evaluations on different signed languages. The inclusion of optical flow and 3D hand normalization enhances the model's robustness in such contexts. In conclusion, this study presents an effective approach to sign language segmentation that significantly improves accuracy and shows promise for real-world applications. The findings suggest that incorporating linguistic cues and advanced modelling techniques can enhance the performance and robustness of sign language processing systems.

On the other hand, Rao et al. (2024) address the need for improvement in gloss representations. They present a simpler yet very effective technique of cross-sentence gloss consistency. Their strategy is to not only focus on the gloss correlation in an individual sentence but also consider the relationship among glosses from different sentences to achieve a richer correlation. This would eventually boost feature extraction performances significantly. To begin with, the proposed framework consists of three main components: gloss prototype (GP), gloss contrastive loss function and an auxiliary similarity fusion strategy (ASFS). The implementation was done by first aligning the visual backbone ResNet18, through which the gloss features are extracted. The GP stores prototypical features of all gloss categories, it serves as a comprehensive feature dictionary and is initialized by empirical/statistical distribution. It serves as a prototype memory bank for referral and to distil representative prototypes from diverse gloss samples. Alongside, it repressively refine the GP via momentum update, providing gloss feature references and gloss contrastive learning. A further use of gloss contrastive loss function reduces the distance between gloss sample point and its belonging prototype making it more distinguishable. ASFS used to fuse the intra-sentence recognition clues and cross-sentence clues using Softmax and CTC. Finally, sequences of words are generated using two-layer BiLSTM. The model was tested using the PHOENIX 14 and CSL Daily datasets and the evaluation metric used was WER. To be precise their model improves the current state of the art performances on the Dev sets of PHOENIX14, PHOENIX14-T, and CSL-Daily by 1.6%, 2.4%, and 5.7% respectively, and by 1.4%, 1.5%, 5.3% on Test sets respectively. The primary limitation of the conducted research was the narrow domain range in the dataset as they mostly focused on weather-type information. To summarise, the idea of cross-sentences validation gives us well-distinguished gloss prototypes. Simultaneously, it improves the gloss discrimination with a gloss contrastive loss and an auxiliary similarity fusion strategy revealing notable improvements amongst the results.

2.3.2 Alignment

Alignment issues have long persisted in this domain of research. The lack of rigid annotation of words to videos is one of its contributing factors. It complicates the process of converting continuous Sign Language to the corresponding phrases of a sentence.Pu et al. (2019) tackled this challenge that presented itself under a CSLR setting. They proposed the idea of dividing the process into two approaches: feature extraction and an encoder-decoder network with CTC for sequence modelling. They conducted their experiments on two public datasets RWITH-PHEONIX and CSL. Firstly, they used a CNN model 3D-resnet to process the video frames to extract important spatial and temporal features. Therefore, this process converted the video into an order of feature vectors. Secondly, these vectors were sent to an Bi-LSTM network which caught dependencies in the video forwards and backwards in time. As a result, this provided an adequate comprehension of the sequence. Thirdly, they decoded the sequence by using a CTC decoder and LSTM decoder. The CTC decoder implemented the alignment of sequences and the LSTM used the attention tool to pinpoint the important parts of the input sequence to generate each word in the output. Lastly, they used Soft-DTW to make sure that the two sequences from CTC and LSTM are aligned perfectly. By merging the uses of the mentioned models, their architecture to recognise Sign Language ensured promising results. After training the models with 94 sentences, they tested their system using different sentences which were not included in the training set but had vocabularies from it. Their system recorded 0.327 WER, which was substantially lower than many state-of-the-art methods such as LSTM CTC, S2VT and HLSTM. In addition, they did an independent signers test where they trained their architecture with videos of 40 signers and tested with videos of 10 signers. In this test, the sentences of the training and testing set were the same. Therefore, they achieved 0.980 BLEU(Bilingual Evaluation Understudy) which was better than existing methods which gave 0.936 - 0.948 BLEU.

Similarly, the weak supervision of sign language labels has been another prominent impediment for CSLR model performances. Hence, Xue et al. (2023) addresses it using an iterative alignment network and attention mechanism. The sign recognition process comprises two main modules: feature extraction and sequence learning modules. The feature extraction based on iterative alignment network consists of spatio-temporal residual network (STRN), time series convolution module and a sequence learning module. Firstly, the sign language is given as input in block, the STRN then extracts the block level features, then fed into the TCN(Temporal Correlation Enhancement) which enhances the correlation between block-level features and a three-layer bidirectional gated neural network. CTC is then used to learn the mapping relationship between the features and final sign language labels. Using this iterative approach the feature extraction is considerably refined, especially at word level feature. The word level features are then used as input into the encoder-decoder network. The encoder is a three layer LSTM unit and the decoder is based on an attention mechanism which generates the target sequence. The designed model is tested primarily on two datasets: RWTH-Weather-Phoenix 2014 (i.e. PHOENIX2014), and CSL. Once again, WER was employed as the evaluation metric. During result comparison, a few compromises were made. The 3D-ResNet uses a large number of parameters. As a result, the processing requires considerable computational power. As a countermeasure, Xue et al. (2023) adopted three-dimensional residual connections which reduces the cost. The final error rate emerged as low as 25.65% compared to the previous state-of-the-art model which had an error rate of 35.7% in the RWTH-Weather-Phoenix 2014 DATASET. In the CSL dataset, the WER(%) achieved was 1.7 compared to 2.1 under the STMC model (discussed later). To conclude, this framework uses its effective sequence mapping technique and the iterative alignment network to boost the accuracy of CSLR. However, it is worth noting that there are further scopes of using multi-modal fusion methods to improve the overall performance.

Amidst growing concerns over the alignment problem, cross-modality augmentation in the CSLR model emerged as a potential turn-around strategy. However, in 2024, a research identifies some of the key problems with this popular approach and provides a tentative solutions to it (Guo et al., 2024). Their work is noteworthy because they do so by not only considering manual markers, but also non-manual markers to some extent. The traditional approach is non optimal because it ignores some crucial facial gestures in the video input during feature extraction(Pu et al., 2020). This proves detrimental for accurate translation. This paper uses denoising-diffusion alignment (DDA) to tackle this problem. The DDA consists of two parts: a partial noise processor and a denoising-diffusion autoencoder. The former adds random Gaussian noise to the video part. This strategy enables the text modality to be a natural condition in the denoising process. Simply, the authors attempted to match up visual representations from videos with sequences of sign language glosses. The Denoising-Diffusion Autoencoder naturally conditions the textual sequence representations to denoising the part-noisy visual representation to achieve global alignment (Guo et al., 2024). The DDA also has the capability of spontaneously enforcing the CSLR model to learn contextual information within and between both modalities, enhancing the generalisation of visual representation. The three datasets they worked on are PHOENIX-2014, PHOENIX-2014T and CSL-Daily. DDA outperforms the key points supervised TwoStream-SLR by 1.1% and 0.9% WERs on the dev and test set of PHOENIX-2014 and even surpasses it by 0.5% and 0.6% WERs on the dev and test set of PHOENIX-2014T (Guo et al., 2024). Furthermore, it shows significant improvement compared to other multi-cue methods, which employ the pre-captured hands, face, key points, or heartmaps as supervision. Overall, the model is sensitive to sign movements and can capture the sign movements of the mouth, palm, head, and other sign events much more effectively to improve the sign recognition ability (Guo et al., 2024).

2.3.3 Translation

Amongst early attempts at conducting continuous sign language translation effectively, weakly-supervised learning had garnered popularity. However, the weakly-supervised learning aproach had problems like weak labels and noisy data. Koller et al. (2020) proposed a notion to resolve these aspects. This involved building a robust system for recognizing sign languages using a multi stream HMM to jointly align and synchronize mouth and hand shape patterns. Their research involved using the RWITH-PHEONIX Weather dataset. However, since this dataset did not have mouth shape annotations, they created an extension of the daatset by adding the mouth shape sequences from observing visible mouth shapes on it. Firstly, they used a parallel sequence learning where they divided the problem of mouth and hand shape recognition into sub tasks which worked parallel to each other as independent streams. Then these independent streams were modeled using a multi-stream HMM. Each stream handled sequence constraints specific to that channel of sign language. Consequently, to leverage parallelism, they created synchronization points among the streams which successfully aligned the independent sequences appropriately. Later CNN-LSTM models were integrated into the HMM streams which created a hybrid CNN-LSTM-HMM system. This, in turn, combined the advantages of deep learning for extracting features alongside utilising a probabilistic approach. The researchers used this hybrid model to learn and extract features from weak labels and noisy data. By applying this method, they got 26.5% WER on the first stream, 25.4% WER on the second stream and 24.1% WER on the third stream . Lastly, they had a 73.4% accuracy on their designed system where a previously published result on the same dataset was 62.8%. Therefore, despite being one of the earlier models, the hybrid CNN-LSTM-HMM system evidently addressed the challenges of weak labels and noisy data in continuous sign language recognition effectively.

Later Yin and Read (2020) emerges with a breakthrough when they contradict with the notion that gloss-based modelling is pivotal to achieve remarkable outcomes in overall SLR translation. While this strategy has been effective in some of the papers we have investigated as well, Yin and Read (2020) state that a reliance on gloss-based systems present a fundamental problem since ground truth glosses are suboptimal for capturing the nuances of sign language. The problem lies in the assumption that GT glosses provide an ideal intermediate representation for sign language, which may overlook the complexities of the language. The lack of universal standard for glosses and imprecise representation of sign language leads to a bottleneck for the multidimensional stream of data that sign language possesses. This issue is addressed through the introduction of the STMC-Transformer model for video-to-text translation, a novel solution aimed at enhancing translation accuracy in SLT systems. Yin and Read propose the use of a Spatial-Temporal Multi-Cue (STMC) network with a self-contained pose estimation branch to break down the input video into spatial features of multiple visual cues. Then, it uses a temporal multi-cue (TMC) module with stacked TMC blocks and temporal pooling (TP) layers to identify temporal correlation among inter-cue and intra-cue features. Leveraging recent advancements in Neural Machine Translation (NMT), particularly a two-layered Transformer maximising the log-likelihood, their technology facilitates more accurate translation by better preserving the intricacies of sign language expressions. Through extensive experiments on the PHOENIX-Weather 2014T and ASLG-PC12 datasets, Yin and Read demonstrate that their STMC-Transformer model outperforms both recurrent networks and Transformer models translating GT glosses, yielding a substantial improvement in translation accuracy. The model achieves significant improvements, surpassing the previous state-of-the-art by over 5 and 7 BLEU scores on gloss-to-text and video-to-text translation tasks, respectively, for the PHOENIX-Weather dataset, and over 16 BLEU on the ASLG-PC12 corpus. Notably, the STMC-Transformer's video-to-text translation outperforms the translation of ground truth (GT) glosses, challenging the assumption that GT gloss translation is the upper bound for SLT performance. This suggests that glosses may be an inefficient representation of sign language. For future work, the authors recommend focusing on end-to-end training of recognition and translation models or experimenting with alternative sign language annotation schemes. They also emphasize the importance of advanced techniques such as weight tying, transfer learning, and ensemble learning in pushing the boundaries of SLT methodologies. The study establishes a new benchmark for SLT and opens new avenues for further research and development.

Continuing with further developments on this model, the translation subsystem has undergone a transformative performance upgrade. Deep Learning has played a pivotal role behind this achievement. DL models are often more popular because they focus on the stronger features to achieve quick convergence. However, certain visual cues may be overlooked in the process. This is an important limitation because the comprehensive analysis of continuous sign language requires addressing the manual (hand shape) and non-manual (facial expressions, upper body posture) features simultaneously (Koller et al., 2020, as cited in Zhou et al., 2020). As such, Zhou et al. (2020) introduces a Spatial MultiCue (SMC) and a Temporal MultiCue (TMC) module dedicated to spatial representations and temporal correlations respectively. Combined, they contribute as part of the Spatial-Temporal MultiCue (STMC) model to achieve a high performing CSLR model. Firstly, the SMC module processes each video frame to extract spatial features from cues (full-frame, hand, face, pose). This involves pose estimation followed by patch cropping to eventually generate feature sequences of cues. The TMC module utilizes these to integrate information about the unique features of each cue (intra-cue path) with learning about the combination of different cues at different timings (inter-cue path) (Zhou et al., 2020). The output from TMC modelling is then passed to BiLSTM encoders and CTC layers. This achieves sequence learning and inference. Additionally, a joint optimization strategy is employed to ensure that STMC network sequence learning is end-to-end. Zhou et al. (2020) implements this network architecture using PyTorch. The resultant framework has remarkable results with large-scale prominent CSLR datasets: PHOENIX-2014, PHOENIX-2014-T, and CSL. It records state-of-the-art performance on all three datasets, achieving a 20.7% WER on PHOENIX-2014 while outperforming its best competitor on CSL by 4.1% on WER (Zhou et al., 2020).

It also manages to have a 17.6% and 5.3% edge over the recent multi-cue models LS-HAN and CNN-LSTM-HMM respectively. Extensive experimentation on the PHOENIX-2014 dataset also reveals that SMC can achieve a 3% improvement on this test set in comparison to the assumed baseline model. However, the unsupervised TMC module showed no such progress over 1D-CNN. Only when CTC loss helps its intra-cue path learn about the temporal dependency amongst cues, there is an improvement of 1.6% and 1.7% upon this development and test sets respectively. The STMC network itself manages to reduce WER by 4.8%. Alongside, the self-contained pose estimation branch plays a crucial role in regularisation and preventing overfitting. It reduces inference time by 44% which is an upgrade even upon the performance of an off-the-shelf model. However, it is worth noting that the model struggled to distinguish sign gestures from upper-body pose features. It performed best when considering hand-gestures and full frames alongside different cues along the inter-cue path. In summary, the STMC model's innovative aggregation of spatial and temporal features into an integrated architecture significantly enhances CSLR performance and efficiency.

In further attempts to improve CSLR accuracy, Min et al. (2021) focuses upon the iterative training scheme to address the prominent issue of overfitting. They discover that for overfitting problems, sufficiently training the feature extractor is necessary. Hence, a Visual Alignment Constraint (VAC) is proposed. This packs two auxiliary losses, one focusing on visual features and the one enforcing prediction alignment between the feature extractor and the alignment module. Before starting to work on the model, CTC-based CSLR models were revisited, which led to an interesting observation: only a few frames played a key role in training. So constraining the feature space would be critical to train a CSLR model. The framework proposed, incorporates three components: a feature extractor, an alignment module and an auxiliary classifier. The feature extractor takes in the images to sort out the frame-wise features before applying 1D-CNN to extract the local visual information. The features extracted are then sent to the alignment module and classifier. Two auxiliary losses are chosen during training. Firstly, the visual enhancement loss which primarily aligns visual features and the target sequences. Secondly, the visual alignment loss which aligns the short and long term context prediction through a method called knowledge distillation. To test the model, the RWTH-PHOENIX-Weather-2014 and CSL dtaatsets are used and alike previous researches we have explored so far, the prediction inconsistency is evaluated using WER. Min et al. (2021) performs the experiment in four parts. For the first experiment, the baseline model is used, resulting in Dev and Test WER(%) of 25.4 and 26.6 respectively. The second experiment using Baseline+VE model obtains Dev and Test WER(%) of 23.3 and 23.8 respectively. Following that, the third experiment consisting of the Baseline+VA model achieves Dev and Test WER(%) of 24.5 and 25.1 respectively. Finally, the proposed Baseline+VAC model gains an impressive Dev and Test WER(%) of 21.2 and 22.3 respectively. Moreover, comparisons made with the state-of-the-art models of the time such as STMC and FCN recorded WER(%) of 2.1, 3 respectively compared to the WER score of 1.6 achieved by the proposed model. To conclude, the VAC end-to-end model proposed addresses the overfitting problem, showing promising results and continuing improved CSLR outcomes.

Zuo and Mak (2022) continue to address the challenges associated with DL-based models. However, their approach is driven by an incentive to enhance the consistency of CSLR backbones. They target the visual and sequential modules and contribute an auxiliary constraint to each. HRNet pre-trained on MPII is used to extract key points before it undergoes a post-processing stage for heatmap refinement. The inclusion of these keypoint heat maps allows the visual module to achieve spatial attention consistency (SAC) constraint. This spatial attention module builds on the concept proposed in CBAM by dynamically assigning weights to imply channel importance. This way, it offers an improvement in terms of feature-extraction effectiveness and fewer parameters over former CSLR models that use an off-the-shelf pose detector and a multi-stream architecture. The module works similar to attention mechanisms and focuses on informative regions. To further improve recognition accuracy, they build on previous models like Visual Alignment Constraint (VAC) and Self-Mutual Knowledge Distillation (SMKD) which portray the benefit of enforcing consistency between visual and sequential modules to improve inter-module cooperation. Since both output features from both these modules represent the same sentence, Zuo and Mak (2022) introduce a sentence embedding consistency (SEC) constraint. The Sentence Embedding Extractor (SEE) is modeled based on QANet and it improves the overall feature representation power. It is also worth noting that the overall loss function of C^{2}SLR is a combination of SAC and SEC loss alongside traditional CTC. The effectiveness of these changes are evident in the results from experiments on CSLR backbones like VTB (VGG11, TCN and BiLSTM), CT (CNN and TCN) and VLT (VGG11 and Local Transformer). C^{2}SLR also achieves the revolutionary feat of being the first end-to-end method that outperforms models utilizing state optimization strategy. It records new state-of-the-art art performance on the PHOENIX-2014 dataset, surpassing STMC. It also achieves comparable performance on CSL against the state-of-the-art model MSeqGraph. With PHONEIX-2014-T, its consistent performance on the development and test sets, alongside it outperforming the state of the art performance on the latter, is evidence to its effectiveness in real-world scenarios for dealing with unseen data (Zuo & Mak, 2022). As such, this approach significantly enhances CSLR backbones by introducing constraints that achieve superior performance across various datasets.

More recent attempts at outperforming existing models like STMC have led to the emergence of attention-based CSLR models. Amongst these, SENet, CBAM, NLNet have yielded positive outcomes, not without their fair share of limitations. Eventually, L. Hu et al. (2023) implements a correlation network (CorrNet) that outperforms all three of these aforementioned models. Besides, they achieve better results at aggregating spatial-temporal information than techniques like I3D and TSM. They address the problem that the prevalent technology for capturing spatial features in CSLR models processes the video frames individually. As such, adjacent frame correlations and cross-frame signing trajectories get overlooked. Even if 3D CNN is used despite its compromised capacity to provide precise gloss boundaries, its success is limited to combining information within nearby video frames. The rigidity of the overall structure also means that it is not adequate for adapting to the role of finding the most important information from different videos. Using temporal convolutions for this task suffers from the same disadvantages. To improve this status quo, CorrNet focuses primarily on body trajectories. The network involves two modules.

Initially the correlation module generates correlation maps to track the movement of spatial feature patches across frames. The size of this correlation map is varied depending on a need for semantic consistency or fetching information from far-apart frames. Alongside, the size of feature patches can be adjusted: reducing it to even one pixel depending on computational constraints. Later these correlation maps are used by the identification module to focus upon the most informative spatial regions (hands and face) and closely correlated spatial-temporal features dynamically (L. Hu et al., 2023). However, this is achieved without any large 3D spatial-temporal kernel. Instead, multiple convolutions along parallel branches, each with progressively increasing dilation rates, are employed. These help reduce computational costs and contribute to model capacity. Consequently, the branches undergo group convolutions to reduce parameter and computation requirements. Eventually, the branch-wise extracted features are regulated using a learnable coefficient, alpha, depending on importance and groupwise convolution of different branches to achieve feature extraction from multiple spatial-temporal neighbourhoods. During training, L. Hu et al. (2023) maintained the same settings as the former state-of-the-art models. Yet, CorrNet achieved new state-of-the-art accuracy on all four of their datasets: PHOENIX14, PHOENIX14-T, CSLDaily and CSL. It records an 18.8% and 19.4% WER on PHOENIX14 development and testing datasets respectively. Similarly, it achieves 18.9% and 20.5% WER on the corresponding datasets of the PHOENIX14-T dataset. It even outperforms normal convolution consisting of more parameters and computation during the ablative study on the development and testing sets of PHOENIX14 dataset. Alongside, it offers an advantage over previous models that relied on hand and facial features from multistream inputs, body keypoints and pre-extracted hand patches. For instance, unlike STMC it does not require an additional pose-estimation network. The need for extra supervision or training stages is also alleviated yet CorrNet can achieve end-to-end dynamic training of the model as before.

Meanwhile, **Zuo2024TowardsOS** has recently approached the limitations in the translation subsystem from a different perspective. They observed that the CTC based models usually took the entire signed video as input for making predictions. Therefore, they developed a system to shift away from offline recognition towards effectively conducting SLR online. They used the same datasets mentioned in the majority of the research discussed so far: Phoenix-2014 (P-2014), Phoenix 2014T (P-2014T) , and CSL-Daily. Firstly, they constructed a sign dictionary where a pre-trained CSLR model TwoStream-SLR was used. It segmented the input videos into isolated signs. Consequently, DTW (Dynamic Time Warping) was employed to identify the signs by aligning the frames with the gloss labels. Since the segmented signs were not consistent in terms of accuracy, augmented signs were incorporated to enhance the training data. Secondly, they used ISLR models for encoding the sign videos. Additionally, gloss sampling was utilised to sample the glosses first and then multiple instances of the glosses were selected to form batches. In order to improve overall sign prediction, two types of cross-entropy loss functions (instance level and gloss level) played an important role. A saliency loss was also incorporated to focus on only the frames with signs. , he went into detecting signs in real time by using a sliding window approach. In addition, He used a voting system to remove any duplicate signs and background noise. Thus, the predictions were better filtered. Lastly, he added a gloss to the text network which translated the signs into actual text phrases or sentences in real time. By implementing his architecture he got 70/90% (dev/test) accuracy in recognizing the isolated signs. The online system here can, in turn, boost the offline models by aligning the dimensions of their features. In summary, this innovative approach not only enhances real-time sign language recognition but also improves the accuracy and efficiency of offline models.

2.3.4 Multilinguality

Maalej (2002) addresses it as a fallacy that there is a universal sign language. It is also widely acclaimed that SLR models can only be designed specific to individual languages. This is because of the noticeable differences in dialect, signing and other nuances that make sign languages in specific regions develop with unique characteristics.

H. Hu et al. (2023) challenges this concept by tackling the problem of multilingual sign language recognition using a unified framework. Their system used a shared combination of technologies including visual encoder and several multiple language dependent sequential modules. This contributed to temporal learning. A probabilistic decoding method was subsequently used to align the sign videos with the sign words. For the datasets, they use PHOENIX-2014, CSL and GSL-SD (which is a Greek sign language dataset). Firstly, the features are extracted from the videos using the shared encoder for all the languages. Each language is then allocated a separate sequential models. Therefore, each language gets its own model to learn the features and then verify them as sign words. Eventually, CNN-TCN is used as an additional shared layer to extract the features and before passing it to LSTM to do sequence modelling. Furthermore, the output is sent through a softmax layer where probabilities and CTC is utilised for training. H. Hu et al. (2023) tests their framework for a case consisting of two languages and another consisting of three languages. In case 1 (recognizing two languages), the model achieved a BLEU score of 0.840 and a WER of 0.191, outperforming state-of-the-art models, which reported BLEU scores ranging from 0.332 to 0.795 and WERs from 0.245 to 0.757. For case 2 (recognizing three languages), the model attained a BLEU score of 0.852 and a WER of 0.181.R.

2.3.5 Dataset

Majority of the significant research that we have covered so far uses one or more of the datasets mentioned below. All of these large public datasets have been instrumental in advancing recent progress in continuous sign language recognition. Hence, we have summarised their key dataset characteristics below.

RWTH-PHOENIX is German weather forecast recordings involving nine actors. It features 6841 sentences with a vocabulary of 1295 signs. It is divided into 5672 training, 540 development, and 629 testing samples. PHOENIX2014 is very similar but it features the nine actors against a clean background with a resolution of 210 × 260. Its sign language data has a specific structure and annotation, making it more suitable for sign language recognition tasks. To summarise, RWTH-PHOENIX is a broader term that may refer to the initial corpus provided by RWTH Aachen University while PHOENIX2014 is a specific version of the dataset. The latter also has a training-development-testing split of 5672-540-629. PHOENIX2014-T comes as an extension of PHOENIX2014, providing additional data and annotations to support both CSLR and translation research. It contains 8247 sentences with a vocabulary of 1085 signs, divided into 7096 training, 519 development, and 642 testing instances. CSL is a Chinese Sign Language dataset. It has 50 signers representing a vocabulary size of 178 signs across 100 sentences gathered in a laboratory environment. The dataset contains 25000 videos, divided into training and testing sets in an 8:2 ratio, gathered in a laboratory environment. In contrast, CSLDaily reflects more natural daily activities. Videos are recorded indoors at 30 frames per second by 10 signers and the dataset includes 20654 sentences, with 18401 training, 1077 development, and 1176 testing samples.

However, as Manoharan and Roy (2022) has mentioned there is not enough continuous datasets available for research. They state how several models perform well with these training datasets but suffer from a case of overfitting and fail to perform according to expectations on unseen data. Besides, Hao, Min et al. (2021) state that most CSLR datasets only have sentence-level annotations owing to frame-level annotations being an expensive alternative. Thus, the current progress in this domain is hindered by the lack of high-quality datasets.

Hence, several research groups are now increasingly more concerned with improving the dataset in this field of research. Kim and O'Neill-Brown (2019) introduced us to the idea of synthetic data, which is the mimic of real world-data created with the help of various algorithms. Their approach utilises a baseline prototype ASLR based on DeepHands using the Kinect Sensor and a graphical user interface. The dataset involved is the Sign Language Lexicon Video Dataset (ASLLVD), comprising of 10,000 ASL signs by 6 native ASL signers and human-annotated linguistic information such as gloss and handshape labels. For feature extraction, they use OpenPose which provides the different models pre-trained on publicly available datasets. Additionally, 25-point body pose and 20-point hand detection models are involved. During preprocessing, two main types of features are extracted from all frames: 2D coordinate of the hand followed by the DeepHand hand-shape. The latter is created using a CNN-based sign language recognizer trained on approximately one million images. Later K-means clustering is used for classification. Consequently, for Data Augmentation, several manipulation strategies such as adding noise, changing brightness change, rotation and zoom were considered. However, for the scope of the proposed experiment, only Rotate and Zoom is used. The experiment is carried out in several stages using varying amounts of synthetic data (0%, 100%, 300%, 500% and 1000%) and accuracy is used as the primary evaluation metric. The best performance was recorded with 1000% synthetic data with K-means cluster size of 3000. However, interestingly, the relationship between performance and synthetic data quantity appears non-linear. Performance dropped significantly until 300% but showed significant improvement when 1000% was surpassed. Overall, the finding of the research was that with careful data augmentation, we can significantly enhance SLR recognition models.

More recently, Kezar, Thomason, et al. (2023) addressed this crucial problem primarily for ISR. We have previously discussed in page 8 the significance of phonology for accurate ISLR in reference to another paper co-authored by Kezar, Pontecorvo, et al. (2023). In this paper, that relevance is reestablished by stating that models considering signs as a sequence of linguistic components improve ISLR by up to 6%. As such, they introduce a new dataset, the Sem-Lex Benchmark, comprising 84,568 videos of isolated sign production, the largest of its kind under the status quo. The objective of this dataset is to cater to the growing number of researchers who no longer treat SLR as an entirely vision-based problem. Alongside its data size, the dataset boasts using ethical and consistent sourcing. Deaf fluent signers are involved to counter heterogeneity and inconsistencies in sign articulation of videos. Additionally, a final inter-rater reliability threshold of only 0.7 was used for selecting labellers. However, its most prominent contribution is adding representative and high-quality data amidst a profound absence of well-annotated datasets.This is achieved by incorporating linguistic information accompanying each sign alongside a new annotation scheme. They rely on videos of ASL signs to label ASL signs themselves Kezar, Thomason, et al. (2023). Utilising the lexical databases in this way allowed their new resultant dataset to have sign labels aligned with other available linguistic resources: ASL-LEX, ASL Citizen, and ASL SignBank. In turn, phonological descriptions that are recorded in ASL-LEX can now be utilised as phonological information of ISR dataset without manually annotating it. Similarly, the cross-compatibility of this dataset with ASL SignBank has enabled labelling of continuous signing videos. However, this benchmark is yet to incorporate grammatical features as well as coarticulation. Simultaneous phonological feature and gloss recognition has already shown to improve few-shot ISR accuracy by 6% and overall ISR accuracy by 2% respectively Kezar, Thomason, et al. (2023). Further inclusion of more aspects of ASL to address the current limitations of this dataset is expected to transform it into a prominent driving force behind the future research on CSLR.

Chapter 3

Work Plan

The work plan for our thesis consists of three main phases, starting with Pre-Thesis 1 and then Pre-Thesis 2, and finally the defense. Pre-Thesis 1 lasted for four months which involved member selection and team building, selecting a topic and supervisor, writing an abstract, reading relevant papers and summarising them into a literature review, and finally writing a report. Secondly, pre-thesis 2 will also likely last for four months and it involves collecting and analyzing data, building and evaluating models, and also writing a report and a poster. The final phase of the thesis will involve analyzing the results, completing the paper, and preparing for the final defense. The Gantt chart [3.1] visually represents the time frame and level of completion for each individual task within the overall work plan. The time frame has also been color coded to signify the three different phases, and we have used green to showcase that the first phase is over, whereas the second phase and final phase have yet to be completed.

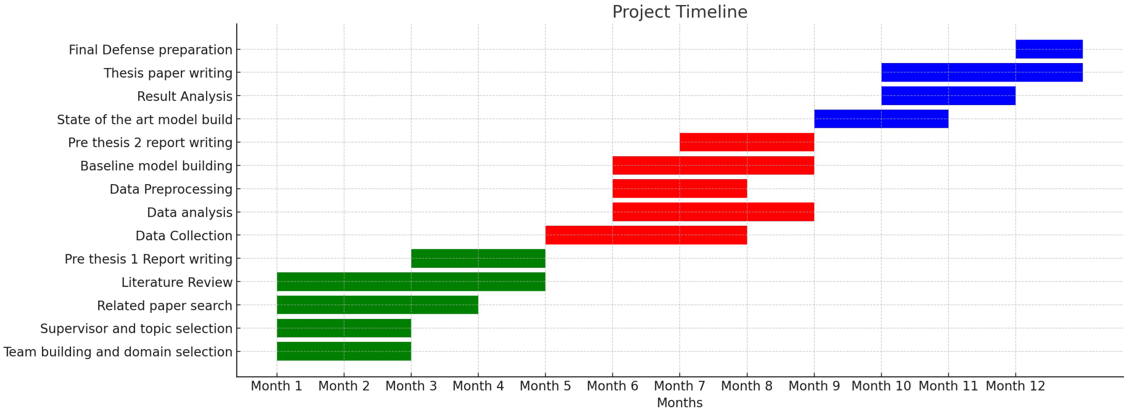

Figure 3.1: Workplan Gantt Chart

Chapter 4

Conclusion

The objective of this research is to contribute to the motivation about enhancing the bridge between the communication of signers and the rest of the world. In this paper, we summarised our findings on the different problems that exist in creating Sign Language translation models and categorised them. For each category, multiple researches have been analysed, drawing conclusions on their limitations, barriers and scope for future work. Upon drawing conclusions on the advancements already established in this area and considering the scopes of further research, we hope to be able to make our own improvements. Now that we have summarised the barriers faced by recent researchers who have contributed to this case, our task remains to utilise the technologies in NLP, neural networks, and image processing to come up with a holistic approach to the recognition and translation issue, while careful consideration of tackling the barriers. In the age of digitalisation, communication, a fundamental living requirement, should not have to be a concern for anyone.Understanding and interpreting Sign Language is not only necessary for easy communication with signers, but also for giving the Deaf community the recognition and ease of life they deserve.

Bibliography

Cui, R., Liu, H., & Zhang, C. (2019). A deep neural framework for continuous sign language recognition by iterative training. *IEEE Transactions on Multimedia*, *21*(7), 1880–1891. https://doi.org/10.1109/TMM.2018.2889563

Fang, Y., Wang, L., Lin, S., et al. (2023). Visual feature segmentation with reinforcement learning for continuous sign language recognition. *International Journal of Multimedia Information Retrieval*, *12*(39). https://doi.org/10.1007/s13735-023-00302-8

Guo, L., Wang, Z., Liu, Q., Feng, W., Wang, L., & Liang, H. (2024). Denoising-diffusion alignment for continuous sign language recognition. *ArXiv*. https://arxiv.org/abs/2305.03614

Hao, A., Min, Y., & Chen, X. (2021). Self-mutual distillation learning for continuous sign language recognition. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 11283–11292. https://doi.org/10.1109/ICCV48922.2021.01111

Hu, H., Pu, J., Zhou, W., & Li, H. (2023). Collaborative multilingual continuous sign language recognition: A unified framework. *IEEE Transactions on Multimedia*, *25*, 7559–7570. https://doi.org/10.1109/TMM.2022.3223260

Hu, L., Gao, L., Liu, Z., & Feng, W. (2023). Continuous sign language recognition with correlation network. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2529–2539.

Jalaja, S., & Shekar, K. (2022). Robotic arm for sign language interpretation with sentiment analysis and auto-complete text features. *International Journal of Engineering Research and Applications*, *12*(10), 92–99. https://www.ijera.com

Jamwal, A., Vasukidevi, G., Malleswari, T., Vijayakumar, T., Reddy, L. S., & Gupta, A. S. A. L. G. G. (2022). Real time conversion of american sign language to text with emotion using machine learning. *2022 Sixth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, 603–609. https://doi.org/10.1109/I-SMAC55078.2022.9987362

Joshi, A., Sierra, H., & Arzuaga, E. (2017). American sign language translation using edge detection and cross correlation. *2017 IEEE Colombian Conference on Communications and Computing (COLCOM)*, 1–6. https://doi.org/10.1109/ColComCon.2017.8088212

Kezar, L., Pontecorvo, E., Daniels, A., Baer, C., Ferster, R., Berger, L., Thomason, J., Sevcikova Sehyr, Z., & Caselli, N. (2023). The sem-lex benchmark: Modeling asl signs and their phonemes. *ArXiv*. https://doi.org/10.48550/arXiv.2310.00196

Kezar, L., Thomason, J., & Sehyr, Z. S. (2023). Improving sign recognition with phonology. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2732–2737.

Kim, J., & O'Neill-Brown, P. (2019). Improving american sign language recognition with synthetic data. *Proceedings of Machine Translation Summit XVII: Research Track*, 151–161.

Koller, O., Camgoz, N. C., Ney, H., & Bowden, R. (2020). 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), 2306–2320. https://doi.org/10.1109/TPAMI.2019.2911077

Maalej, Z. (2002). Book review: Language, cognition, and the brain: Insights from sign language research. *The Linguist List*. http://www.linguistlist.org/issues/13/13-1631.html

Manoharan, M., & Roy, P. (2022). A comprehensive review of sign language recognition: Different types, modalities, and datasets.

Matlani, R., Dadlani, R., Dumbre, S., Mishra, S., & Tewari, M. A. (2022). Real-time sign language recognition using machine learning and neural network. *Proceedings of the 2022 International Conference on Electronics and Renewable Systems (ICEARS)*, 1381–1386. https://doi.org/10.1109/ICEARS53579.2022.9752213

Min, Y., Hao, A., Chai, X., & Chen, X. (2021). Visual alignment constraint for continuous sign language recognition. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 11522–11531.

Moryossef, A., Jiang, Z., Müller, M., Ebling, S., & Goldberg, Y. (2023). Linguistically motivated sign language segmentation. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 12703–12724.

Mukushev, M., Sabyrov, A., Imashev, A., Koishybay, K., Kimmelman, V., & Sandygulova, A. (2020). Evaluation of manual and non-manual components for sign language recognition. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 6073–6078.

Niu, Z.-Y., & Mak, B.-K. (2020). Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition. *Proceedings of the European Conference on Computer Vision*.

Pu, J., Zhou, W., Hu, H., & Li, H. (2020). Boosting continuous sign language recognition via cross modality augmentation. *ArXiv*. https://doi.org/10.1145/3394171.3413931

Pu, J., Zhou, W., & Li, H. (2019). Iterative alignment network for continuous sign language recognition. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4160–4169.

Rao, Q., Sun, K., Wang, X., Wang, Q., & Zhang, B. (2024). Cross-sentence gloss consistency for continuous sign language recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, *38*(5), 4650–4657.

Wali, A., Shariq, R., Shoaib, S., Amir, S., & Farhan, A. A. (2023). Recent progress in sign language recognition: A review. *Machine Vision and Applications*, *34*, 1–20.

Xue, C., Yu, M., Yan, G., et al. (2023). Continuous sign language recognition based on iterative alignment network and attention mechanism. *Multimedia Tools* *and Applications*, *82*(15), 17195–17212. https://doi.org/10.1007/s11042-022-13705-7

Yin, K., & Read, J. (2020). Better sign language translation with stmc-transformer. *Proceedings of the 28th International Conference on Computational Linguistics*, 5975–5989.

Zhou, H., Zhou, W., Zhou, Y., & Li, H. (2020). Spatial-temporal multi-cue network for continuous sign language recognition. *Proceedings of the AAAI Conference on Artificial Intelligence.*

Zuo, R., & Mak, B.-K. (2022). C^{2}slr: Consistency-enhanced continuous sign language recognition. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5121–5130. https://doi.org/10.1109/CVPR52688.2022.00507