International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

A Study on Techniques and Challenges in Sign Language Translation

Prachi P. Waghmare[1], Ashwini M. Deshpande[1]
[1]Electronics and Telecommunication Department,
MKSSS's Cummins College of Engineering for Women
Pune, India
prachi.waghmare@cumminscollege.in, ashwini.deshpande@cumminscollege.in

Abstract: Sign Language Translation (SLT) plays a pivotal role in enabling effective communication for the Deaf and Hard of Hearing (DHH) community. This review delves into the state-of-the-art techniques and methodologies in SLT, focusing on its significance, challenges, and recent advancements. The review provides a comprehensive analysis of various SLT approaches, ranging from rule-based systems to deep learning models, highlighting their strengths and limitations. Datasets specifically tailored for SLT research are explored, shedding light on the diversity and complexity of Sign Languages across the globe. The review also addresses critical issues in SLT, such as the expressiveness of generated signs, facial expressions, and non-manual signals. Furthermore, it discusses the integration of SLT into assistive technologies and educational tools, emphasizing the transformative potential in enhancing accessibility and inclusivity. Finally, the review outlines future directions, including the incorporation of multimodal inputs and the imperative need for co-creation with the Deaf community, paving the way for more accurate, expressive, and culturally sensitive Sign Language Generation systems.
.

Keywords-sign language, sign language translation, sign language recognition, Deep learning

I. INTRODUCTION

Communication plays a central role in our daily lives and social interactions. However, for individuals with hearing impairments, particularly deaf people, the inability to hear poses a significant challenge to their ability to communicate naturally within a society dominated by spoken language.

Sign language (SL) serves as a crucial mode of communication for the deaf community, involving a complex system of gestures, hand movements, and facial expressions. These nonmanual elements, including raised eyebrows, smiles, and specific mouth and tongue movements, form the essence of sign language, transforming sentences and words into visual gestures. Achieving effective communication between individuals with hearing loss and those without necessitates bilateral translation, bridging the gap between natural signs and spoken language.

Sign language translation (SLT) is a fundamental task within sign language interpretation, aiming to convert continuous signing videos into text-based translations in natural language. This process involves understanding the grammatical structure of sign language and translating it into visual frame streams. SLT comprises two essential components: grammar conversion and word translation. However, the decomposition of continuous video sequences into sign words presents challenges in terms of accuracy and speed, exacerbated by the significant semantic gap between visual and written contexts.

This review delves into the realm of machine translation, pose estimation, and video generation techniques, exploring methods to generate and process sign language components. The focus encompasses machine translation, pose estimation, and video generation techniques, each incorporating specific criteria tailored to their respective applications.

A. Background

Sign language processing involves two key applications: sign language translation (SLT) and sign synthesis. Sign language encompasses diverse elements, including hand movements, fingerspelling, and crucial nonmanual features expressed through facial expressions. For instance, conveying happiness involves reflecting joy on the face while signing the word 'person' requires specific hand movements, and signing 'fast' involves swift and continuous hand and head motions.

Recognizing these intricate signs and performing corresponding actions presents a challenge. The educational focus often revolves around oralism and lip reading, potentially leaving deaf children isolated if not introduced to sign language early on. Early exposure to sign language can mitigate this isolation, preventing its adverse effects on the overall personality development of deaf children.

To facilitate meaningful dialogue between the hearing and the deaf, it is imperative to develop efficient systems capable of translating spoken languages into sign languages and vice versa. Such systems play a vital role in bridging the communication gap and fostering inclusive interactions within a diverse society Another example is the Indian sign for the word "AUNT", as shown in Figure 1. To demonstrate the gesture for the word "aunt", make a sign with your index finger near your nose, showing a sign for women, and make a c shape near your face.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

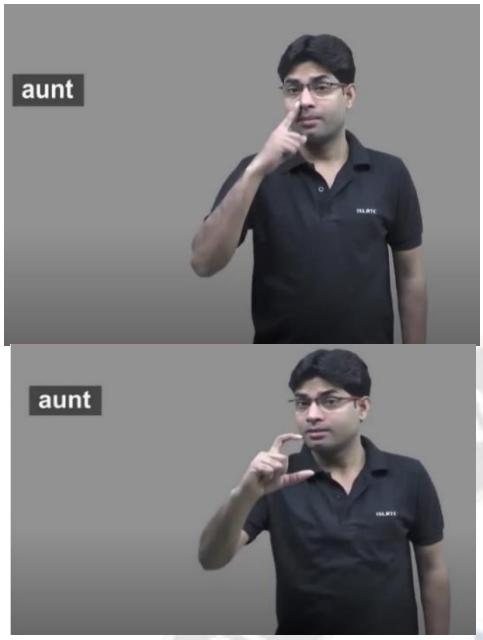

Figure 1. Sign for the word aunt [1]

It is challenging to recognize these signs and perform the corresponding actions. The focus of SL learning in school is oralism or lip reading. A child may fail to grasp SL and may become totally isolated if not made aware of it at an early age. To cater to their learning ability, introducing them to SL at an early age may suffice for the problem of isolation. Isolation can have a deep impact on the overall personality traits of deaf children. To achieve effective dialog between the hearing and the deaf, it is necessary to build efficient systems that can translate spoken languages into SLs and vice versa

B. RESEARCH QUESTIONS:

The primary aim of this review is to categorize the existing literature focusing on Sign Language Machine Translation (SLMT), sign synthesis, and the criteria used to evaluate translation methods. The specific study topics and questions are outlined in Table I, each accompanied by its rationale.

TABLE I: RESEARCH QUESTIONS AND MOTIVATION

Research questions	**Motivation**
What are the current state-of-the-art methodologies in SL processing?	Understanding early research efforts in the field, including their advantages and limitations
What are the different platforms/apps that are available for generating synthesized SL videos?	Exploring web-based and mobile applications that aid the deaf community.
What datasets are available to process SL?	Identifying benchmark datasets crucial for synthesizing signs.
What are the evaluation metrics for SL recognition and translation?	Understanding assessment metrics and performance parameters used to gauge system efficiency.

In this section, we delve into the components of Statistical Machine Translation (MT) and Neural MT (NMT) necessary for sequential translation tasks from natural languages like English, Chinese, Spanish, and Arabic to Sign Language (SL). MT combines insights from natural language, statistics, translation theory, and artificial intelligence (AI) to develop systems translating human languages effectively and reliably. Various methods for MT, as illustrated in Figure 2, are discussed chronologically in the following section.

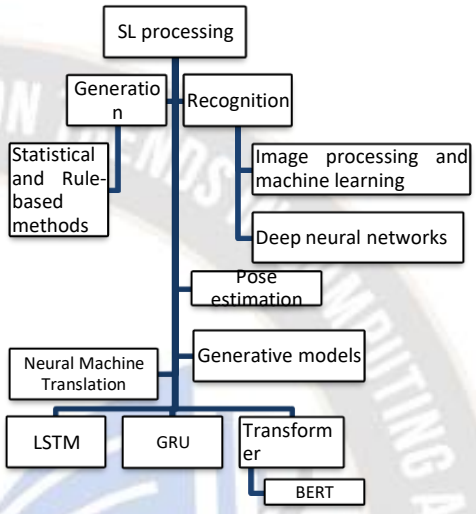

Figure 2.SL Processing techniques

2.Sign Language Translation Methods (SLTM): This task involves translating text into gloss, where gloss represents the written syntax depicting how a corresponding sign is formed or how word order is structured to convey a sign. Gloss generation serves as an intermediary step in synthesizing Sign Language (SL). SLMT technologies aid the deaf community by transforming sign presentations into signing avatars or synthesized videos, enabling information reception and communication. In this section, methods for translating text from one language to another, including SL, are explored.

For computer translation from spoken words to SL, language processing models are essential. The primary language processing paradigm includes components such as input text parser, word eliminator, stemmer, and phrase reordering. Tokenization is utilized by the parser to obtain part of speech (POS) labels for the incoming text. These tokens are then processed by the eliminator module, which removes undesired tokens. The stemmer module focuses on verbs, converting each word into its basic present tense form. Notably, the sequencing differs between spoken languages and SLs. The phrase reordering model, rooted in SL grammar, rearranges arguments to align with this distinction. A significant challenge in SL generation and processing revolves around weakly annotated data. The subsequent section investigates efforts to process existing datasets, extracting textual features, and examines text-processing steps employed by both statistical methods and Neural Machine Translation (NMT).

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

2.1 Statistical Machine Translation Approaches for Sign Language

Hung-Yu Su and Chung-Hsien Wu [2] introduced a structural statistical machine translation (SSMT) model for translating Chinese into Taiwanese Sign Language (SL) using synchronous context-free grammar (SCFG). This innovative approach demonstrated superior performance, particularly for lengthy sentences. An evaluation was conducted using the Mean Opinion Score (MOS).

Yu-Hsien Chiu and Chung-Hsien Wu [3] proposed a syntax- and phrase-level alignment model to facilitate translation between Chinese sentences and Taiwanese sign sequences. Additionally, they presented a maximum posterior (MAP) algorithm for sign video synthesis.

A statistical machine translation method [4] for both SL and written language was presented, focusing on the linguistic interplay between German Sign Language (DGS) and German. This study examined a dataset comprising 1399 words in DGS and German, employing various statistical models, including IBM Models 1-4. Evaluation metrics included the Word Error Rate (WER) and Position-Independent WER (PER). However, the reliance on rule-based techniques limited its adaptability to new language combinations.

Ian Marshall and Eva Safar [5] introduced a statistical approach to English text parsing using the Carnegie Mellon University (CMU) parser. They utilized a Discourse Representation Structure (DRS) to isolate specific semantic elements, such as nominal, morphological, and adjectival-based predicates, discourse contexts, and temporal relationships. The DRS was subsequently translated into a related Semantic Head-Driven Phrase Structure Grammar (HPSG), which formed the basis for SL generation. HPSG, a phrase structure grammar, drew from Ferdinand de Saussure's sign theory and ideas from computer engineering, making it suitable for natural language processing due to its conventional formalism and modular structure.

Lynette van Zijl and Guillaume Olivri [6] developed a system for resolving word problems using the GNApp program. This unique approach augmented the language front-end to address differences in perception, thereby improving the accuracy of rule-based translation from English to South African Sign Language. It also mitigated the challenge of limited available datasets.

Achraf Othman and Mohamed Jemni [7] proposed a statistical Sign Language Machine Translation (SLMT) model to automatically translate between English and American Sign Language (ASL) gloss. This model employed the GIZA++ statistical word alignment toolkit and the MOSES phrase tool to generate a 3D avatar for interpreting SL from English and ASL gloss.

Tirthankara Dasgupta et al. [8] converted English text into International Sign Language (ISL) gloss using the lexical functional grammar (LFG) structure. The system encoded the grammatical relations, subject, object, and tense of input sentences through an f-structure, which represented the internal organization of sentences. While the system operated effectively overall, it encountered challenges primarily in handling compound sentences and directed verbs.

In the realm of statistical machine translation, the focus lies on extracting morphological, semantic, and syntactical information through parsers. Some approaches rely on alignment scores derived from word mappings, while others address discourse and word sense disambiguation to capture context-dependent word meanings, as seen in languages like English. Achieving robust translation necessitates consideration of all stages of natural language processing, including morphological, semantic, discourse, and pragmatic stages. Tackling the complex task of context capture within sentences is a key challenge, and exploring neural machine translation (NMT) and deep learning approaches holds promise in addressing this challenge.

2.2 Neural Machine Translation Methods

Neural Machine Translation (NMT) models leverage artificial neural networks to estimate the probability of word sequences. In this approach, the entire input sentence is condensed into a fixed-length vector. An encoder dissects phrases into constituent words, representing their meanings with vectors. These representations are collectively interpreted and decoded through weighted distributions across these encoded vectors.

D. Bahdanau et al. [9] pioneered an NMT method utilizing an encoder-decoder architecture for translating English to French sentences, surpassing phrase-based translation methods. Their model, evaluated on sentences ranging from 30 to 50 words, employed a recurrent neural network (RNN)-based search model, with evaluation conducted using the bilingual evaluation understudy (BLEU) score.

Biao Fu et al. [10] introduced a token-level contrastive framework for Sign Language Translation (ConSLT). This method involved two stages: first, a continuous SL recognition model extracted SL gloss, and then token-level contrastive learning calculated similarities between positive and negative sentence pairs. The objective was to minimize losses during training, resulting in a BLEU score improvement on the PHOENIX 14T dataset.

Nada B. Ibrahim et al. [11] developed an automatic visual Sign Language Recognition System (SLRS) for translating isolated Arabic word signs into text. The system achieved a remarkable 97% recognition rate, outperforming other methods, owing to its precise hand and head positioning using hand segmentation, tracking, feature extraction, and classification.

J. Zheng et al. [12] proposed a frame stream density compression algorithm to eliminate redundant frames in video frames using the RWTH-PHOENIX-Weather 2014T dataset. This technique, employing a temporal convolution unit, improved semantic information extraction during temporal tokenization.

To address the challenge of vision-based sequence learning, Hao Zhou et al. [13] introduced a spatial-temporal multicue (STMC) network, demonstrating the potential of various cue sources for SL detection and translation. The proposed method outperformed other benchmark techniques, showcasing new state-of-the-art performance.

In a different paradigm, the task of translating SL videos to text was divided into two tasks. A visual action recognition task converted sign videos to semantics using a visual language mapper, combining spatiotemporal information from sign films with semantics from text transcriptions, enhancing translation accuracy [14].

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

Necati Cihan Camgoz [15] proposed a method to generate SL videos from spoken language sentences using autoregressive transformer decoder models trained on the PHOENIX-14T dataset. Their approach, integrating spatial and temporal representations, achieved notable accuracy in translation and recognition, demonstrating the effectiveness of their framework. Convolutional and sequential networks incorporate neural SLT by learning tokens [16]. The tokenization scheme proposed by Camgoz et al. extracts features from a trained convolutional neural network (CNN) to give good representations of sentence-level annotations. The CNN is trained on two datasets, where the first dataset contains millions of images with blur and noise, while the second set is small and less noisy.

Researchers have explored various neural architectures for SL recognition (SLR). For instance, the bidirectional spatial-temporal long short-term memory fusion attention network (Bi-ST-LSTM-A) [17] employed an attention mechanism in a bidirectional ST-LSTM encoder-decoder framework, achieving an accuracy of 81.22%.

Kayo and Read Jesse [18] improved sign gloss to text translation using spatiotemporal multicue networks. These networks incorporated spatiotemporal graph neural networks and kinetic 3D action datasets for accurate action recognition in SL, enhancing SL generation systems. The authors in [19] proposed a method to convert English text to American SL (ASL) with an ASL corpus and implemented a sequential neural machine network with a transformer model, achieving an accuracy of 97.89\% with different network layers.

The study in [20] tested an RNN-based neural network and a transformer with two additional techniques, namely, byte pair encoding (BPE) and back-translation, on the PHOENIX dataset. The BPE tokenization network based on the n-gram model achieved a score improvement.

The survey [21] presented different combinations of approaches for converting text to sign language video. This survey examined NMT models as well as methodologies with different modalities, such as end-to-end translation from text to 3D sign poses, which can be animated. It also highlighted the scarcity of the data needed to process regional language for various language pairs.

The first statistical MT models were built to accomplish this task. The article [22] proposed a hybrid MT approach that combines rule-based and phrase-based MT algorithms to produce glosses in Indian SL that are in sync with equivalent English sentences. The issue of subject phrase generation will be addressed by this new strategy. Prior to phrase-based MT, POS tagging and named entity recognition were used to determine the content of a sentence. In an effort to enable automatic SLT, [23] translators work similarly to other digital translators, allowing users to type words and then access the translation in another language as a 3D avatar. The words are processed using the grammar rules of the language of interest and then correctly changed (in terms of word orders and verb tenses) using the sign database the authors of [24] presented the frame stream density compression (FSDC) technique. An improved architecture that comprises a T-Conv unit, an NMT module, and a sequentially dynamic hierarchical bidirectional GRU (DH-BiGRU) replaces the encoder model. More detailed information is gathered by using an enhanced component that incorporates temporal tokenization data. LSTM and GRUs are two time series learning techniques.

In the realm of text-to-SL translation, researchers proposed innovative strategies. For instance, Luqman [25] combined convolutional neural networks (CNN) with LSTM and GRU models to effectively learn spatial and temporal information, achieving a low Word Error Rate (WER). A method [26] for converting Marathi text into a representation in Indian SL (ISL) was described. Due to the limited linguistic resources for this language pair, the researchers adopted a rule-based method. “Gloss” is the name of the matching ISL representation. To dynamically interpret SL, the gloss is translated to the System of Hamburg Notation, also known as Signing Gesture Markup Language (SGML). For deaf people to communicate in the future, Iranian SL (IrSL) [27] should be taught to children who do not have any hearing issues. To do this, the components of eight different IrSL gestures—either with one hand or both—have been created in a virtual reality game using the Unity game engine. Real-time recognition using a deep learning system built on a visual algorithm is what causes fluid movements. After that, the player records were kept, and the authors made progress in learning the selected IrSL signals under examination.

Furthermore, researchers introduced a neural-based SLT framework with multimodal feature fusion based on a dynamic graph, enhancing translation models [28]. The SL video-to-spoken language text translation model is initialized using pre-trained bidirectional encoder representations from transformers (BERT)-base and mBART-50 models. The authors of [29] used the frozen pre trained transformer technique to reduce overfitting by freezing the bulk of the training-time parameters. Pretrained language models can be employed to enhance SLT performance

Additionally, syntax-aware approaches, such as the Text2Gloss transformer described in [30], utilized word dependence tags to enhance discriminative abilities in embeddings. These models extended transformers to include lexical dependency aspects, leading to improved performance in SLT.

Table II summarize significant strides in the field of neural machine translation for sign languages, employing diverse architectures and innovative techniques to bridge the communication gap between spoken and sign languages.

TABLE II: SUMMARY OF THE APPROACHES USED TO GENERATE GLOSS SEQUENCES FROM A DATASET

Approach	**Dataset**	**Methodology**	**Metrics**	**Limitations/ Future scope**	**Output**
Stoll et.al[40]	RWTH-PHOE	Encoder and decoder (NMT)	BLEU 50.67	Difficulty conveying complex spatiotemp	Gloss

	NIX-2014-T			oral syntax in text.	
Camgoz et.al[41]	RWTH-PHOENIX-2014-T	Multichannel transformer	BLEU 17.25	Translation loss due to overfitting. Future work involves a simplified, lightweight graph-based neural SLT model	Gloss
Saunders et.al[41]	RWTH-PHOENIX-2014	Utilizing a sequence-to-sequence learning loss function, CTC, provides an alternative method of reduced supervision	WER 24.59	Enhancing networks' understanding of language connections among sign articulators (faces, hands, bodies).	Gloss
Zheng et.al[28]	RWTH-PHOENIX-2014	Dynamic graph-based neural SLT	BLEU 46.24	Proposed future work: Develop a simplified, lightweight graph-based neural SLT model	Gloss+Text
De Coster et.al[29]	RWTH-PHOENIX-2014	Pretrained BERT network	BLEU 22.25	Advancements in feature extraction and architectural enhancements can significantly improve model quality.	Gloss+Text
Gómez et.al[30]	RWTH-PHOENIX-2014-T	Provide word dependence tags to the model to make it syntax-aware and to improve the discriminative ability of the embeddings supplied to the encoder	BLEU 53.52	Minor deterioration in METEOR metric. Potential improvements can be explored in text-to-gloss translation models.	Text to gloss

Table II summaries of various methods are provided. Stoll et al. used an Encoder and Decoder (NMT) approach on RWTH-PHOENIX-2014-T, facing challenges in conveying spatiotemporal syntax in text. Camgoz et al. employed a Multichannel Transformer on the same dataset, dealing with translation loss due to overfitting. Saunders et al. utilized Sequence-to-sequence learning with CTC on RWTH-PHOENIX-2014 to enhance networks' understanding of sign articulators. Zheng et al. introduced a Dynamic Graph-Based Neural SLT on RWTH-PHOENIX-2014, suggesting future work with a simplified graph-based model. De Coster et al. utilized a Pretrained BERT network on RWTH-PHOENIX-2014, emphasizing improvements in feature extraction and architecture. Gómez et al. enhanced syntax-awareness with word dependence tags on RWTH-PHOENIX-2014-T, exploring potential improvements in text-to-gloss translation

2.3 Pose Estimation Methods: Enhancing Sign Language Video Sequences

In this section, diverse strategies for extracting crucial points from human poses, pivotal for processing Sign Language (SL) videos, are explored.

G. Rogez et al. [31] pioneered a localization classification-regression network for human pose detection using the MPII human pose dataset, overcoming challenges like occlusion and boundary constraints. Chaitanya Ahuja and Louis-Philippe Morenci [32] introduced JL2P, employing an encoder-decoder system validated with the KIT Motion Language Dataset. It generated phrase animations evaluated by average position error (APE). Masoud Zadghorban and Manoochehr Nahvi [33] devised a vision-based method for Persian SL films, employing hand tracking and shape features for sign word extraction.

Researchers [34] employed graph neural networks using OpenPose and HRnet, leveraging spatiotemporal features for SL. Key body points were extracted for face, left hand, and right hand, enhancing prediction accuracy. Razieh Rastgoo et al. [35] developed a cascaded model incorporating deep learning methods like SSD, CNN, and LSTM for SL recognition,

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

achieving superior hand sign recognition results on the isoGD dataset. To mitigate mistakes in sign, pose sequence generation (G2P), a vector-quantized diffusion algorithm was proposed [38]. The Pose-VQVAE model represented posture sequences through latent codes, enhancing the quality of pose sequences.

A STFE-Net was developed [39] combining spatial and temporal features for Chinese Sign Language Translation (CSLT). The model, integrating Transformer-based relative position encoding, demonstrated improved BLEU scores by incorporating spatial and temporal features.

Table III provides an overview of these diverse methods, illustrating advancements in generating intermediate pose sequences essential for SL video creation.

TABLE III: SUMMARY OF THE AVAILABLE APPROACHES FOR POSE ESTIMATION

Approach	Dataset	Methodology	Metrics	Limitations/ Future scope	Output
Razieh Rastgoo et al.[35]	NYU, First-Person, and RKS-PERSIAN SIGN	SSD, CNN, and LSTM	Accuracy 99.8	-	Hand action recognition
Nicati Camoz et.al[36]	RWTH-PHOENIX-Weather	A pretrained 2D hand pose estimator	BLEU 12.18	-	Pose
Pan Xie et.al[38]	RWTH-PHOENIX-Weather	Pose vector-quantized diffusion (PoseVQ-Diffusion) model	WER 78.21 BLEU 16.03	Challenging training process; Focus on pose improvement	Pose
Jiwei Hu et.al[39]	RWTH-PHOENIX-Weather 2014T and Chinese SL	STFE-Net	Accuracy 95.7	Real-time performance and effective key frame extraction methods	Pose

This table outlines various studies in Sign Language (SL) processing. Razieh Rastgoo et al. employed SSD, CNN, and LSTM models on diverse datasets for Hand Action Recognition, achieving 99.8% accuracy. Necati Camoz et al. utilized a pretrained 2D Hand Pose Estimator on the RWTH-PHOENIX-Weather dataset to generate Pose sequences, measured by BLEU (12.18). Pan Xie et al. developed a Pose Vector-Quantized Diffusion model, facing training challenges but achieving improvements in pose accuracy (WER 78.21, BLEU 16.03). Jiwei Hu et al. applied STFE-Net on multiple datasets, focusing on real-time performance and effective key frame extraction methods, attaining an accuracy of 95.7%.

Another method called Text2sign employs a generative adversarial network (GAN) to generate SL videos from spoken language sentences.

2.4 Video generation with generative models

In this section, the generative models that have been proposed to generate synthetic videos are studied.

The authors in [47] tackled large-scale SLP using a method that produces smooth signs, while SIGNGAN is a position-conditioned human synthesis model that creates realistic SL movies from a skeletal pose while scaling to unrestricted domains of discourse. It uses a novel frame selection network (FS-NET) to enhance its ability to temporally match interpolated dictionary signs to continuous signing sequences, allowing for sign coarticulation. Table IV shows a summary of the approaches to video sequences from text and a video dataset

Video Generation with Generative Models

In this section, we delve into the innovative approaches proposed for generating synthetic sign language (SL) videos utilizing generative models.

Stoll et al. [40] S employed Neural Machine Translation (NMT) and a Generative Adversarial Network (GAN) to convert German text to SL videos. They utilized a motion graph to convert gloss into poses, combining it with real pose images as label data for generating synthetic videos. Although they used the PHOENIX-14T SLT dataset, the resulting videos lacked expressiveness.

Saunders [41] developed SIGNGAN system to transform spoken languages into photorealistic SL videos. They utilized a transformer framework with a mixture density network (MDN) formulation to manage the translation process. Quality evaluation was done using SSIM and Fréchet inception distance (FID) metrics on the PHOENIX-14T dataset.

Dongxu Wei et al. [42] introduced GAC-GAN, an approach for appearance-controllable human video motion transfer. This method supported controllable human video motion transmission with superior video quality, thanks to a general-purpose appearance synthesis approach.

Ian J. Goodfellow et al. [43] proposed GANs for generating data samples. Combining generators and discriminators, the models generate realistic data, offering a reference for the deaf community's learning process when tuned with SL datasets.

Doyeon Kim et al. [44] presented a model for generating authentic video based on given text descriptions. By integrating extra label information, this method enhanced the quality of generated videos.

Ventura et al. [45] proposed EDN approach was employed to create animated signer videos from a series of keypoints. OpenPose was utilized to extract keypoints from the original video, training a GAN to generate each frame of the video.

Authors in [46] Annotated subsets of the PHOENIX-14T dataset were created with varying intensities. These subsets were used to train cutting-edge transformer models, improving the production of expressive SL.

Authors in [47] address large-scale Sign Language Processing (SLP), this approach focused on creating smooth signs. Using SIGNGAN as a basis, it incorporated a frame selection network (FS-NET) for enhanced temporal synchronization, allowing for

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

sign coarticulation in continuous signing sequences. Table IV Summary of the approaches for forming video sequences from text and video datasets

TABLE IV: SUMMARY OF THE APPROACHES FOR FORMING VIDEO SEQUENCES FROM TEXT AND VIDEO DATASETS

Approach	**Dataset**	**Methodology**	**Metrics**	**Limitations/Future scope**	**Output**
Stoll et al.[40]	RWTH-PHOENIX-Weather	NMT +GANs	SSIM PSNR MSE 0.9, 23.56, 303.0417	The expressiveness of the video generated is less accurate	Sign video sequences
Ventura Lucas et.al[45]	English to ASL	GANs	Accuracy 91.6% MOS 2.58 BLEU 12.38	The hands were not accurately synthesized	Sign video
Inan Mert et.al[46]	RWTH-PHOENIX-Weather	Transformer with intensification	BLEU -26.01 ROGUE 24.98 BS-	The absence of spatial and temporal context in the automatic back-translation evaluation was one of the study's limitations	Sign video
Saunders Ben et.al[47]	Meine DGS (mDGS)	Frame selection networks	BLEU -21.10 ROGUE-42.57	In opposition to spoken language The techniques, architectures, and datasets for NMT must be improved In future research, the authors will concentrate on spatial coarticulation between dictionary signs	Sign video sequences

These methods highlight the evolving landscape of generative models in the realm of sign language video synthesis

3.SL recognition:

The best-performing and most extensively used techniques from the literature are discussed in this part to help readers attain a better understanding of how different automatic SLR approaches behave. The chosen methods cover all of the method types that have been put forth thus far. The subsequent quantitative comparison evaluation makes it easier to provide insightful information for each SLR methodology.

The simultaneous alignment and recognition challenges were addressed by Camgoz et al. [48] using a DNN-based method known as "sequence-to-sequence" learning. In particular, the overall issue is broken down into a number of specialized systems known as SubUNets. The frames of the video are handled independently by each SubUNet. Their model adopts a 2D CNN-LSTM architecture and LSTM-CTC in place of HMM. To complete the task at hand, the general objective is to model the spatiotemporal interactions among these SubUNets. More particularly, the SubUNets enable the system to include domain-specific expert knowledge about appropriate intermediate representations. They also enable the implicit transfer of learning between several tasks that are related to one another. Cui et al. [49] offered a model that incorporates an additional temporal module (TConvs) following the feature extractor, which contrasts with conventional 2D CNN-based approaches that use HMMs (such as GoogLeNet). Two 1D CNN layers and two max pooling layers make up the TConv module. It is intended to collect the fine-grained relationships that occur within a gloss (intragloss dependencies) between adjacent frames into small per-window feature vectors. The representations of the intermediate segments approximate the typical length of a gloss. To capture the context information between gloss segments, bidirectional RNNs are then deployed. To utilize the expressive power of DNN models for sparse data, the entire architecture is trained in an iterative manner. The proposed methodology consists of a feature extractor, an alignment module, and an auxiliary classifier. The feature extractor extracts local visual information from a given temporal receptive field and takes an image sequence to abstract framewise features. The alignment model [50] and auxiliary classifier identify visual features. During training, two auxiliary losses are used: a visual enhancement loss, which aligns visual characteristics and the target sequence, and a visual alignment loss, which arranges the short-term visual predictions and long-term context predictions.

The ASL signs made by a deaf person were non intrusively recorded by DeepASL in [51], which is a light-based sensor that can extract skeletal joint information from fingers, palms, and forearms using Leap Motion. It uses its domain knowledge of ASL and a hierarchical bidirectional deep RNN (HB-RNN) to accurately simulate the spatial structure and temporal dynamics of the extracted ASL features. DeepASL uses a probabilistic framework based on the Connectionist Temporal Classification (CTC) [52], a tool that enables users to translate entire sentences into individual words and allows DeepASL to perform the task of compiling all possible ASL sentences without pausing between neighboring signs. Few SL references for the regional

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

Marathi language have been found.The authors in [53] were successful in achieving an accuracy of 99.28% in terms of recognizing the representations of Marathi letters with SL using CNNs. Figure 3 shows the SL generation techniques and methods that are reviewed in this study.

4. Sign Language generation approaches

Sign Language (SL) generation involves converting text or spoken language into sign language, allowing communication with deaf or hard-of-hearing individuals. Several approaches are employed for Sign Language generation:

1. Rule-Based Approaches:

Linguistic Rules: Use linguistic rules to convert grammatical structures and vocabulary from the source language to sign language.

Lexical Semantics: Map words to their corresponding signs using a dictionary or lexicon.

Syntactic Rules: Apply rules for sentence structure and word order in sign language.

2. Data-Driven Approaches:

Corpus-Based Learning: Utilize large datasets of sign language videos or motion capture data to train machine learning models. Deep learning techniques, such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, are often employed.

Gesture Recognition: Use computer vision techniques to recognize gestures from video input and translate them into sign language.

3. Avatar-Based Approaches:

Virtual Avatars: Create computer-generated avatars capable of signing. These avatars are controlled by algorithms that convert text input into sign language gestures.

Motion Capture: Use motion capture technology to record human sign language gestures and transfer these movements to virtual avatars.

4. Hybrid Approaches:

Combination of Rule-Based and Data-Driven Methods: Use rule-based methods for grammatical and syntactic aspects, while employing data-driven techniques for capturing natural signing gestures and variations.

Rule Learning from Data: Extract linguistic rules from large datasets of sign language utterances, combining the strengths of both approaches.

5. Sign Language Animation Software:

Animation Tools: Use specialized software that allows animators to manually create sign language animations based on provided scripts. Animators can control the avatar's movements, facial expressions, and body language.

Pre-recorded Libraries: Use pre-recorded sign language libraries, where a set of standard signs and expressions are captured and can be combined to form sentences dynamically.

6. Human-in-the-Loop Approaches:

Crowdsourcing: Involve human signers to generate sign language content. Crowdsourced platforms allow deaf or hard-of-hearing individuals to contribute sign language translations and interpretations.

Human Verification: Use human experts to verify and improve the quality of machine-generated sign language outputs, ensuring accuracy and cultural appropriateness. These approaches can be combined or tailored to specific applications, such as educational software, communication devices, or online platforms, to provide effective and natural sign language communication experiences. Figure 3 summarizes the SL approaches.

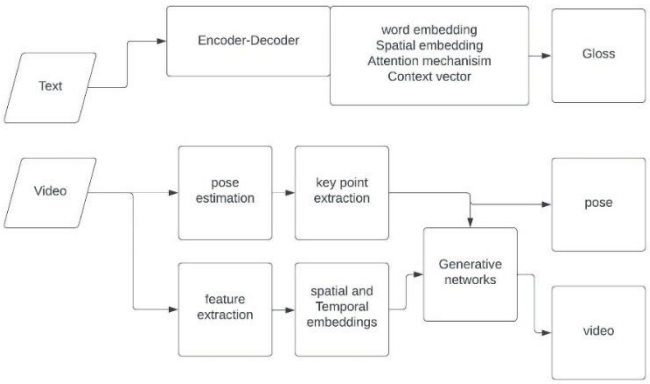

Figure. 3. Approaches for SL generation

5. SL datasets

This analysis underscores the various challenges in sign language video synthesis, highlighting the need for improvements in accuracy, expressiveness, and understanding of spatial and temporal nuances within sign languages. Addressing these challenges is pivotal for advancing the field and enhancing the quality of synthesized sign language videos. In this section, we delve into various datasets designed primarily for Sign Language (SL) recognition and translation. The use of diverse datasets demonstrates the adaptability of systems. This approach contrasts with avatar and motion capture methods, where avatars require specialized animation skills and motion capture technology proves unscalable and costly. Among these datasets is the Purdue SL dataset [54], encompassing 2576 videos featuring 14 native signers. Notably, this database includes motion traces, aiding parameter estimation related to motion. Another valuable resource is the SIGNUM [55] database, capturing both discrete and continuous sign expressions. Utilizing a vision-based method, this corpus is saved as a series of images, facilitating easy access to specific frames.

DGS, comprising 450 fundamental signs representing diverse word types, formed the basis for 780 sentences varying from two to eleven signs each. Recorded by 25 native signers, this corpus includes all 450 fundamental signals and all 780 words. The publicly accessible PSL Kinect 30 dataset offers 30 classes of word-level data derived from a single-depth Kinect camera [56]. Gesture recognition studies benefit from the IIITA-ROBITA ISL [57] Gesture Database, featuring 23 distinct gestures recorded at 30 frames per second and 320x240 pixels using various lighting conditions.

An essential contribution is the CSL-Daily dataset [58], providing gloss-level annotations and spoken language

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

translations, focusing on everyday scenarios where SLT finds applications. Additionally, datasets for Chinese SL, captured using Kinect 2.0, offer isolated and continuous SL recognition data, encompassing signer's 3D joints, RGB videos, and depth videos. The British SL (BSL) Corpus [59] comprises videos by deaf individuals using BSL, supplemented by textual explanations and signers' biographical details.

The ASLG-PC12 [60] project proposed a rule-based methodology for creating a parallel corpus of written English texts and American SL gloss. The RWTH-PHOENIX-Weather dataset [61], collected over two years, includes seven signers' videos annotated with American SL videos recorded by 1-6 native ASL signers, featuring gloves and providing front and above views at 30 fps and a 1600x1200 pixel resolution. ASSLLVD [62] boasts 73,300 American SL videos recorded by 1-6 native ASL signers, featuring gloves and providing front and above views at 30 fps and a 1600x1200 pixel resolution.

For continuous SL recognition, the LSFB-continuous dataset [63] offers videos at 720x576 pixels and 5 fps, allowing extraction of facial and pose landmarks. The WLASL word-level ASL dataset [64] contains 34,404 video samples of 3126 glosses, each lasting 2.41 seconds. The INCLUDE [65] dataset for the Indian subcontinent comprises 4287 videos across categories like adjectives, animals, colors, and transport, each with a duration of 2-4 seconds and a resolution of 1920x1080 at 2 FPS. Lastly, How2Sign [66] stands out as a multimodal and multiview dataset, providing a rich resource for SL research. gloss and central hand/palm points, aiding pattern recognition.

TABLE V: SUMMARY OF THE AVAILABLE SL DATASETS FOR SIGN RECOGNITION AND GENERATION

Dataset	SL	Type	Link to dataset
Purdue ASL [54]	ASL	Videos+Glosses	Purdue ASL Database (ohio-state.edu)
SIGNUM [55]	German SL	Videos+Glosses	SIGNUM Database (uni-muenchen.de)
PSL Kinect [56]	Polish sign Language	Videos, depth from Kinect camera	vision.kia.prz.edu.pl/dynamickinect.php
IIITA-ROBITA [57]	ISL	Videos	IIITA-Robotics
CSL-Daily [58]	Chinese SL (CSL)	Videos +Glosses+Text	CSL-Daily: A Continuous Chinese SL Dataset (ustc.edu.cn)
BSL [59]	BSL	Videos+Text	Home - BSL Corpus Project
ASLG-PC-12 [60]	ASL	Text+Glosses	English-ASL Gloss Parallel Corpus 2012: ASLG-PC12 | Dr. Achraf Othman
RWTH - Phoenix-Weather dataset [61]	DGS	Text+Glosses +Videos	Research - RWTH-PHOENIX-Weather Database (rwth-aachen.de)
ASSLLVD [62]	• AMERICAN SL LEXICON VIDEO DATASET (ASLLVD)	Videos	American SL Linguistic Research Project (ASLLRP) Sign Bank (rutgers.edu)
LSFB [63]	French Belgian SLs	Videos	LSFB dataset (unamur.be)
WASL [64]	Word-Level American SL	Videos	GitHub - dxli94/WLASL: WACV 2020 "Word-level Deep Sign Language Recognition from Video: A New Large-scale Dataset and Methods Comparison"
INCLUDE [65]	Indian SL	Videos	INCLUDE: A Large Scale Dataset for Indian SL Recognition | Zenodo

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

How2Sign [66]	American SL	Videos+Text +Glosses+Speech	How2Sign Dataset

5. Evaluation metrics for text and video processing

The popular metrics for evaluating how well translation is performed by machine-based algorithms are BLEU [67], the metric for the evaluation of translation with explicit ordering (METEOR) [68], and recall-oriented understudy for gisting evaluation (ROGUE) [69]. BLEU is based on n-gram precision to measure translation quality.

$$precision\ n-gram = \frac{predicted\ n-gram\ precsion}{total\ predicted\ n-gram\ precision} \quad (1)$$

$$Geometric\ Average\ precision(GAP) = \prod_{n=1}^{N} p_n^{w_n} \quad (2)$$

$$Brevity\ penalty = \begin{cases} 1 & if\ c > r \\ e^{\left(1-\frac{r}{c}\right)} & if\ c \leq r \end{cases} \quad (3)$$

Where,

c is *predicted length = number of words in the predicted sentence* and

r is *target length = number of words in the target sentence*

Bleu score=GAP X Brevity penalty

It uses a modified n-gram precision measure to clip the number of matches for a given word. It is solely based on precision and does not cater to recall, semantics, or word order. METEOR is another metric that correlates machine- and human-generated references. METEOR is an autonomous measure for evaluating MT performance based on unigram matching between machine-generated and human-generated translation reference translations. It can be expanded to add more complex matching strategies, and it generates a score after a match is discovered. It is evaluated by calculating the correlation between the metric scores and the human translation quality evaluations.

$$P=\frac{m}{w_t} \quad (4)$$

$$R=\frac{m}{w_r} \quad (5)$$

$$F_{mean} = \frac{10PR}{R+9P} \quad (6)$$

$$p = 0.5\left(\frac{c}{u_m}\right)^3 \quad (7)$$

$$M = F_{mean}(1-p) \quad (8)$$

Where ,P: Precision
R: Recall
m: Number of unigrams in the candidate translation also found in reference
w_t: Number of unigrams in candidate translation
w_r: Number of unigrams in reference translation
c: Number of chunks in candidate
U_m: Unigrams in candidate

The ROGUE metric that compares unigrams is called ROGUE-1; it counts the number of matching words in the reference and machine-generated outputs for short sentences. For sentences, ROGUE-2 may be used as it considers bigrams. Another ROGUE variant is ROGUE-L, which does not consider bigrams but treats each summary as a sequence of words and looks for the longest common sequences. This is a sequence that appears in the same relative order. ROGUE-L captures sentence structure more accurately.we can compute ROGUE-1 ,ROGUE-2 and ROGUE-L by

$$Recall=\frac{Total\ number\ of\ overlapping\ unigrams}{Total\ number\ of\ unigrams\ in\ the\ reference\ summary} \quad (9)$$

$$Recall = \frac{Total\ number\ of\ overlapping\ bigrams}{Total\ number\ of\ bigrams\ in\ the\ reference\ summary} \quad (10)$$

$$F=\frac{(1+b^2)R_{lcs}P_{lcs}}{R_{lcs}+b^2_{\square}P_{lcs}} \quad (11)$$

where,

lcs is the longest common subsequence

b is the controlling parameter for precision

An automated evaluation method termed BERTScore [70] is used to gauge how well text generation algorithms work. Instead of focusing on computing token-level syntactical similarities, as done by other prominent techniques, BERT Score calculates the semantic similarity between the reference and hypothesis tokens. It performs better than human judgments in picture captioning and MT tests.

The quality of each frame can subsequently be collected and aggregated over time to determine the overall quality of a video sequence. Certain video quality evaluation metrics (such as the PSNR or SSIM) are simply picture quality models, with an output calculated for each frame of a video sequence. The sequence is evaluated by comparing every frame of the

degraded and original video signals. The PSNR [71] is the most extensively used image quality metric.

However, because of the human vision system's complicated, extremely nonlinear behavior, PSNR values do not correspond well with reported picture quality levels. The SSIM [72] is a perception-based model that integrates critical perceptual phenomena such as luminance and contrasts masking terms, and it considers image deterioration to be an observed shift in structural information. The main issue is that the reference is subjected to various tokenization and normalization schemes. To address this, a new tool, SACREBLEU1, was proposed [73]. WER has also been utilized in [74]. Furthermore, the edit distance *d (t; r)* (quantities of insertions, deletions, and substitutions) between both the produced translation t and one predetermined reference translation r can be calculated. Because the underlying data and algorithm are always the same, the edit distance has the substantial advantage of being automatically computable. As a consequence, its results are both affordable and reproducible.

$$\text{WER}=\frac{Insertion+Deletion+substitution}{Total\ number\ of\ words\ in\ reference} \quad (12)$$

The minimum opinion score MOS [75] is a commonly used measure for evaluating video, audio, and audiovisual quality, but it is not limited to those modalities.

Challenges: Translating spoken or written language into sign language presents several challenges due to the complexity and richness of sign languages. Here are some of the key challenges faced in sign language translation:

1. Linguistic Complexity:

Sign languages have complex grammatical structures that differ significantly from spoken languages. Translators must understand these differences and accurately convey the intended meaning.

Sign languages often rely on facial expressions, body movements, and spatial referencing, making translation more intricate.

2. Cultural and Regional Variations:

Sign languages vary across cultures and regions. There are numerous sign language dialects and regional variations, requiring translators to be aware of these differences to ensure accurate communication.

3. Ambiguity and Context:

Signs can have multiple meanings depending on the context. Translators need to consider the surrounding context to choose the correct interpretation.

Ambiguity arises when a sign or gesture could have multiple interpretations, leading to potential misunderstandings.

4. Limited Lexical Resources:

Sign languages often have a smaller vocabulary compared to spoken languages, especially in specialized fields. Finding appropriate signs for technical or specific terms can be challenging.

The development of sign language dictionaries and lexicons is an ongoing process to expand resources for translators.

5. Facial Expressions and Non-Manual Signals:

Facial expressions, tone, and other non-manual signals are integral to sign language communication. Translating these nuanced expressions into written or spoken language can be challenging as they carry essential meaning in sign languages.

6. Real-time Translation:

In real-time situations such as live events or interpreting services, sign language translators must work quickly and accurately, making it a high-pressure task.

Latency in translation systems can disrupt the natural flow of conversation, impacting communication.

7. Lack of Standardization:

Unlike many spoken languages, sign languages often lack standardization. Variations in signs and grammar among different communities can complicate the translation process, especially in global or cross-cultural contexts.

8. Accessibility and Technology:

Access to sign language interpretation services can be limited in certain regions or situations, restricting the availability of accurate translations.

Developing reliable and user-friendly sign language translation technologies, such as apps or devices, is an ongoing challenge.

9. Ethical and Cultural Considerations:

Translators must be culturally sensitive, understanding the social and cultural nuances of the deaf community.

Ethical considerations include respecting the privacy and preferences of individuals when translating sensitive or personal information.

Addressing these challenges requires collaboration between linguists, technologists, and the deaf community, emphasizing the importance of cultural understanding and linguistic expertise in sign language translation efforts.

Future scope: sign language generation holds significant potential for advancements and applications in various fields. Here are some areas where we can expect growth and innovation in sign language generation:

1. Education and E-Learning:

Sign language generation can enhance educational experiences for deaf and hard-of-hearing students. Interactive e-learning platforms with real-time sign language translation can make educational content more accessible.

Virtual classrooms and online courses can integrate sign language interpretation, promoting inclusive education.

2. Communication Devices and Apps:

Sign language generation can be integrated into smartphones, tablets, and other communication devices, allowing deaf individuals to communicate seamlessly through sign language.

Mobile apps with sign language translation capabilities can facilitate communication in various settings, such as healthcare, customer service, and social interactions.

3. Accessibility Services:

Sign language generation services can be integrated into public spaces, transportation systems, and government offices to enhance accessibility for the deaf community.

Sign language interpretation services can be made available in real-time during live events, conferences, and public announcements, ensuring inclusivity.

4. Human-Computer Interaction:

Sign language generation can be incorporated into human-computer interfaces, enabling natural and intuitive interactions with computers and smart devices.

5.Research and Development:

Continued research in sign language linguistics and technology can lead to the development of more accurate and expressive sign language generation systems.

Advancements in machine learning, computer vision, and natural language processing will contribute to the improvement of sign language generation algorithms.

6.Global Collaboration and Standardization:

Collaboration between researchers, developers, and the deaf community on a global scale can lead to the standardization of sign language generation systems, ensuring consistency and interoperability.

International efforts can address regional variations in sign languages, making sign language generation accessible and relevant worldwide.

The future of sign language generation lies in the ongoing development of innovative technologies, increased awareness of accessibility needs, and active involvement of the deaf community in the design and implementation of these systems. As these advancements progress, sign language generation will play a vital role in promoting inclusivity and bridging communication gaps for deaf individuals

.

References

[1] Indian sign Language Dictionary. [Online]. Available: Indian Sign Language Research and Training Center (ISLRTC), Government of India [Accessed: September 3, 2021].

.

[2] Su, Hung-Yu, and Chung-Hsien Wu. "Improving structural statistical machine translation for sign language with small corpus using thematic role templates as translation memory." IEEE Transactions on Audio, Speech, and Language Processing vol.17, no. 7, pp.1305-1315, sep. 2009,doi.org/10.1109/TASL.2009.2016234

[3] Chiu, Y.H., Wu, C.H., Su, H.Y. and Cheng, C.J "Joint optimization of word alignment and epenthesis generation for Chinese to Taiwanese sign synthesis." IEEE Transactions on Pattern Analysis and Machine Intelligence vol.29,no.1,pp.28-39,Jan 2007,doi.org/10.1109/TPAMI.2007.250597

[4] Bungeroth, J., & Ney, H. "Statistical sign language translation." In *sign-lang@ LREC* European Language Resources Association (ELRA), *2004*, pp. 105-108.

[5] Marshall, I, and É. Sáfár. "Extraction of semantic representations from syntactic SMU link grammar linkages." Proc of Recent Advances in Natural Language Processing, 2001, pp.154-159,.

[6] Van Zijl, Lynette, and Dean Barker. "South African sign language machine translation system." Proc. of the 2nd international conference on Computer graphics, virtual Reality, visualisation and interaction in Africa, 2003, pp. 49-52.

[7] Othman, A. and Jemni, M., "Statistical sign language machine translation: from English written text to American sign language gloss." , 2011, doi.org/10.48550/arXiv.1112.0168.

[8] Dasgupta. T. and Basu. A. "Prototype machine translation system from text-to-Indian sign language." Proc. 13th international conference on Intelligent user interfaces, 2008, pp. 313-316.

[9] Fu. B, Ye. P, Zhang. L, Yu. P, Hu. C, Shi. X. and Chen. Y, "A Token-Level Contrastive Framework for Sign Language Translation." Proc ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1-5.

[10] Bahdanau. D, Cho. K and Bengio. Y, "Neural machine translation by jointly learning to align and translate." 2014, doi.org/10.48550/arXiv.1409.0473

[11] Ibrahim, N. B., Zayed, H. H., & Selim, M. M, "Advances, challenges and opportunities in continuous sign language recognition." Journal of Engineering and Applied Sciences 15, no. 5, 2020 pp.1205-1227.

[12] Zheng. J, Zhao. Z, Chen. M, Chen. J, Wu. C, Chen. Y, & Tong. Y. "An improved sign language translation model with explainable adaptations for processing long sign sentences." *Computational Intelligence and Neuroscience* 2020.

[13] Zhou H, Zhou W, Zhou Y, Li H. "Spatial-temporal multi-cue network for continuous sign language recognition." Proc. of the AAAI Conference on Artificial Intelligence Vol. 34, No. 07, Apr. 2020. pp 13009-13016. doi.org/10.1609/aaai.v34i07.7001

[14] Chen Y, Wei F, Sun X, Wu Z, Lin S. "A simple multi-modality transfer learning baseline for sign language translation." Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 5120-5130.

[15] Camgoz NC, Koller O, Hadfield S, Bowden R. "Sign language transformers: Joint end-to-end sign language recognition and translation." Proc. of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10023-10033

[16] Orbay A, Akarun L. "Neural sign language translation by learning tokenization." Proc. 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) Nov 2020, pp. 222-228. IEEE.

[17] Xiao Q, Chang X, Zhang X, Liu X. "Multi-information spatial–temporal LSTM fusion continuous sign language neural machine translation". Ieee Access. Nov 2020; 8:216718-28.

[18] Yin K, Read J. "Attention is all you sign: sign language translation with transformers." In Sign Language Recognition, Translation and Production (SLRTP) Workshop-Extended Abstracts 2020 Aug Vol. 4.

[19] Mohamed A, Hefny H. "A deep learning approach for gloss sign language translation using transformer." Journal of Computing and Communication. 2022 Aug 15;1(2):1-8. doi.org/10.21608/jocc.2022.254979

[20] Angelova G, Avramidis E, Möller S. "Using neural machine translation methods for sign language translation." Proc. of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. May 2022, pp. 273-284.

[21] Núñez-Marcos A, Perez-de-Viñaspre O, Labaka G. "A survey on Sign Language machine translation." Expert Systems with Applications. Oct 2022, 13:118993.

[22] Sugandhi, Kumar P, Kaur S. "Sign language generation system based on Indian sign language grammar." ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP). Apr 2020, 19(4):1-26. doi.org/10.1145/3384202

[23] Oliveira T, Escudeiro P, Escudeiro N, Rocha E, Barbosa FM. "Automatic sign language translation to improve communication." Proc. IEEE Global Engineering Education Conference (EDUCON) Apr 2019, pp. 937-942, doi.org/10.1109/EDUCON.2019.8725244

[24] Zheng J, Zhao Z, Chen M, Chen J, Wu C, Chen Y, Shi X, Tong Y. "An improved sign language translation model with explainable adaptations for processing long sign sentences." Computational Intelligence and Neuroscience. Oct 2020, doi.org/10.1155/2020/8816125

[25] Luqman H. "ArabSign: A Multi-modality Dataset and Benchmark for Continuous Arabic Sign Language Recognition." Proc. IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG), Jan. 2023, pp. 1-8, doi.org/10.1109/FG57933.2023.10042720

[26] Bhagwat SR, Bhavsar RP, Pawar BV. "Translation from simple Marathi sentences to Indian sign language using phrase-based approach." Proc.International Conference on Emerging Smart Computing and Informatics, 2021, pp. 367-373, doi.org/10.1109/ESCI50559.2021.9396900

[27] Mazhari A, Esfandiari P, Taheri A. "Teaching Iranian Sign Language via a Virtual Reality-Based Game." Proc. 10th RSI International Conference on Robotics and Mechatronics (ICRoM), Nov. 2022, pp. 146-151, doi.org/10.1109/ICRoM57054.2022.10025347

[28] Zheng J, Li S, Tan C, Wu C, Chen Y, Li SZ. "Leveraging Graph-based Cross-modal Information Fusion for Neural Sign Language Translation". arXiv preprint arXiv:2211.00526. Nov. 2022.

[29] De Coster M, D'Oosterlinck K, Pizurica M, Rabaey P, Verlinden S, Van Herreweghe M, Dambre J. "Frozen pretrained transformers for neural sign language translation." Proc.18th Biennial Machine Translation Summit

(MT Summit 2021) 2021, pp. 88-97. Association for Machine Translation in the Americas.

[30] Gómez SE, McGill E, Saggion H. "Syntax-aware transformers for neural machine translation: The case of text to sign gloss translation." Proc. of the 14th Workshop on Building and Using Comparable Corpora (BUCC 2021) Sep. 2021 pp. 18-27.

[31] Rogez G, Weinzaepfel P, Schmid C. "Lcr-net++: Multi-person 2d and 3d pose detection in natural images." IEEE transactions on pattern analysis and machine intelligence. Jan. 201913;42(5):1146-61. doi.org/10.1109/TPAMI.2019.2892985

[32] Ahuja C, Morency LP. "Language2pose: Natural language grounded pose forecasting." Proc. 2019 International Conference on 3D Vision (3DV) Sep. 2019(pp. 719-728). IEEE. https://doi.org/10.1109/3DV.2019.00084

[33] Zadghorban M, Nahvi M. "An algorithm on sign words extraction and recognition of continuous Persian sign language based on motion and shape features of hands." Pattern Analysis and Applications. May. 2018, ;21:323-35,doi.org/10.1007/s10044-016-0579-2

[34] Kan J, Hu K, Hagenbuchner M, Tsoi AC, Bennamoun M, Wang Z. Sign language translation with hierarchical spatio-temporal graph neural network. InProceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision 2022 (pp. 3367-3376).

[35] Rastgoo R, Kiani K, Escalera S. "Hand sign language recognition using multi-view hand skeleton." Expert Systems with Applications, Jul. 2020 ;150:113336

[36] Saunders B, Camgoz NC, Bowden R. "Everybody sign now Translating spoken language to photo realistic sign language video." arXiv preprint arXiv:2011.09846. Nov. 2020. doi.org/10.48550/arXiv.2011.09846

[37] Toshev A, Szegedy C. "Deeppose Human pose estimation via deep neural networks." Proc. of the IEEE conference on computer vision and pattern recognition 2014 pp. 1653-1660. `

[38] Xie P, Zhang Q, Li Z, Tang H, Du Y, Hu X. "Vector quantized diffusion model with codeunet for text-to-sign pose sequences generation." arXiv preprint arXiv:2208.09141. Aug. 2022.

[39] Hu J, Liu Y, Lam KM, Lou P. "STFE-Net: A Spatial-Temporal Feature Extraction Network for Continuous Sign Language Translation." IEEE Access. Jan. 2023, doi.org/10.1109/ACCESS.2023.3234743

[40] Stoll S, Camgoz NC, Hadfield S, Bowden R. "Text2Sign: towards sign language production using neural machine translation and generative adversarial networks." International Journal of Computer Vision. Apr. 2020;128(4):891-908.

[41] Saunders B, Camgoz NC, Bowden R. Everybody sign now: "Translating spoken language to photo realistic sign language video." arXiv preprint arXiv:2011.09846. Nov. 2020.

[42] Wei D, Xu X, Shen H, Huang K. GAC-GAN: A general method for appearance-controllable human video motion transfer. IEEE Transactions on Multimedia. Jul. 2020 24;23:2457-70. doi.org/10.1109/TMM.2020.3011290

[43] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. "Generative adversarial nets." Advances in neural information processing systems. 2014;27.

[44] Kim D, Joo D, Kim J. Tivgan: Text to image to video generation with step-by-step evolutionary generator. IEEE Access. Aug. 2020;8:153113-22. doi.org/10.1109/ACCESS.2020.3017881

[45] Ventura L, Duarte A, Giró-i-Nieto X. "Can everybody sign now? Exploring sign language video generation from 2D poses." arXiv preprint arXiv:2012.10941. Dec. 2020.

[46] Inan M, Zhong Y, Hassan S, Quandt L, Alikhani M. "Modeling Intensification for Sign Language Generation: A Computational Approach." arXiv preprint arXiv:2203.09679. Mar. 2022.

[47] Saunders B, Camgoz NC, Bowden R. "Signing at scale: Learning to co-articulate signs for large-scale photo-realistic sign language production". Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2022 pp. 5141-5151.

[48] Cihan Camgoz N, Hadfield S, Koller O, Bowden R. "Subunets: End-to-end hand shape and continuous sign language recognition." Proc. of the IEEE international conference on computer vision 2017, pp. 3056-3065.

[49] Cui R, Liu H, Zhang C. "A deep neural framework for continuous sign language recognition by iterative training." IEEE Transactions on Multimedia. Jan 20191;21(7):1880-91.

[50] Min Y, Hao A, Chai X, Chen X. Visual alignment constraint for continuous sign language recognition. InProceedings of the IEEE/CVF International Conference on Computer Vision 2021 pp. 11542-11551.

[51] Fang B, Co J, Zhang M. "Deepasl: Enabling ubiquitous and non-intrusive word and sentence-level sign language translation." Proceedings of the 15th ACM conference on embedded network sensor systems Nov.2017. pp. 1-13.

[52] Nakatani T. "Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration." Proc. Interspeech 2019 (Vol. 2019).

[53] Deshpande A.M, Kalbhor S.R. "Video-based marathi sign language recognition and text conversion using convolutional neural network." Emerging Trends in Electrical, Communications, and Information Technologies: Proc. of ICECIT-2018 Jan.2020 pp. 761-773. Springer Singapore. doi: 10.1007/978-981-13-8942-9_65

[54] Martínez AM, Wilbur RB, Shay R, Kak AC. "Purdue RVL-SLLL ASL database for automatic recognition of American Sign Language." Proc. Fourth IEEE International Conference on Multimodal Interfaces Oct. 2002 pp. 167-172. IEEE.

[55] Von Agris U, Zieren J, Canzler U, Bauer B, Kraiss KF. "Recent developments in visual sign language recognition." Universal Access in the Information Society. Feb. 2008;6:323-62.

[56] Oszust M, Wysocki M. Polish "sign language words recognition with Kinect." Proc.2013 6th International Conference on Human System Interactions (HSI) Jun. 2013 pp. 219-226. IEEE.

[57] Nandy A, Mondal S, Prasad JS, Chakraborty P, Nandi GC. "Recognizing & interpreting Indian sign language gesture for human robot interaction." Proc. International Conference on computer and communication technology (ICCCT) Sep. 2010, pp. 712-717. IEEE.

[58] Zhou H, Zhou W, Qi W, Pu J, Li H. "Improving sign language translation with monolingual data by sign back-translation." Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2021. pp. 1316-1325.

[59] Stamp R, Schembri A, Fenlon J, Rentelis R, Woll B, Cormier K. "Lexical variation and change in British Sign Language." PLoS One. Apr. 2014 ;9(4):e94053.

[60] Othman A, Jemni M. "English-asl gloss parallel corpus 2012: Aslg-pc12." Insign-lang@ LREC 2012 May 2012 (pp. 151-154). European Language Resources Association (ELRA).

[61] Forster J, Schmidt C, Hoyoux T, Koller O, Zelle U, Piater JH, Ney H. "RWTH-PHOENIX-weather: A large vocabulary sign language recognition and translation corpus." InLREC May 2012 (Vol. 9, pp. 3785-3789.

[62] Neidle C, Thangali A, Sclaroff S. "Challenges in development of the american sign language lexicon video dataset (asllvd) corpus." Proc5th workshop on the representation and processing of sign languages: interactions between corpus and Lexicon, LREC May 2012.

[63] Fink J, Frénay B, Meurant L, Cleve A. "LSFB-CONT and LSFB-ISOL: Two new datasets for vision-based sign language recognition." Proc. International Joint Conference on Neural Networks (IJCNN) Jul 2021, pp. 1-8. IEEE.

[64] Li D, Rodriguez C, Yu X, Li H. "Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison." Proc. of the IEEE/CVF winter conference on applications of computer vision 2020. pp. 1459-1469.

[65] Sridhar A, Ganesan RG, Kumar P, Khapra M. "Include: A large scale dataset for indian sign language recognition." Proc. of the 28th ACM international conference on multimedia Oct. 2020, pp. 1366-1375.

[66] Duarte A, Palaskar S, Ventura L, Ghadiyaram D, DeHaan K, Metze F, Torres J, Giro-i-Nieto X. "How2sign: a large-scale multimodal dataset for continuous american sign language." Proc. of the IEEE/CVF conference on computer vision and pattern recognition 2021, pp. 2735-2744.

[67] Papineni K, Roukos S, Ward T, Zhu WJ. "Bleu: a method for automatic evaluation of machine translation." Proc. of the 40th annual meeting of the Association for Computational Linguistics Jul. 2002 pp. 311-318.

[68] Banerjee S, Lavie A. "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments." Proc. of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization Jun. 2005 pp. 65-72.

[69] Lin CY. "Rouge: A package for automatic evaluation of summaries." InText summarization branches out Jul. 2004pp. 74-81.

[70] Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. "Bertscore: Evaluating text generation with bert." arXiv preprint arXiv:1904.09675. Apr. 2019.

[71] Huynh-Thu Q, Ghanbari M. "Scope of validity of PSNR in image/video quality assessment." Electronics letters. 2008 Jun 19;44(13):800-1.

[72] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. "Image quality assessment: from error visibility to structural similarity." IEEE transactions on image processing. Apr. 2004.13;13(4):600-12.

[73] Post M. A call for clarity in reporting BLEU scores. arXiv preprint arXiv:1804.08771. Apr. 2018 .

[74] Nießen S, Och FJ, Leusch G, Ney H. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. InLREC May. 2000.

[75] ITU-T Rec. P.10/G.100 (2017) Vocabulary for performanceG. Casey and E. Lecolinet, "A survey of methods and strategies in character segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 18, no. 7, pp. 690–706, Jul. 1996, doi: 10.1109/34.506792