Machine Learning with Applications 14 (2023) 100504

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

A survey on sign language literature

Marie Alaghband [a,b,*], Hamid Reza Maghroor [b], Ivan Garibay [b]

[a] VaxCare LLC, Orlando, FL, USA
[b] Department of Industrial Engineering and Management Systems, University of Central Florida, Orlando, FL, USA

ARTICLE INFO

Keywords:
Sign language recognition
Gesture recognition
Facial expression recognition
Sign language translation
Artificial intelligence
Machine learning
Deep learning
Hardware-based
Vision-based

ABSTRACT

Individuals with hearing impairment encounter various types and levels of difficulties, highlighting the need for more research to provide effective support. One significant difficulty is communication and interaction with others. Given that these individuals employ sign language as their primary mode of communication, there exists a notable information void among those who can hear in comprehending and interpreting sign language. Consequently, to bridge this gap, the field of sign language research has seen significant growth. In this study, we emphasize the importance of sign language recognition and translation and provide a comprehensive review of relevant research conducted in this field. Our examination encompasses multiple perspectives, including sign language recognition, translation, and the availability of datasets. By exploring these aspects, we aim to contribute to the advancement of sign language literature and its practical applications.

1. Introduction

Language plays a pivotal role in everyday life, serving as a complex system for expressing our personality and facilitating effective communication with others. Through words, gestures, and vocal tones, we interact with people in various contexts, conveying our emotions, desires, and inquiries. Individuals with severe or profound hearing loss naturally rely on sign language as their mode of communication. According to the World Health Organization (WHO), over 5% of the global population, approximately 466 million people, have some form of hearing impairment. By 2050, this number is expected to rise to 900 million, equivalent to one in every ten individuals (World Health Organization, 2023).

The concept of sign language underwent a significant shift during the first half of the 20th century. Initially, it was perceived as a form of imitation of spoken language or a method of visually conveying verbal signals. However, later in the 20th century, it was recognized that sign language is not a direct translation of spoken words. Instead, it employs gestures to convey meaning independently. Different countries or regions have unique sign languages, which are not universally shared. Across the globe, more than 135 distinct sign languages exist, and even countries that share a spoken language, such as Australia, the United States, and England, have their own distinct sign languages (Ai-Media, 2023).

Another fascinating aspect of sign language is its distinct grammar, syntax, and structure. Unlike spoken languages, sign language has a separate syntax. A single sign can represent an alphabet, a word, or even an entire spoken sentence. Furthermore, sign languages, like any other language, undergo development and evolution over time (National Institute on Deafness and Other Communication Disorders (NIDCD), 2023).

Hand gestures and facial expressions, which serve as integral parts of both manual and non-manual signals, are the primary components of sign language. Manual signals involve hand shape, position, location, and movement, to convey meanings through signs. Non-manual signals are produced by other body parts, such as eye gaze and movement, lip patterns, mouthing, body movements, and head orientations.

While manual signals are commonly used to convey words or phrases, non-manual signals are used to convey grammatical and emotional information (Elakkiya, Vijayakumar, & Kumar, 2021; Koller, Zargaran, Ney, & Bowden, 2016, 2018; Kumar, Roy, & Dogra, 2018; Yang & Lee, 2011). Signers utilize both manual and non-manual signals to ensure accurate communication of meaning and emotions, considering the potential for multiple interpretations of a single hand gesture. Non-manual signals, such as facial expressions, provide additional information that complements the manual signals, resulting in a comprehensive and accurate message, thereby avoiding misinterpretation. Therefore, the combination of these signals gives sign language its richness and sophistication (Neiva & Zanchettin, 2018). Both signals are interdependent and incomplete without one another (Adaloglou et al., 2021; Kelly, Reilly Delannoy, Mc Donald, & Markham, 2009).

In this paper, we extend the research conducted in chapters 1 and 2 of the doctoral dissertation titled in "Analysis of Sign Language Facial

* Correspondence to: 3113 Lawton Rd, # 250, Orlando, FL, 32803, USA.
E-mail addresses: marie.alaghband@knights.ucf.edu (M. Alaghband), hmaghroor@ucf.edu (H.R. Maghroor), igaribay@ucf.edu (I. Garibay).

https://doi.org/10.1016/j.mlwa.2023.100504
Received 23 June 2023; Received in revised form 18 August 2023; Accepted 6 October 2023
Available online 24 October 2023

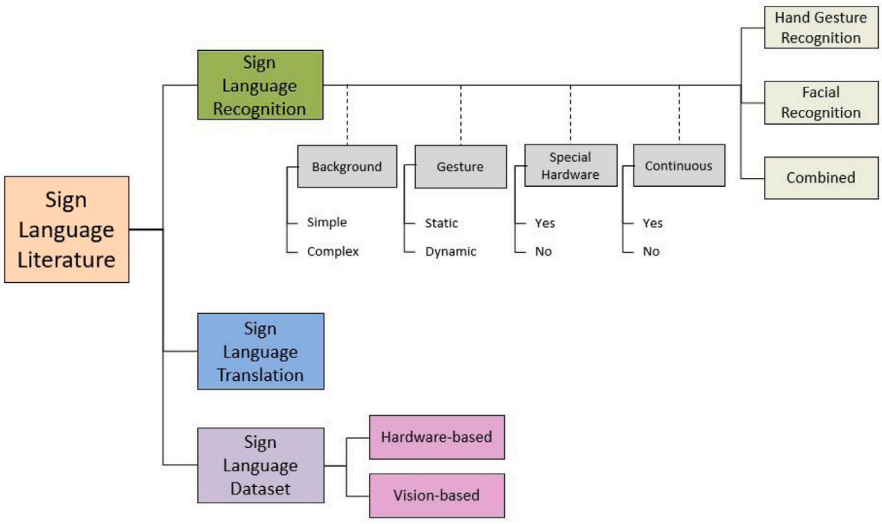

Fig. 1. The diagram presents a comprehensive overview of sign language literature, highlighting its main components. Within the broader scope of sign language literature, three main areas of focus are sign language recognition, sign language translation, and sign language translation. These three components form integral parts of the overall landscape of sign language literature.

Expressions and Deaf Students' Retention Using Machine Learning and Agent-based Modeling" (Alaghband, 2021). The findings and analysis from these chapters serve as the foundation for the current study. This article aims to provide researchers with a better understanding of the conducted research in the field of sign language studies. For additional study, we recommend the following recent surveys on sign language (Adaloglou et al., 2021; Ardiansyah, Hitoyoshi, Halim, Hanafiah, & Wibisurya, 2021; Bahia & Rani, 2023; Elakkiya, 2021; Guo, Lu, & Yao, 2021; Núñez-Marcos, Perez-de Viñaspre, & Labaka, 2022; Oudah, Al-Naji, & Chahl, 2020; Rastgoo, Kiani, & Escalera, 2021b; Sharma, Kumar, Sehgal, & Kaushik, 2022; Wadhawan & Kumar, 2021).

The field of sign language literature encompasses three primary genres: sign language recognition, sign language translation, and sign language datasets. Fig. 1 provides a visual summary representation of the research and studies conducted in this area. Following the structure shown in Fig. 1, this work is organized in three main sections: Section 2 focuses on research related to sign language recognition, Section 3 explores works addressing sign language translation studies and finally, Section 4 enumerates upon listing sign language datasets available in both hardware-based and vision-based categories.

2. Sign language recognition

Sign language recognition poses significant challenges due to the intricate hand gestures, body postures, and facial expressions involved, which often incorporate rapid and complex movements (Jiang et al., 2021). Hand gesture recognition, in particular, is a complex aspect of sign language recognition, characterized by high inter-class similarities, significant intra-class variation, and frequent obstructions in hand morphologies, leading to substantial complexity and variability (Tao, Leu, & Yin, 2018).

Given the complexity of sign language recognition, researchers have approached the problem from various perspectives. As depicted in Fig. 1, sign language recognition, which can be divided in three main categories of hand gesture recognition, facial recognition, and combined recognition methodologies, can be approached through four primary angles: background, gesture, special hardware utilization, and continuity. Different combinations of these components can be employed to address the challenges of three different categories of sign language recognition.

However, these investigations can be categorized into two main types based on the use of specialized hardware. Specifically, there are studies focused on vision-based recognition (without specialized hardware) and hardware-based recognition (utilizing special hardware such as depth cameras, sensors, special gloves, etc.) for sign language recognition (Ansari & Harit, 2016; Bulugu, 2021; Freitas, Peres, Lima, & Barbosa, 2017; Kadous, 2002; Pugeault & Bowden, 2011; Ren, Yuan, & Zhang, 2011; Ronchetti, Quiroga, Estrebou, Lanzarini, & Rosete, 2016).

In the following subsections, we will discuss studies that fall into the aforementioned categories, providing a detailed exploration of sign language recognition techniques. Specifically, we will delve into the subcategories of hand gestures, facial expressions, and combined recognition, as indicated in Fig. 1, in Sections 2.1, 2.2, and 2.3, respectively.

2.1. Hand gesture recognition

Hand gesture recognition systems play a crucial role in various applications, including natural Human-Computer Interaction (HCI) (Barbhuiya, Karsh, & Jain, 2021; Tan, Lim, & Lee, 2021; Wang & Wang, 2007), virtual object manipulation, multimedia and gaming interaction (Wong, Juwono, & Khoo, 2021), smart homes, in-vehicle infotainment systems (Chevtchenko, Vale, & Macario, 2018), and sign language recognition (Birk, Moeslund, & Madsen, 1997; Saxena, Paygude, Jain, Memon, & Naik, 2022).

Hand gestures constitute an essential element of sign language, encompassing hand movement, shape, orientation, alignment, and finger position in relation to the hands and body (Saha, Tapadar, Ray, Chatterjee, & Saha, 2018). While some signs involve sweeping hand movements across a wide area, others rely on the precise positioning of a single finger (Luqman & Mahmoud, 2017). In the following subsections, we will explore published studies on hand gesture recognition, categorized based on the use of specialized hardware and non-specialized hardware (vision-based), respectively.

2.1.1. Hardware-based studies

Various types of hardware have been employed in the literature to facilitate the recognition of hand gestures in sign language. These include depth sensors such as Kinect and Intel RealSense cameras, leap motion sensors, and colored gloves. Fig. 2 demonstrates some special gloves and depth image outputs in first and second rows, respectively.

In a study conducted by Tao et al. (2018), the recognition of static alphabets and numerals in American Sign Language (ASL) was examined. The researchers proposed a novel method that utilized multi-view

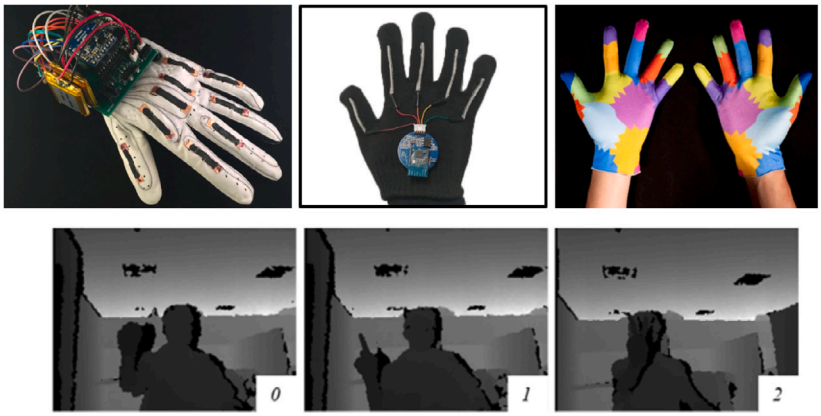

Fig. 2. The top row illustrates some specific types of specialized hardware, specifically gloves, employed in sign language studies. The two images on the left depict wired gloves presented in O'Connor et al. (2017) and Zhou, Chen, et al. (2020), respectively. The colored glove on the right was introduced in a study by Wang and Popović (2009). The second row showcases sample depth images from the NTU digit dataset (Ren et al., 2011).

augmentation and inference fusion using depth images from two publicly available ASL static sign datasets captured by Microsoft Kinect. Their approach involved extracting 3D information from depth images and generating additional data from various perspectives to simulate realistic sign gestures. The system then processed data from each perspective and provided the final prediction for gesture recognition. The experimental results, based on the ASL benchmark dataset (Pugeault & Bowden, 2011), showed an accuracy rate ranging from 93% to 100%. Similarly, on the NTU digit dataset (Ren et al., 2011), the accuracy rate achieved was 100%.

In recent years, advancements in artificial intelligence (AI)-driven approaches have significantly impacted the field of sign language hand gesture recognition. Researchers have leveraged the combination of specialized hardware and Machine Learning (ML) and Deep Learning (DL) frameworks to develop more advanced models. For instance, in the research conducted by Beena, Namboodiri, and Dean (2017), the automatic recognition of static alphabets and numerals in ASL, specifically from 0 to 9, was addressed. Kinect sensors were utilized to capture signals, taking advantage of depth images. The proposed system employed a Convolutional Neural Networks (CNN) classifier, and although it achieved an accuracy rate of 94.68%, the study suggested that further improvements could be made by increasing the number of training data images from various subjects.

By utilizing the Microsoft Kinect sensor, Huang, Zhou, Li, and Li (2015) developed a novel 3D CNN framework to extract spatial–temporal features from raw video streams. The sensor enabled the simultaneous capture of multichannel video streams, including color images, depth maps, and joint locations. A comparison between their proposed method (3D CNN) and the baseline method (Gaussian Mixture Model with Hidden Markov Model (GMM-HMM)) revealed an average accuracy rate of 90.8% for the baseline method and an increased accuracy of 94.2% for the 3D CNN method. This demonstrated the capability of the 3D CNN approach to improve accuracy while leveraging multiple channels.

In the recent study by Ismail, Dawwd, and Ali (2022), the challenges of gesture recognition in videos within the field of computer vision were discussed, particularly focusing on environmental factors. The authors highlighted the limitations of previous single deep network approaches in effectively capturing shape information and spatio-temporal variation simultaneously. To address this, the researchers combined multiple models to capture these features more effectively. They collected a dynamic dataset of 20 meaningful words in Arabic sign language using a Microsoft Kinect v2 camera, which consisted of 7,350 Red, Green, and Blue (RGB) videos and 7,350 depth videos. Four deep neural network models, incorporating 2D and 3D CNNs, were proposed for feature extraction, with sequence classification performed using Long Short Term Memory (LSTM) and Gated Recurrent Units (GRU). Various fusion techniques were evaluated, and the best-performing multi-model achieved 100% accuracy in recognizing Arabic sign language gestures. Among the proposed models, the model that employed pre-trained models, outperformed the others, with ResNet50-LSTM being the most successful multi-model. Fusion methods at the feature level achieved test accuracies above 99% for all approved fusion types. The research introduced the optimal multi-model, ResNet50-BiLSTM-Normalization, which achieved a 100% test accuracy without any incorrect training or validation.

Sahoo, Prakash, Pławiak, and Samantray (2022) addresses the challenges associated with training deep CNN networks, such as AlexNet, VGG-16, and ResNet, from scratch when dealing with the limited availability of labeled image samples for static hand gestures. To overcome this issue, the authors propose an end-to-end fine-tuning approach, which involves utilizing a pre-trained CNN model and applying a score-level fusion technique. This method aims to recognize hand gestures using a dataset with a small number of gesture images. The effectiveness of the proposed technique is evaluated through Leave-One-subject-Out Cross-Validation (LOO CV) and regular CV tests on two benchmark datasets. Furthermore, the authors develop and test a real-time system for ASL recognition using the proposed approach.

In another work by Kolivand, Joudaki, Sunar, and Tully (2021), authors introduce a new technique called ASLNN for hand posture recognition in the ASL alphabet. This technique utilizes a three-dimensional depth-based sensor camera and a neural network to extract geometrical features from hands. The authors compare their proposed DGSLR framework for feature extraction with other methods such as discrete cosine transform and moment invariant. They highlight that the DGSLR framework enhances recognition accuracy by leveraging the invariant nature of the features against hand orientation. The results of their iterations demonstrate that combining the extracted features leads to higher accuracy rates. The authors then employ a neural network to achieve the desired results, showing that ASLNN is proficient in hand posture recognition and achieves a precision rate of up to 96.78%.

Beside Kinect depth cameras, Intel RealSense depth cameras have also become a valuable tool for researchers aiming to facilitate the identification of sign language gestures in images and videos. Some instances of this approach can be observed in the studies of Liao, Li, Ju, and Ouyang (2018) and Mistry and Inden (2018). In the research conducted by Liao et al. (2018), both Intel RealSense camera depth images and RGB images are utilized. To bridge the gap between these two data sources, which lack a direct one-to-one pixel correspondence, a recognition system is developed that leverages the generalized Hough transform for mapping depth images to color images. This alignment aids in segmenting the hand from complex backgrounds in color images using depth information. The recognition process is then facilitated by

a dual-channel CNN, with one channel dedicated to RGB images and the other to depth images.

In addition to depth cameras, leap motion sensors have also been utilized by researchers for sign language recognition. In a study by Hisham and Hamouda (2017), authors employ Leap motion depth sensors to recognize both static and dynamic Arabic sign language gestures. The dataset used in the study includes 28 alphabets, the initial 10 numerals, and Arabic sign language words encompassing nouns and verbs. The authors also discuss a segmentation method for continuous signal sequences. For implementation, well-known ML algorithms such as Support Vector Machine (SVM), K-Nearest-Neighbor (KNN), Artificial Neural Network (ANN), and Dynamic Time Warping (DTW) are employed. The proposed models consider two primary feature sets: palm features set and the bone (hand palm bone) features set. These sets share certain features such as palm position, palm direction, and fingertips direction for each finger in three dimensions of x, y, and z. Through various experiments conducted on the proposed models, it is observed that KNN outperforms other techniques in recognizing palm and bone feature sets, achieving accuracy rates of 99.9% and 98.8%, respectively. On the other hand, DTW outperforms other methods when using the same feature sets, with accuracy rates of 97.4% and 96.4%, respectively.

In a recent article, Lee et al. (2021) addresses the lack of real-time educational applications for ASL and proposes a prototype of a new ASL learning application using the leap motion controller. The application includes both static and dynamic gestures for ASL alphabets. The authors introduce a novel approach combining LSTM recurrent neural network with the KNN method. By training the model with 100 samples for each alphabet (total of 2600 samples) obtained from leap motion controller, they achieve an accuracy rate of 99.44% and 91.82% in 5-fold cross-validation.

In another study, Kumar, Gauba, Roy, and Dogra (2017a) present a multisensor framework for recognizing Indian sign language terms using both Leap motion and Kinect sensors. The combination of these two depth sensors allows the framework to accurately capture sign language gestures from multiple angles. The authors compile a dataset consisting of 25 dynamic isolated Indian sign language terms performed by 10 unique signers, with each sign repeated eight times. By employing the proposed Coupled HMM (CHMM) approach, the model achieves an accuracy of 90.80%.

Similarly, Kumar, Gauba, Roy, and Dogra (2017b) conducts a study using a multi-model framework to capture finger and palm positions for the recognition of 50 words in signed gestures. Leap Motion and Kinect datasets are utilized, and the collected data from both sensors are combined to enhance the accuracy of the system. The results show that the system achieves an accuracy of 97.85% for single hand gestures and 94.55% for double hand gestures. The recognition is performed using HMM and bidirectional LSTM neural network (BLSTM-NN) for single and double hand gestures, respectively.

In a similar vein, another study done by Rastgoo, Kiani, and Escalera (2018) proposes a novel Restricted Boltzmann Machine framework for automatic hand gesture recognition in sign language. The study utilized RGB and depth images as input modalities for their model. Experimental results showcased the effectiveness of the proposed framework, achieving state-of-the-art performance. In a more recent work by the same authors (Rastgoo, Kiani, & Escalera, 2021a), an LSTM model is proposed to address the same modalities in videos. Utilizing four well known dataset of Montalbano II (Escalera et al., 2013), MSR Daily Activity 3D (Wang, Liu, Wu, & Yuan, 2012), and CAD-60 (Sung, Ponce, Selman, & Saxena, 2012), authors discuss the experimental results efficiency increase of 1.64% to 7.6%.

In a study done by Rasines, Remazeilles, and Bengoa (2014), authors present a method for systematically selecting features for sign language recognition, specifically focusing on the first ten ASL numerals (0 to 9). The computational results of sign language recognition were demonstrated using images from the Massey University dataset for hand gestures introduced by Barczak, Reyes, Abastillas, Piccio, and Susnjak (2011). The proposed feature selection method aims to minimize the feature vector while maximizing the F1 score of the classification system, achieving an accuracy rate of 97.7%.

To explore a novel combination of features and optimization technique, Chevtchenko et al. (2018) introduces the Non-dominated Sorting Genetic Algorithm II (NSGA-II), which combines ANNs and genetic algorithms. The Massey University dataset by Barczak et al. (2011) is used to identify 36 hand gestures, including 26 letters and 10 numerals of ASL. Using this approach, the authors achieve an accuracy rate of up to 98.61%. Additionally, they compare the accuracy and computational cost of various commonly used feature extraction methods on the same dataset.

In a similar study utilizing the same dataset, Masood, Chandra, and Srivastava (2018) focuses on applying CNNs for the recognition of sign language hand gestures. They evaluate both static and dynamic ASL alphabets and numerals, totaling 36 gestures, and demonstrate that their proposed CNN model achieves a 96% accuracy rate.

The use of glove-based recognition methods is widespread in sign language recognition due to their advantages, such as independence from locomotion during the recognition process and adaptability to varying lighting conditions. Several studies have explored the application of data gloves for sign language recognition. For both static and dynamic hand gesture recognition, Ong, Lim, Lu, Ng, and Ong (2018) presents SIGMA, a system that utilizes a data glove equipped with a 6-DOF IMU and nine flex sensors. SIGMA focuses on continuous gesture recognition, where static and dynamic gestures are separated by pauses. The method considers four states for static gestures and varying numbers of states (five, six, or seven) for dynamic gestures with different hand forms and movements. Since this limitation is worth noting, the study employs a dataset of 30 samples for each of the 17 featured gestures and achieves an accuracy rate of 79.44% using a HMM for recognition.

In a different study by Kakoty and Sharma (2018), the authors utilize a data glove to recognize single-handed Indian sign language alphabets (8 letters), ASL alphabets (A to Z), and sign numerals (0 to 9) based on hand kinematics. After identifying the alphabets and numerals, they are translated into speech using label matching and a speech database. The proposed technique achieves an accuracy rate of 95.7%.

Alternatively, instead of relying on existing specialized hardware, some researchers have developed and introduced novel hardware solutions. In this regard, Wong et al. (2021) presents a new wearable capacitive sensor device that is cost-effective and designed to capture and record capacitance values. The authors analyze the data collected from these sensors and extract 15 features for training and testing purposes. To enhance the efficiency of the ML algorithms, they propose a feature compression approach based on correlation analysis, and employ two algorithms, namely Error Correction Output Code SVM (ECOC-SVM) and KNN classifiers. The experimental results demonstrate that the proposed approach achieves an accuracy rate of 99% for intra-participant data, while utilizing the mentioned ML methodologies achieves an accuracy rate of 97%.

2.1.2. Vision-based studies

Utilization of specialized hardware, such as depth cameras or special gloves, offers several challenges as well as benefits. Benefits could be listed as simplifying the process of capturing unique characteristics, and increasing the obtained accuracy of the proposed models (Rathi, Pasari, & Sheoran, 2022). While in the opposite side, the challenges that it introduces are sensitiveness towards user motions, limitations of the methods while using special hardware, and the expensive prices for some of the hardware (Park, Lee, & Ko, 2021).

The challenges discussed in the field of hardware-based sign language recognition have motivated researchers to further investigate

vision-based sign language recognition area of research (Shin, Matsuoka, Hasan, & Srizon, 2021). As a result, several vision-based sign language recognition systems have been proposed, utilizing datasets captured by standard cameras and without utilization of any special hardware (Alaghband, Yousefi, & Garibay, 2021; Caselli, Sehyr, Cohen-Goldberg, & Emmorey, 2017; Forster et al., 2012; Joze & Koller, 2018; Kadous, 1995; Koller, Forster, & Ney, 2015; Martínez, Wilbur, Shay, & Kak, 2002). These systems employ AI techniques, such as ML and DL, to develop effective sign language recognition models. Overcoming the challenges in vision-based sign language and gesture recognition has paved the way for the development of real-time interpreter systems that bridge the communication gap between individuals who are deaf and those who do not understand sign language (Nyaga & Wario, 2018).

In a recent study by Masood, Srivastava, Thuwal, and Ahmad (2018), the focus is on real-time sign language recognition using ML and DL techniques, specifically CNN and recurrent neural networks (RNN). The proposed approach utilizes a deep CNN to capture spatial characteristics and an RNN to model temporal characteristics. By employing a dataset consisting of 46 Argentinean sign language (LSA) word gestures, the proposed method achieves an accuracy rate of 95.2%.

Unlike most vision-based studies that focus on either static or dynamic gestures, the study by Kumar and Manjunatha (2017) examines both static and dynamic letters in ASL. They capture webcam image data from input webcams and store it in a database. To handle hybrid configurations and recognize single hand movements, they propose a SVM-HMM algorithm. The performance of the system is evaluated based on metrics such as accuracy, sensitivity, precision, FNR (False Negative Rate), and FDR (False Discovery Rate). The results obtained from the proposed system show a consistent improvement in proficiency and reliability.

In a study by Koller, Ney, and Bowden (2016), a novel approach is proposed that combines a CNN with a HMM to achieve end-to-end embedding. The CNN's outputs are treated as true Bayesian posteriors. The computational results using three publicly accessible benchmarks on continuous sign language datasets (RWTH-PHOENIX-Weather 2012 (Forster et al., 2012), RWTH-PHOENIX-Weather Multisigner 2014 (Forster, Schmidt, Koller, Bellgardt, & Ney, 2014; Koller et al., 2015), and SIGNUM single signer (Von Agris, Knorr, & Kraiss, 2008)) demonstrate a Word Error Rate (WER) that is approximately 15% lower. The best-performing WER is achieved on the first two datasets, with rates of 30% and 7.4%, respectively. For the third dataset (RWTH-PHOENIX-Weather 2014 Multisigner), which contains over 1000 vocabularies, the WER is reduced to 38.8%. Furthermore, the authors provide a dataset of one million hand images, which serves as a valuable benchmark for evaluating state-of-the-art methods.

A year later, Camgoz, Hadfield, Koller, and Bowden (2017) introduced a network called SubUNets as an improvement to the work of Koller, Ney, and Bowden (2016). SubUNets are a collection of domain-specific expert systems that aim to address sequence-to-sequence learning problems and allow for the incorporation of domain-specific expert knowledge. The computational results of SubUNets were evaluated using a dataset of one million hand images, showing an improvement in frame-level accuracy by approximately 30%.

In the same year, Cui, Liu, and Zhang (2017) proposed a weakly supervised deep architecture with a recurrent CNN for continuous sign language recognition. They trained their model on the RWTH-PHOENIX-Weather multisigner 2014 benchmark dataset and achieved comparable results to state-of-the-art models in various contexts without the need for additional supervision.

Another study by Koller, Zargaran, and Ney (2017) presented an iterative re-alignment strategy for labeling visual sequences. They embedded a deep hybrid CNN-BLSTM network within an HMM algorithm. The results on two publicly accessible datasets with over 1000 classes showed an improvement of approximately 10% in the WER compared to the state-of-the-art models.

In a study by Cui, Liu, and Zhang (2019), a recurrent convolutional neural network was proposed for continuous sign language recognition. Unlike previous models, their framework utilized a recurrent neural network for the sequence learning module. The proposed method was evaluated on two publicly accessible datasets: RWTH-PHOENIX-Weather multisigner 2014 and SIGNUM, resulting in improved representation and performance.

Two years later, Sharma and Kumar (2021) focus on dynamic sign language recognition using ASL hand gesture videos from the Boston ASL Lexicon Video Dataset (LVD). They propose a novel 3D-CNN framework for classifying 100 words in the Boston ASL LVD dataset. The dataset is split into 70% for training and 30% for testing. The experimental results show precision of 3.7%, recall of 4.3%, and f-measure of 3.9%.

In another study by Athira, Sruthi, and Lijiya (2022), authors introduce a vision-based gesture recognition system for Indian Sign Language. Their system is capable of recognizing single-handed static and dynamic gestures, double-handed static gestures, and finger spelling words. Key frame extraction is performed using Zernike moments to reduce computation time, and an improved method is introduced to address co-articulation in finger spelling alphabets. The gesture recognition module comprises preprocessing, feature extraction, and classification steps. Experimental results demonstrate accurate recognition of finger spelling alphabets (91% accuracy) and single-handed dynamic words (89% accuracy). The proposed system outperforms existing methods in terms of recognition rate. Moreover, it is cost-effective and can be implemented using a mobile camera, making it user-friendly for everyday use. The inclusion of a trajectory-based method enhances dynamic gesture recognition, and a new dataset is created to address the limited availability of Indian Sign Language data. The experimental results demonstrate high accuracy in recognizing static and dynamic gestures, finger spelling words, and detecting and eliminating co-articulation.

In a recent study focused on ASL, Li, Liu, Shen, and Sun (2022) employ ML and DL methods to recognize and classify 24 English letter signs. They utilize Principal Component Analysis (PCA) and manifold algorithms for dimension reduction. Multiple ML techniques including Random Forest Classification (RFC), KNN, Gaussian Naïve Bayes (GNB), SVM, and Stochastic Gradient Descent (SGD) are employed. The study finds that the manifold algorithm is the most effective for dimension reduction when combined with KNN using the MNIST dataset. Moreover, PCA is shown to be more suitable than KNN for other ML algorithms. CNN and Deep Neural Networks achieve the highest accuracy among the models tested. The research suggests further exploration of evaluation methods and broader applications of these models in ML tasks.

Tyagi and Bansal (2022) address the limitations of existing feature extraction techniques, such as SIFT, SURF, FAST, and ORB, which often suffer from computational inefficiency and classification errors due to insignificant or excessive highlights. To overcome these challenges, they propose a novel approach called Hand mark analysis of sign language (HMASL) that combines feature extraction and hand geometry. HMASL reduces computation and focuses on meaningful regions within complex backgrounds. Experimental results on their proposed Indian sign language dataset, which includes 3,000 images with 300 images for each of the Indian Sign Language words "afraid", "agree", "bad", "become", "chat", "college", "from", "today", "which", and "you", demonstrate that HMASL significantly improves feature selection and enhances recognition accuracy compared to classical feature extraction methods.

Lately, professionals in the domain have adopted Generative Adversarial Network (GAN) models to advance and push the boundaries of sign language research. This includes expanding the scope and quantity of available datasets, as well as refining the existing methodologies in sign language recognition and translation. In a conducted study by Wang, Zeng, Sun, and Liu (2020), authors tackle the challenge of

limited data availability and introduce a novel GAN model to enhance diversity and create data for Chinese sign language.

In response to advancing models and methodologies, Elakkiya et al. (2021) proposes a novel continuous sign language recognition model considering both manual and non-manual signals while utilizing hyperparameter based optimized GANs (H-GANs). Designing the proposed model in three phases of preprocessing data with reduced feature dimensions using Stacked Variational Auto-Encoders (SVAE) and PCA, creating generator and discriminator with deep LSTM and LSTM with a 3D Convolutional Neural Network (3D-CNN), and optimizing and regularizing hyperparameters with Deep Reinforcement Learning (DRL) employments. The proposed H-GANs are evaluated using two benchmark vocabulary sign corpora of continuous sign videos; RWTH-PHOENIX-Weather 2014 and ASLLVD dataset (Athitsos et al., 2008). Experimental findings show that the proposed H-GAN achieves superior performance on benchmark datasets containing multilingual corpora. By incorporating multimodal features (manual, non-manual, and high-level features), the performance and efficiency of the proposed H-GAN are enhanced. This approach outperforms other state-of-the-art techniques through fine-tuned, optimized, and regularized hyperparameters, leading to increased accuracy, reduced error rates, and improved computation time.

2.2. Facial expression recognition

The importance of facial expressions in communicating human emotions and intentions was first addressed by Darwin and Prodger (1998). Due to its importance, facial expression recognition has grasped practical significance in various research fields (i.e., sign language recognition, sociable robotics, driver fatigue surveillance, and HCI systems), leading to numerous studies in this area.

Facial expressions, including eye gaze recognition, eyebrows, eye blinks, and mouth movements, play a critical role in sign language to convey emotions, sentiments, and grammar (Tolba & Elons, 2013). A group of facial expressions called Grammatical Facial Expressions (GFEs) are used in sign language to support grammatical constructions and eliminate sign ambiguity (Kumar et al., 2018). Therefore, sign language recognition systems that do not consider facial expression recognition are insufficient.

Deep neural networks have shown superior performance compared to traditional models in various computer vision tasks, including facial expression capture and recognition. Fan, Lam, and Li (2018) introduce a novel multi-region ensemble CNN to capture global and local features from multiple sub-regions of the face. Jain, Kumar, Kumar, Shamsolmoali, and Zareapoor (2018) propose a CNN-Recurrent Neural Network (CNN-RNN) method for facial expression recognition using two datasets: the MMI facial expression dataset and the Japanese Female Facial Expression (JAFFE) Database. Barsoum, Zhang, Ferrer, and Zhang (2016) demonstrate a Deep CNN method for learning from noisy labels, using facial expression recognition as an example. Jung, Lee, Yim, Park, and Kim (2015) present a joint framework for fine-tuning deep CNNs to address the facial expression recognition problem. The framework consists of two models: one for capturing temporal appearance features from image sequences and another for extracting temporal geometry features from facial landmark points. For additional research on facial expression recognition techniques, we recommend surveys conducted in Li, Wang, Liu, and Feng (2018), Revina and Emmanuel (2021), Zheng, Guo, Huang, Li, and He (2020).

2.3. Combined recognition systems

Sign language recognition systems that aim for a comprehensive understanding must consider both manual and non-manual signals. While most research focuses on analyzing these signals separately, there have been few studies proposing multimodal frameworks that combine hand gestures and facial expressions. These frameworks, known as multimodal methods or multi-semantics, treat different sections of the human body as various signal sources for gesture or sign language recognition (Zhou, Zhou, Zhou, & Li, 2020).

Several recent investigations have explored multimodal approaches in sign language recognition. Kelly and colleagues (Kelly et al., 2009) developed a multimodal system based on HMMs that simultaneously recognizes hand gestures and head movements. Yang and Lee (2011) proposed a multimodal framework using Hierarchical Conditional Random Fields (H-CRF) and SVMs to identify hand gestures and facial expressions, respectively. In a subsequent work, they introduced a second multimodal framework utilizing three cameras to capture different orientations (Yang & Lee, 2013). Chen, Tian, Liu, and Metaxas (2013) presented a novel multimodal framework that combines face and body gestures, employing Histogram of Oriented Gradients (HOG) features and an SVM classifier to recognize ten distinct expressions.

Recent research by Huang, Zhou, Zhang, Li, and Li (2018) introduced a multimodal framework for continuous sign language recognition. Their approach incorporates global hand locations/motions and local hand gesture details, utilizing a Hierarchical Attention Network with Latent Space (LS-HAN). This framework demonstrated improved accuracy on the RWTH-PHOENIX-Weather 2014 benchmark dataset. Zhou, Zhou, et al. (2020) proposed a spatial–temporal multimodal network with joint optimization for end-to-end sequence learning. Their model achieved high precision and performance on three benchmark datasets: RWTH-PHOENIX-Weather 2014, CSL, and RWTH-PHOENIX-Weather T.

These multimodal approaches highlight the importance of integrating different signal sources in sign language recognition systems. By considering multiple modalities and employing advanced models and optimization strategies, researchers have made significant progress in improving the precision and performance of sign language recognition systems.

In a recent study by Min, Hao, Chai, and Chen (2021), the iterative training approach used in recent works on Continuous Sign Language Recognition (CSLR) is examined using the RWTH-PHOENIX-Weather dataset. The study identifies the importance of proper training of the feature extractor to address overfitting. To tackle this issue, the authors propose a novel Visual Alignment Constraint (VAC) that incorporates alignment supervision to enhance the feature extractor. The VAC includes two auxiliary losses, one focusing on visual features and the other enforcing prediction alignment between the feature extractor and alignment module. The study also introduces metrics to evaluate overfitting by measuring prediction inconsistency. Experimental results on the PHOENIX14 and CSL datasets demonstrate that the proposed VAC enables end-to-end training of CSLR networks and achieves competitive performance (Min et al., 2021).

In addition to vision-based models, some research in multimodal sign language recognition utilizes specialized sensors. Kumar et al. (2018) evaluate facial expressions (captured by Kinect) and hand gestures (captured by Leap Motion) simultaneously to enhance recognition performance. They employ the HMM method for recognition assignment and an independent Bayesian classification combination strategy. The authors provide a publicly available dataset for Indian sign language, showcasing high recognition accuracy for single and double hand gestures (Kumar et al., 2018). In another work, Quesada, Marín, and Guerrero (2016) presents a multimodal framework using Intel RealSense depth cameras for capturing face and hand gestures. The work by Agrawal, Jalal, and Tripathi (2016) provides a comprehensive survey on both manual and non-manual sign language recognition systems.

In the most recent work by Aditya et al. (2022), a novel approach for CSLR is proposed, leveraging a spatio-temporal attentive multi-feature network. Existing methods often rely on limited RGB features or focus on individual frames, limiting their learning capability. The proposed method incorporates keypoint features and employs attention mechanisms to emphasize important information. Experimental results

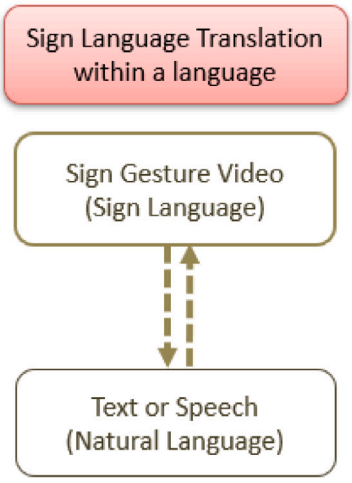

Fig. 3. Schematic representation of frameworks for sign language translation, encompassing translation from sign gesture videos to text or speech and vice versa.

on the CSL and PHOENIX datasets demonstrate the superiority of the proposed approach, with significant improvements in WER scores. The model combines spatial and temporal modules, capturing spatial–temporal correlations and achieving enhanced performance (Aditya et al., 2022). Future work includes exploring transfer learning and applying the proposed model in practical industry settings.

3. Sign language translation

Sign language translation encompasses the transformation of sign language into either or both spoken language and written text, and vice versa. This dynamic field combines technology, linguistics, and AI to overcome communication barriers between sign language users and non-users. Diverse strategies such as neural networks, generative models, and tailored datasets are harnessed to create precise and effective translation mechanisms. Fig. 3 represents a schematic schema of the sign language translation systems.

In a study by Jin, Omar, and Jaward (2016), authors introduced a mobile platform that enables the translation of ASL into text or speech, marking the pioneering application of sign language recognition on a mobile device. The proposed framework focuses on recognizing 16 static ASL alphabets, achieving an average accuracy of 97.13%. To detect hand gestures against complex backgrounds, the authors employed edge detection and region expansion techniques. SURF (Speeded-Up Robust Features) and SVM methods were utilized for feature extraction and classification. It is important to note that the computational performance of the system is sensitive to variations in illumination and background conditions, which can potentially impact the overall accuracy of the recognition process.

Similarly, Park et al. (2021) discusses the challenges of using depth cameras vs. RGB videos in the problem of sign language recognition and translation. To overcome the challenges of using depth cameras, authors propose a novel framework using 3D-CNN to classify sign language gestures and translate them to texts within a language. By collecting their own data from 20 individuals signing 50 Korean sign language words, authors shown that the proposed method could reach the maximum accuracy of 91% and is enable to be utilized as a mobile solution for sign language translation.

A software-based sign language converter was developed by Warrier, Sahu, Halder, Koradiya, and Raj (2016) to recognize ASL cue symbols without relying on ML or DL techniques. The system utilizes a geometric matching algorithm to identify the symbols and generates corresponding discrete text outputs through a decision-making process. The final output can be either converted text or audio, providing accessibility for the end user.

In the case of Indian sign language, Loke, Paranjpe, Bhabal, and Kanere (2017) introduces a system that translates static hand gestures into texts in the Indian spoken language. The system captures images of the signs and employs a transition between color spaces (from RGB to Hue, Saturation, and Value (HSV)) to execute hand tracking and segmentation (feature extraction). These images are then transmitted to a web hosting server and processed by a Matlab neural network. The converted text is sent back to the user's mobile device after mapping the image to its equivalent in the Indian spoken language. To demonstrate the system, an Android application is developed using Android Studio. However, it should be noted that the system's performance may be affected by variations in lighting conditions, as the captured images are obtained in controlled environments with specific background settings, which may differ from real-world scenarios.

A real-time system for sign language recognition and audio output conversion using gesture recognition is presented by Shubhangi and Shinde (2017). The hand gesture recognition process consists of three steps: image preprocessing, feature extraction, and gesture classification. Local Binary Pattern is employed in the initial stage of image preprocessing, while Gray Level Co-occurrence Matrix is used for feature extraction. The final stage involves classification using KNN classifiers. By using a dataset of 10 ASL alphabets, the proposed method achieves an accuracy rate of 92.5% for k=3 and 94% for k=5.

Alternatively, converting speech or text into sign language offers a different approach to address translation challenges. In a work done by Al-Barahamtoshy and Al-Barhamtoshy (2017), an automatic speech recognition (ASR) module is utilized to convert speech into text, which is then combined with a model that translates the textual scripts into Arabic sign language. To achieve this objective, 3D Avatar signers are created and employed. The 3D avatar's hand movements are controlled by four parameters, including hand shape, location, orientation, and movement. Testing on a set of 10 words demonstrates that the model effectively translates the meaning of textual scripts and achieves a detection rate of up to 85%.

In a recent study by Wen, Zhang, He, and Lee (2021), the limitations of existing glove-based solutions for sign language recognition are discussed. These solutions can only recognize discrete single gestures, such as numerals, letters, or words, rather than complete sentences. To address this, the authors propose an AI-based sign language recognition and communication system. The system consists of sensor mittens, a DL block, and a virtual reality interface. The segmentation technique divides complete sentence signals into word units, allowing the DL model to recognize all word elements and reconstruct sentences in reverse order. The proposed model achieves an average accuracy rate of 86.67% in recognizing novel sentences formed by recombining word elements. The results of sign language recognition are projected into virtual space and translated into text and sound, facilitating remote bidirectional communication between signers and non-signers.

In the most recent study by Chakraborty, Sarkar, Paul, Bhattacharjee, and Chakraborty (2023), a novel recognition model is proposed using MediaPipe Holistic followed by two types of recurrent neural networks (Gated Recurrent Unit and Long Short-Term Memory). The authors highlight that the utilization of DL models enables the extraction of temporal and spatial features from sign language videos. The computational results on ASL demonstrate an accuracy of 99%.

In the study by Zhou, Zhou, Qi, Pu, and Li (2021), the limited availability of sign language datasets with parallel sign-text data is addressed. The authors propose a new approach and dataset for sign language translation called sign Back-Translation (signBT) and CSL-Daily, respectively. The method involves using a text-to-gloss translation model to back-translate monolingual text into gloss sequences. These gloss sequences are then combined with segments from a gloss-to-sign bank at the feature level to generate paired sign sequences. This synthetic parallel data is used to enhance the training of the

Fig. 4. Sample continuous frames of benchmark datasets of RWTH PHOENIX Weather 2014 T (Camgoz, Hadfield, Koller, Ney, & Bowden, 2018) and CSL-Daily (Zhou et al., 2021).

encoder–decoder sign language translation (SLT) framework. Additionally, the article introduces CSL-Daily, a large-scale continuous sign language translation dataset that includes spoken language translations and gloss-level annotations related to daily activities. The proposed sign back-translation method is evaluated using extensive experiments on CSL-Daily, demonstrating significant improvements in performance.

In another study by Jiang et al. (2021), a novel approach called the Skeleton Aware Multi-modal Sign Language Recognition (SAM-SLR) framework is proposed. The framework utilizes both RGB and depth data to achieve improved recognition rates. It incorporates the Sign Language Graph Convolution Network (SL-GCN) to capture the embedded dynamics and the Separable Spatial-Temporal Convolution Network (SSTCN) to leverage skeleton features. By including RGB and depth modalities, the framework provides comprehensive information that complements the skeleton-based methods of SL-GCN and SSTCN. Experimental results on the 2021 Looking at People Large Scale Signer Independent Isolated Sign Language Recognition Challenge demonstrate the superior performance of the proposed approach, achieving the highest accuracy in both the RGB (98.42%) and RGB-D (98.53%) tracks.

Establishing a reference dataset and framework, Camgoz et al. (2018) addresses limitations in prior sign language recognition and translation approaches. Through the introduction of a novel methodology named Neural Machine Translation (NMT) systems and a fresh dataset called RWTH PHOENIX Weather 2014 T dataset, the study aims to address the challenges and establish a new foundation for the task of sign language translation. The authors additionally claim that their suggested NMT framework represents the initial exploration into translating sign gesture videos to text. Figs. 4 and 5 demonstrate sample RWTH PHOENIX Weather 2014 T and high level schema of the baseline NMT systems proposed by Camgoz et al. (2018).

As previously discussed, researchers in the field have embraced GANs to advance the existing approaches in the field of sign language recognition, translation, and generation. Through a fusion of GANs, NMT systems, and image generation techniques, Stoll, Camgoz, Hadfield, and Bowden (2018, 2020) introduce innovative methods for automatically crafting sign language videos from spoken language sentences (text). Unlike many prior investigations in this domain, the authors emphasize that their proposed models in both studies do not heavily rely on extensively annotated data. Additionally, the authors assert that the model introduced in the latter work is the first of its kind, generating sign videos without employing traditional graphical avatars. This achievement is realized through two pivotal steps: firstly, translating spoken language sentences into sign pose sequences using NMT in conjunction with a Motion Graph (MG); secondly, generating sign language videos from these sign pose sequences using generative models. Computational results on the benchmark dataset, RWTH PHOENIX Weather 2014 T dataset, reveal a foundational BLEU-4 score of 16.34 on the development set and 15.26 on the test set, laying the groundwork for text-to-gloss translation. For a comprehensive understanding of the BLEU evaluation metrics applied in machine translation tasks, you can refer to the study conducted by Papineni, Roukos, Ward, and Zhu (2002).

Continuing from prior studies that share the idea of converting text into sign gesture videos and employing identical datasets, Camgoz, Koller, Hadfield, and Bowden (2020a, 2020b) introduce novel techniques grounded in the concept of NMT systems. While (Camgoz et al., 2020a) overcomes the dependency on gloss information, Camgoz et al. (2020b) jointly learns continuous sign language recognition and translation. Both works set new baseline for the task of text to sign gesture video tasks. Another research conducted by Vasani, Autee, Kalyani, and Karani (2020) employs sentence processing and GANs to produce sign gesture videos from sentences (text). The entire model is structured to transform sentences into concise gloss notations, subsequently generating synthetic video frames for each gloss through GANs. The outcomes of the GAN model are subsequently refined to create the final sentence video. The authors elaborate that the computational outcomes of their model on an Indian sign language dataset exhibit more authentic video outputs.

One year later, a study conducted by Papastratis, Dimitropoulos, and Daras (2021) introduces an innovative approach to context-aware continuous sign language recognition and translation called Sign Language Recognition GAN (SLRGAN). This innovative network architecture comprises a generator that deciphers sign language glosses by extracting spatial and temporal features from video sequences. Additionally, a discriminator evaluates the generator's predictions by modeling text information at both sentence and gloss levels. To enhance recognition accuracy, contextual information derived from hidden states of previous sentences is incorporated into the bidirectional LSTM module of the generator. In the final phase, sign language translation is executed by a transformer network that converts sign language glosses into natural language text. The proposed method yielded word error rates of 23.4%, 2.1%, and 2.26% on the RWTH-Phoenix-Weather-2014, Chinese Sign Language (CSL), and Greek Sign Language (GSL) Signer Independent (SI) datasets, respectively.

In a recent study by Natarajan, Elakkiya, and Prasad (2023), authors examine the progress and limitations of NMT systems for sign language translation. Acknowledging the challenges of NMT systems in dealing with longer sentences, new vocabulary, and understanding complex multilingual language structures and word relationships, authors introduce an innovative NMT system based on a deep stacked Gated Recurrent Unit (GRU) algorithm. This system is designed to translate spoken sentences into sign language words (glosses). Additionally, authors link the glosses with sign gesture images to automatically generate sign gesture videos through deep GAN models. The effectiveness of their proposed model, named Sentence2SignGesture, is evaluated using three benchmark datasets: RWTH PHOENIX Weather 2014 T dataset, How2Sign (Duarte et al., 2021), and ISL-CSLTR Dataset (Elakkiya & Natarajan, 2021). Computational results indicate that Sentence2SignGesture overcomes the limitations of previous models, excelling in handling multilingual datasets and producing more precise and higher-quality translations with BLEU Score of 38 on RWTH

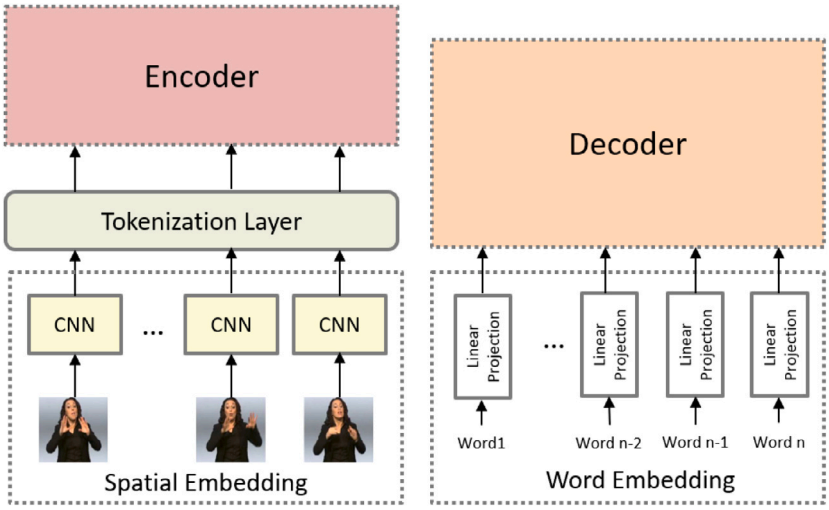

Fig. 5. An overview of baseline NMT model proposed by Camgoz et al. (2018). In this proposed approach, authors suggest the fusion of CNNs with attention-based encoder–decoders to capture the conditional probabilities. By training the proposed method in an integrated manner, authors enable simultaneous learning of alignment and translation from sign language videos to spoken language sentences.

Table 1
An overview of hardware-based sign language datasets.

Authors	Name	Language	Gesture type	Language level	Classes	Data type
Adaloglou et al. (2021)	–	Greek	Static & Dynamic	Words & Sentence	–	Video
Duarte et al. (2021)	How2Sign	American	Static & Dynamic	Words & Sentence	–	Video
Kadous (1995)	UCI Australian Auslan Sign Language dataset[a]	Australian	Dynamic	Alphabets	95	Data Glove
Pugeault and Bowden (2011)	ASL Finger Spelling A[b]	American	Static	Alphabets	24	Depth Images
Pugeault and Bowden (2011)	ASL Finger Spelling B[c]	American	Static	Alphabets	24	Depth Image
Kurakin, Zhang, and Liu (2012)	MSRGesture 3D[d]	American	–	Words	12	Depth Video
Escalera et al. (2014)	CLAP14	Italian	–	Words	20	Depth Video
Wan et al. (2016)	ChaLearn LAP IsoGD[e] & ConGD[f]	–	Static & Dynamic	–	249	RGB & Depth Video
Kapuscinski, Oszust, Wysocki, and Warchol (2015)	PSL Kinect 30[g]	Polish	Dynamic	Words	30	Kinect Video
Ansari and Harit (2016)	ISL[h]	Indian	static	Alphabets, Numbers, Words	140	Depth Images

[a] Stands for Australian Sign Language. Available at: https://archive.ics.uci.edu/ml/datasets/Australian+Sign+Language+signs+(High+Quality).
[b] Available at: http://empslocal.ex.ac.uk/people/staff/np331/index.php?section=FingerSpellingDataset.
[c] Available at: http://empslocal.ex.ac.uk/people/staff/np331/index.php?section=FingerSpellingDataset.
[d] Available at: https://www.uow.edu.au/~wanqing/Datasets.
[e] Available at: http://www.cbsr.ia.ac.cn/users/jwan/database/isogd.html.
[f] Available at: http://www.cbsr.ia.ac.cn/users/jwan/database/congd.html.
[g] Available at: http://vision.kia.prz.edu.pl/dynamickinect.php.
[h] Available at: https://github.com/zafar142007/Gesture-Recognition-for-Indian-Sign-Language-using-Kinect.

PHOENIX Weather 2014 T, 38.6 in How2Sign, and 39.1 in ISL-CSLTR dataset.

Despite the existence of well-known benchmark datasets like PHOENIX-2014T and CSL-Daily for sign language translation, these datasets remain significantly smaller than the parallel data typically used to train sign language translation models. Acknowledging the scarcity of data, multiple studies have been conducted to address this issue (Chen, Wei, Sun, Wu, & Lin, 2022; Orbay & Akarun, 2020). Employing transfer learning to explore semi-supervised tokenization methods without the need for extra labeling, Chen et al. (2022) introduces a novel approach to establish a straightforward yet efficient multi-modality baseline. The approach suggested entails a stepwise pretraining process employing both general-domain and within-domain datasets. To elaborate, the visual network responsible for sign-to-gloss conversion undergoes pretraining with extensive human action data and dedicated sign-to-gloss datasets. Similarly, the gloss-to-text translation network is pretrained with multilingual corpora and domain-specific gloss-to-text datasets. The introduction of a visual-language mapper component additionally bolsters the model's effectiveness. This direct method outperforms previous state-of-the-art achievements on prominent sign language translation benchmarks, emphasizing the effectiveness of employing transfer learning to tackle the challenge of limited data availability.

Regarding sign gesture video to text translation, a multitude of studies are available, extensively covered in a recent survey by Núñez-Marcos et al. (2022) on sign language machine translation. Based on thorough investigation in this domain, authors highlight the absence of a definitive optimal solution and emphasize the necessity for further research efforts.

For more recent comprehensive insights into sign language translation literature, we suggest referring to the following manuscripts (Ananthanarayana et al., 2021; Ardiansyah et al., 2021; Farooq, Rahim, Sabir, Hussain, & Abid, 2021; Núñez-Marcos et al., 2022).

4. Sign language datasets

Insufficient availability of comprehensive annotated datasets for sign language has hindered progress in the fields of sign language recognition and translation. Hence, this section aims to review existing sign language datasets in order to enhance their quantity and quality.

Table 2
An overview of vision-based sign language datasets.

Authors	Name	Language	Gesture type	Language level	Classes	Data type
Martínez et al. (2002), Wilbur and Kak (2006)	Purdue RVL-SLLL ASL Database[a]	American	–	Alphabets, Numbers, Words, Paragraphs	104	Image, Video
Shanableh, Assaleh, and Al-Rousan (2007)	–	Arabic	–	Words	23	Video
Athitsos et al. (2008)	Boston ASLLVD[b]	American	Dynamic	Words	>3300	Video
Von Agris et al. (2008)	SIGNUM[c]	German	–	Words, Sentences	450 Words, 780 Sentence	Video
Barczak et al. (2011)	–	American	Static	Alphabets, Numbers	36	Image
Pugeault and Bowden (2011)	ASL Finger Spelling A[d]	American	Static	Alphabets	24	Image
Forster et al. (2012)	RWTH-PHOENIX-Weather 2012[e]	German	–	Sentence	1200	Image
Forster et al. (2014), Koller et al. (2015)	RWTH-PHOENIX-Weather Multisigner 2014[f]	German	Dynamic	Sentence	>1000	Video
Ronchetti et al. (2016)	LSA16[g]	Argentinian	–	Alphabets, Words	16	Image
Ronchetti et al. (2016)	LSA64[h]	Argentinian	–	Words	64	Video
Ansari and Harit (2016)	the ISL dataset[i]	Indian	static	Alphabets, Numbers, Words	140	Image
Oliveira et al. (2017)	ISL hand shape dataset[j]	Irish	Static & Dynamic	–	23 Static & 3 Dynamic	Image Video
Hosoe, Sako, and Kwolek (2017)	Japaneese Finger spelling sign language dataset	Japan	–	–	41	Image
Feng et al. (2017)	HUST-ASL[k]	American	Static	Alphabets, Numbers	34	RGB & Kinect Image
Caselli et al. (2017)	ASL-LEX[l]	American	–	Words	Nearly 1000	Video
Joze and Koller (2018)	MS-ASL[m]	American	Dynamic	–	1000	Video
Camgoz et al. (2018)	RWTH-PHOENIX-Weather 2014T[n]	German	Continuous	Sentence	>7000	Video
Ko, Kim, Jung, and Cho (2019)	KETI sign language dataset[o]	Korean	Continuous	Sentence	>14,000	Video
Elakkiya and Natarajan (2021)	ISL–CSLTR[p]	Indian	Continuous	Sentence	100	Video
Duarte et al. (2021)	How2Sign	American	Static & Dynamic	Words & Sentence	–	Video
Alaghband et al. (2021)	FePh[q]	German	Static & Dynamic	Words & Sentence	–	Video
Zhou et al. (2021)	CSL-Daily[r]	Chinese	Continuous	Sentence	>20,000	Video

[a] Available at: http://www2.ece.ohio-state.edu/~aleix/ASLdatabase.htm and https://engineering.purdue.edu/RVL/Database/ASL/asl-database-front.htm.
[b] Stands for American Sign Language Lexicon Video Dataset. Available at: http://www.bu.edu/av/asllrp/dai-asllvd.html.
[c] Available at: http://www.phonetik.uni-muenchen.de/Bas/SIGNUM/.
[d] Available at: http://empslocal.ex.ac.uk/people/staff/np331/index.php?section=FingerSpellingDataset.
[e] Available at: https://www-i6.informatik.rwth-aachen.de/~forster/database-rwth-phoenix.php.
[f] Available at: https://www-i6.informatik.rwth-aachen.de/~koller/RWTH-PHOENIX/.
[g] Available at: http://facundoq.github.io/unlp/lsa16/index.html.
[h] Available at: http://facundoq.github.io/unlp/lsa64/index.html.
[i] Available at: https://github.com/zafar142007/Gesture-Recognition-for-Indian-Sign-Language-using-Kinect.
[j] Available at: https://github.com/marlondcu/ISL.
[k] Stands for Huazhong University of Science & Technology.
[l] Available at: http://asl-lex.org/.
[m] Available at: https://www.microsoft.com/en-us/download/details.aspx?id=100121.
[n] Suitable for sign language translation. Available at: https://www-i6.informatik.rwth-aachen.de/~koller/RWTH-PHOENIX-2014-T/.
[o] Suitable for sign language translation.
[p] Available at: https://data.mendeley.com/datasets/kcmpdxky7p/1.
[q] Available at: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/358QMQ.
[r] Suitable for sign language translation. Available at: http://home.ustc.edu.cn/~zhouh156/dataset/csl-daily/.

As previously discussed in Section 2.1, conducted research on sign language recognition systems can be categorized into two main types: hardware-based and vision-based recognition systems. Hardware-based systems utilize datasets obtained using specialized colored gloves (Kadous, 2002; Ronchetti et al., 2016; Wang & Popović, 2009), specialized sensors, or depth cameras (such as Microsoft Kinect and Leap Motion) to capture distinct features of the signer's gestures. Table 1 provides a compilation of well-known hardware-based sign language datasets.

While the use of hardware simplifies the capture of specific features, its limited availability restricts its applicability. Therefore, vision-based sign language recognition systems are proposed, which utilize datasets collected using standard cameras (Forster et al., 2012; Joze & Koller, 2018; Koller et al., 2015; Martínez et al., 2002; Mehdi & Khan, 2002). Table 2 presents a selection of notable vision-based sign language datasets.

5. Conclusion

Significant progress has been made in the field of hardware-based sign language recognition, however, it is important to acknowledge that this methodology has its limitations and challenges. With this in mind, researchers should explore vision-based techniques, which have not received as much emphasis in the literature. Expanding research efforts in vision-based approaches can help address the limitations

of hardware-based methods and contribute to a more comprehensive understanding of sign language recognition.

While there have been advancements in sign language recognition techniques that consider individual modalities (e.g., hand gestures or facial expressions), the exploration of combined sign language techniques incorporating multiple modalities is still relatively limited. Considering that sign language involves various elements, including hand movements, facial expressions, and body postures, it is essential to conduct further research that integrates and analyzes these combined modalities. By focusing on combined modality features, researchers can gain a deeper understanding of sign language communication and develop more accurate and comprehensive recognition techniques.

The field of sign language translation is a significant area of study within sign language literature. However, one of the challenges in this field is the lack of benchmark datasets and comprehensive research specifically focused on sign language translation systems that translates one sign language to another. For example, translating ASL to Persian sign language is a form of translation between languages. It is worth noting that many deaf individuals have limited or no knowledge of spoken languages and may struggle with written or spoken language comprehension (Luqman & Mahmoud, 2017). Therefore, translating sign language into a spoken language may not be sufficient to meet the needs of the deaf community. As a result, there is a need for development of standardized datasets that encompass various sign languages and linguistic contexts. With benchmark datasets and rigorous research, advancements in sign language translation can be achieved, leading to improved accessibility and communication for individuals who use sign language.

Finally, despite the progress made in sign language recognition and translation research, there remains a noticeable scarcity of sign language recognition/translation software and applications in practical settings. The development of user-friendly and accessible software and applications for sign language recognition and translation is crucial to bridging the gap between research and real-world implementation. Creating intuitive and reliable software tools will benefit deaf or hard of hearing individuals and has the potential to improve their everyday interaction and overall communication.

CRediT authorship contribution statement

Marie Alaghband: Conceptualization, Data curation, Writing – original draft, Visualization, Investigation, Editing. **Hamid Reza Maghroor:** Visualization, Writing – review & editing. **Ivan Garibay:** Supervision, Validation, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: No relationships to declare.

Data availability

No data was used for the research described in the article.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT in order to edit the writing of the paper. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

Adaloglou, N., Chatzis, T., Papastratis, I., Stergioulas, A., Papadopoulos, G. T., Zacharopoulou, V., et al. (2021). A comprehensive study on deep learning-based methods for sign language recognition. *IEEE Transactions on Multimedia*, *24*, 1750–1762.

Aditya, W., Shih, T. K., Thaipisutikul, T., Fitriajie, A. S., Gochoo, M., Utaminingrum, F., et al. (2022). Novel spatio-temporal continuous sign language recognition using an attentive multi-feature network. *Sensors*, *22*(17), 6452.

Agrawal, S. C., Jalal, A. S., & Tripathi, R. K. (2016). A survey on manual and non-manual sign language recognition for isolated and continuous sign. *International Journal of Applied Pattern Recognition*, *3*(2), 99–134.

Ai-Media (2023). Sign language alphabets. https://www.ai-media.tv/sign-language-alphabets/.

Al-Barahamtoshy, O. H., & Al-Barhamtoshy, H. M. (2017). Arabic text-to-sign (ArTTS) model from automatic SR system. *Procedia Computer Science*, *117*, 304–311.

Alaghband, M. (2021). *Analysis of Sign language facial expressions and deaf students' retention using machine learning and agent-based modeling*. (Doctoral dissertation), University of Central Florida, https://stars.library.ucf.edu/etd2020/1317/.

Alaghband, M., Yousefi, N., & Garibay, I. (2021). Facial expression phoenix (FePh): An annotated sequenced dataset for facial and emotion-specified expressions in sign language. *International Journal of Electronics and Communication Engineering*, *15*(3), 131–138.

Ananthanarayana, T., Srivastava, P., Chintha, A., Santha, A., Landy, B., Panaro, J., et al. (2021). Deep learning methods for sign language translation. *ACM Transactions on Accessible Computing (TACCESS)*, *14*(4), 1–30.

Ansari, Z. A., & Harit, G. (2016). Nearest neighbour classification of Indian sign language gestures using kinect camera. *Sadhana*, *41*(2), 161–182.

Ardiansyah, A., Hitoyoshi, B., Halim, M., Hanafiah, N., & Wibisurya, A. (2021). Systematic literature review: American sign language translator. *Procedia Computer Science*, *179*, 541–549.

Athira, P., Sruthi, C., & Lijiya, A. (2022). A signer independent sign language recognition with co-articulation elimination from live videos: an Indian scenario. *Journal of King Saud University-Computer and Information Sciences*, *34*(3), 771–781.

Athitsos, V., Neidle, C., Sclaroff, S., Nash, J., Stefan, A., Yuan, Q., et al. (2008). The American sign language lexicon video dataset. In *Computer vision and pattern recognition workshops, 2008. CVPRW'08. IEEE computer society conference on* (pp. 1–8). IEEE.

Bahia, N. K., & Rani, R. (2023). Multi-level taxonomy review for sign language recognition: Emphasis on indian sign language. *ACM Transactions on Asian and Low-Resource Language Information Processing*, *22*(1), 1–39.

Barbhuiya, A. A., Karsh, R. K., & Jain, R. (2021). CNN based feature extraction and classification for sign language. *Multimedia Tools and Applications*, *80*(2), 3051–3069.

Barczak, A. L. C., Reyes, N. H., Abastillas, M., Piccio, A., & Susnjak, T. (2011). A new 2D static hand gesture colour image dataset for ASL gestures. *Research Letters in Information Mathematical Sciences*, *15*, 12–20.

Barsoum, E., Zhang, C., Ferrer, C. C., & Zhang, Z. (2016). Training deep networks for facial expression recognition with crowd-sourced label distribution. In *Proceedings of the 18th ACM international conference on multimodal interaction* (pp. 279–283).

Beena, M., Namboodiri, M. A., & Dean, P. (2017). Automatic sign language finger spelling using convolution neural network: Analysis. *International Journal of Pure and Applied Mathematics*, *117*(20), 9–15.

Birk, H., Moeslund, T. B., & Madsen, C. B. (1997). Real-time recognition of hand alphabet gestures using principal component analysis. In *Proceedings of the scandinavian conference on image analysis, vol. 1* (pp. 261–268). PROCEEDINGS PUBLISHED BY VARIOUS PUBLISHERS.

Bulugu, I. (2021). Sign language recognition using Kinect sensor based on color stream and skeleton points. *Tanzania Journal of Science*, *47*(2), 769–778.

Camgoz, N. C., Hadfield, S., Koller, O., & Bowden, R. (2017). Subunets: End-to-end hand shape and continuous sign language recognition. In *2017 IEEE international conference on computer vision* (pp. 3075–3084). IEEE.

Camgoz, N. C., Hadfield, S., Koller, O., Ney, H., & Bowden, R. (2018). Neural sign language translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7784–7793).

Camgoz, N. C., Koller, O., Hadfield, S., & Bowden, R. (2020a). Multi-channel transformers for multi-articulatory sign language translation. In *Computer vision–ECCV 2020 workshops: Glasgow, UK, August 23–28, 2020, proceedings, part IV 16* (pp. 301–319). Springer.

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

Caselli, N. K., Sehyr, Z. S., Cohen-Goldberg, A. M., & Emmorey, K. (2017). ASL-LEX: A lexical database of American sign language. *Behavior Research Methods*, *49*(2), 784–801.

Chakraborty, S., Sarkar, S., Paul, P., Bhattacharjee, S., & Chakraborty, A. (2023). Sign language recognition using landmark detection, GRU and LSTM. *American Journal of Electronics and Communication*.

Chen, S., Tian, Y., Liu, Q., & Metaxas, D. N. (2013). Recognizing expressions from face and body gesture by temporal normalized motion and appearance features. *Image and Vision Computing*, *31*(2), 175–185.

Chen, Y., Wei, F., Sun, X., Wu, Z., & Lin, S. (2022). A simple multi-modality transfer learning baseline for sign language translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5120–5130).

Chevtchenko, S. F., Vale, R. F., & Macario, V. (2018). Multi-objective optimization for hand posture recognition. *Expert Systems with Applications*, *92*, 170–181.

Cui, R., Liu, H., & Zhang, C. (2017). Recurrent convolutional neural networks for continuous sign language recognition by staged optimization. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7361–7369).

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.

Darwin, C., & Prodger, P. (1998). *The expression of the emotions in man and animals*. USA: Oxford University Press.

Duarte, A., Palaskar, S., Ventura, L., Ghadiyaram, D., DeHaan, K., Metze, F., et al. (2021). How2sign: a large-scale multimodal dataset for continuous American sign language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2735–2744).

Elakkiya, R. (2021). RETRACTED ARTICLE: Machine learning based sign language recognition: a review and its research frontier. *Journal of Ambient Intelligence and Humanized Computing*, *12*(7), 7205–7224.

Elakkiya, R., & Natarajan, B. (2021). ISL-CSLTR: Indian Sign Language Dataset for Continuous Sign Language Translation and Recognition. http://dx.doi.org/10.17632/kcmpdxky7p.1.

Elakkiya, R., Vijayakumar, P., & Kumar, N. (2021). An optimized generative adversarial network based continuous sign language classification. *Expert Systems with Applications*, *182*, Article 115276.

Escalera, S., Baró, X., Gonzalez, J., Bautista, M. A., Madadi, M., Reyes, M., et al. (2014). Chalearn looking at people challenge 2014: Dataset and results. In *Workshop at the European conference on computer vision* (pp. 459–473). Springer.

Escalera, S., Gonzàlez, J., Baró, X., Reyes, M., Lopes, O., Guyon, I., et al. (2013). Multi-modal gesture recognition challenge 2013: Dataset and results. In *Proceedings of the 15th ACM on international conference on multimodal interaction* (pp. 445–452).

Fan, Y., Lam, J. C., & Li, V. O. (2018). Multi-region ensemble convolutional neural network for facial expression recognition. In *International conference on artificial neural networks* (pp. 84–94). Springer.

Farooq, U., Rahim, M. S. M., Sabir, N., Hussain, A., & Abid, A. (2021). Advances in machine translation for sign language: approaches, limitations, and challenges. *Neural Computing and Applications*, *33*(21), 14357–14399.

Feng, B., He, F., Wang, X., Wu, Y., Wang, H., Yi, S., et al. (2017). Depth-projection-map-based bag of contour fragments for robust hand gesture recognition. *IEEE Transactions on Human-Machine Systems*, *47*(4), 511–523.

Forster, J., Schmidt, C., Hoyoux, T., Koller, O., Zelle, U., Piater, J. H., et al. (2012). RWTH-PHOENIX-weather: A large vocabulary sign language recognition and translation corpus. In *LREC* (pp. 3785–3789).

Forster, J., Schmidt, C., Koller, O., Bellgardt, M., & Ney, H. (2014). Extensions of the sign language recognition and translation corpus RWTH-PHOENIX-weather. In *LREC* (pp. 1911–1916).

Freitas, F. A., Peres, S. M., Lima, C. A., & Barbosa, F. V. (2017). Grammatical facial expression recognition in sign language discourse: a study at the syntax level. *Information Systems Frontiers*, *19*(6), 1243–1259.

Guo, L., Lu, Z., & Yao, L. (2021). Human-machine interaction sensing technology based on hand gesture recognition: A review. *IEEE Transactions on Human-Machine Systems*, *51*(4), 300–309.

Hisham, B., & Hamouda, A. (2017). Arabic static and dynamic gestures recognition using leap motion. *Journal of Scientific Computing*, *13*(8), 337–354.

Hosoe, H., Sako, S., & Kwolek, B. (2017). Recognition of JSL finger spelling using convolutional neural networks. In *Machine vision applications (MVA), 2017 fifteenth IAPR international conference on* (pp. 85–88). IEEE.

Huang, J., Zhou, W., Li, H., & Li, W. (2015). Sign language recognition using 3d convolutional neural networks. In *Multimedia and expo (ICME), 2015 IEEE international conference on* (pp. 1–6). IEEE.

Huang, J., Zhou, W., Zhang, Q., Li, H., & Li, W. (2018). Video-based sign language recognition without temporal segmentation. arXiv preprint arXiv:1801.10111.

Ismail, M. H., Dawwd, S. A., & Ali, F. H. (2022). Dynamic hand gesture recognition of Arabic sign language by using deep convolutional neural networks. *Indonesian Journal of Electrical Engineering and Computer Science*, *25*, 952–962.

Jain, N., Kumar, S., Kumar, A., Shamsolmoali, P., & Zareapoor, M. (2018). Hybrid deep neural networks for face emotion recognition. *Pattern Recognition Letters*, *115*, 101–106.

Jiang, S., Sun, B., Wang, L., Bai, Y., Li, K., & Fu, Y. (2021). Skeleton aware multi-modal sign language recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3413–3423).

Jin, C. M., Omar, Z., & Jaward, M. H. (2016). A mobile application of American sign language translation via image processing algorithms. In *Region 10 symposium* (pp. 104–109). IEEE.

Joze, H. R. V., & Koller, O. (2018). MS-ASL: A large-scale data set and benchmark for understanding American sign language. arXiv preprint arXiv:1812.01053.

Jung, H., Lee, S., Yim, J., Park, S., & Kim, J. (2015). Joint fine-tuning in deep neural networks for facial expression recognition. In *Proceedings of the IEEE international conference on computer vision* (pp. 2983–2991).

Kadous, M. W. (1995). *GRASP: Recognition of Australian sign language using Instrumented gloves*. Citeseer.

Kadous, M. W. (2002). *Temporal classification: Extending the classification paradigm to multivariate time series*. University of New South Wales Kensington.

Kakoty, N. M., & Sharma, M. D. (2018). Recognition of sign language alphabets and numbers based on hand kinematics using a data glove. *Procedia Computer Science*, *133*, 55–62.

Kapuscinski, T., Oszust, M., Wysocki, M., & Warchol, D. (2015). Recognition of hand gestures observed by depth cameras. *International Journal of Advanced Robotic Systems*, *12*(4), 36.

Kelly, D., Reilly Delannoy, J., Mc Donald, J., & Markham, C. (2009). A framework for continuous multimodal sign language recognition. In *Proceedings of the 2009 international conference on multimodal interfaces* (pp. 351–358). ACM.

Ko, S.-K., Kim, C. J., Jung, H., & Cho, C. (2019). Neural sign language translation based on human keypoint estimation. *Applied Sciences*, *9*(13), 2683.

Kolivand, H., Joudaki, S., Sunar, M. S., & Tully, D. (2021). A new framework for sign language alphabet hand posture recognition using geometrical features through artificial neural network (part 1). *Neural Computing and Applications*, *33*, 4945–4963.

Koller, O., Forster, J., & Ney, H. (2015). Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, *141*, 108–125.

Koller, O., Ney, H., & Bowden, R. (2016). Deep hand: How to train a cnn on 1 million hand images when your data is continuous and weakly labelled. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3793–3802).

Koller, O., Zargaran, S., & Ney, H. (2017). Re-sign: Re-aligned end-to-end sequence modelling with deep recurrent CNN-HMMs. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4297–4305).

Koller, O., Zargaran, O., Ney, H., & Bowden, R. (2016). Deep sign: hybrid CNN-HMM for continuous sign language recognition. In *Proceedings of the British machine vision conference 2016*.

Koller, O., Zargaran, S., Ney, H., & Bowden, R. (2018). Deep sign: Enabling robust statistical continuous sign language recognition via hybrid CNN-HMMs. *International Journal of Computer Vision*, *126*(12), 1311–1325.

Kumar, P., Gauba, H., Roy, P. P., & Dogra, D. P. (2017a). Coupled HMM-based multi-sensor data fusion for sign language recognition. *Pattern Recognition Letters*, *86*, 1–8.

Kumar, P., Gauba, H., Roy, P. P., & Dogra, D. P. (2017b). A multimodal framework for sensor based sign language recognition. *Neurocomputing*, *259*, 21–38.

Kumar, B. P., & Manjunatha, M. (2017). A hybrid gesture recognition method for American sign language. *Indian Journal of Science and Technology*, *10*(1).

Kumar, P., Roy, P. P., & Dogra, D. P. (2018). Independent Bayesian classifier combination based sign language recognition using facial expression. *Information Sciences*, *428*, 30–48.

Kurakin, A., Zhang, Z., & Liu, Z. (2012). A real time system for dynamic hand gesture recognition with a depth sensor. In *EUSIPCO, vol. 2, no. 5* (p. 6).

Lee, C. K., Ng, K. K., Chen, C.-H., Lau, H. C., Chung, S., & Tsoi, T. (2021). American sign language recognition and training method with recurrent neural network. *Expert Systems with Applications*, *167*, Article 114403.

Li, L., Liu, D., Shen, C., & Sun, J. (2022). American sign language recognition based on machine learning and neural network. In *2022 international conference on machine learning and intelligent systems engineering* (pp. 452–457). IEEE.

Li, Y., Wang, X., Liu, W., & Feng, B. (2018). Deep attention network for joint hand gesture localization and recognition using static RGB-D images. *Information Sciences*, *441*, 66–78.

Liao, B., Li, J., Ju, Z., & Ouyang, G. (2018). Hand gesture recognition with generalized hough transform and DC-CNN using realsense. In *2018 eighth international conference on information science and technology* (pp. 84–90).

Loke, P., Paranjpe, J., Bhabal, S., & Kanere, K. (2017). Indian sign language converter system using an android app. In *Electronics, communication and aerospace technology (ICECA), 2017 international conference of, vol. 2* (pp. 436–439). IEEE.

Luqman, H., & Mahmoud, S. A. (2017). Transform-based Arabic sign language recognition. *Procedia Computer Science*, *117*, 2–9.

Martínez, A. M., Wilbur, R. B., Shay, R., & Kak, A. C. (2002). Purdue RVL-SLLL ASL database for automatic recognition of American sign language. In *Multimodal interfaces, 2002. proceedings. fourth IEEE international conference on* (pp. 167–172). IEEE.

Masood, S., Chandra, H. T., & Srivastava, A. (2018). American sign language character recognition using convolution neural network. In *Smart computing and informatics: Proceedings of the first international conference on SCI 2016, volume 2* (pp. 403–412).

Masood, S., Srivastava, A., Thuwal, H. C., & Ahmad, M. (2018). Real-time sign language gesture (word) recognition from video sequences using CNN and RNN. In *Intelligent engineering informatics* (pp. 623–632). Springer.

Mehdi, S. A., & Khan, Y. N. (2002). Sign language recognition using sensor gloves. In *Proceedings of the 9th international conference on neural information processing, vol. 5* (pp. 2204–2206). IEEE.

Min, Y., Hao, A., Chai, X., & Chen, X. (2021). Visual alignment constraint for continuous sign language recognition. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 11542–11551).

Mistry, J., & Inden, B. (2018). An approach to sign language translation using the intel realsense camera. In *2018 10th computer science and electronic engineering* (pp. 219–224). IEEE.

Natarajan, B., Elakkiya, R., & Prasad, M. L. (2023). Sentence2SignGesture: a hybrid neural machine translation network for sign language video generation. *Journal of Ambient Intelligence and Humanized Computing*, *14*(8), 9807–9821.

National Institute on Deafness and Other Communication Disorders (NIDCD) (2023). American sign language. https://www.nidcd.nih.gov/health/american-sign-language.

Neiva, D. H., & Zanchettin, C. (2018). Gesture recognition: A review focusing on sign language in a mobile context. *Expert Systems with Applications*, *103*, 159–183.

Núñez-Marcos, A., Perez-de Viñaspre, O., & Labaka, G. (2022). A survey on sign language machine translation. *Expert Systems with Applications*, Article 118993.

Nyaga, C. N., & Wario, R. D. (2018). Sign language gesture recognition through computer vision. In *2018 IST-Africa week conference* (pp. 1–8). IEEE.

O'Connor, T. F., Fach, M. E., Miller, R., Root, S. E., Mercier, P. P., & Lipomi, D. J. (2017). The language of glove: Wireless gesture decoder with low-power and stretchable hybrid electronics. *PLoS One*, *12*(7), Article e0179766.

Oliveira, M., Chatbri, H., Ferstl, Y., Farouk, M., Little, S., & OConnor, N. (2017). A dataset for Irish sign language recognition. In *Proceedings of the Irish machine vision and image processing conference, vol. 8*.

Ong, C., Lim, I., Lu, J., Ng, C., & Ong, T. (2018). Sign-language recognition through gesture & movement analysis (SIGMA). In *Mechatronics and machine vision in practice 3* (pp. 235–245). Springer.

Orbay, A., & Akarun, L. (2020). Neural sign language translation by learning tokenization. In *2020 15th IEEE international conference on automatic face and gesture recognition* (pp. 222–228). IEEE.

Oudah, M., Al-Naji, A., & Chahl, J. (2020). Hand gesture recognition based on computer vision: A review of techniques. *journal of Imaging*, *6*(8), 73.

Papastratis, I., Dimitropoulos, K., & Daras, P. (2021). Continuous sign language recognition through a context-aware generative adversarial network. *Sensors*, *21*(7), 2437.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311–318).

Park, H., Lee, Y., & Ko, J. (2021). Enabling real-time sign language translation on mobile platforms with on-board depth cameras. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *5*(2), 1–30.

Pugeault, N., & Bowden, R. (2011). Spelling it out: Real-time ASL fingerspelling recognition. In *Computer vision workshops (ICCV Workshops), 2011 IEEE international conference on* (pp. 1114–1119). IEEE.

Quesada, L., Marín, G., & Guerrero, L. A. (2016). Sign language recognition model combining non-manual markers and handshapes. In *International conference on ubiquitous computing and ambient intelligence* (pp. 400–405). Springer.

Rasines, I., Remazeilles, A., & Bengoa, P. M. I. (2014). Feature selection for hand pose recognition in human-robot object exchange scenario. In *Emerging technology and factory automation* (pp. 1–8). IEEE.

Rastgoo, R., Kiani, K., & Escalera, S. (2018). Multi-modal deep hand sign language recognition in still images using restricted Boltzmann machine. *Entropy*, *20*(11), 809.

Rastgoo, R., Kiani, K., & Escalera, S. (2021a). Hand pose aware multimodal isolated sign language recognition. *Multimedia Tools and Applications*, *80*, 127–163.

Rastgoo, R., Kiani, K., & Escalera, S. (2021b). Sign language recognition: A deep survey. *Expert Systems with Applications*, *164*, Article 113794.

Rathi, A., Pasari, S., & Sheoran, S. (2022). Live sign language recognition: Using convolution neural networks. In *2022 8th international conference on advanced computing and communication systems, vol. 1* (pp. 502–505). IEEE.

Ren, Z., Yuan, J., & Zhang, Z. (2011). Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera. In *Proceedings of the 19th ACM international conference on multimedia* (pp. 1093–1096). ACM.

Revina, I. M., & Emmanuel, W. S. (2021). A survey on human face expression recognition techniques. *Journal of King Saud University-Computer and Information Sciences*, *33*(6), 619–628.

Ronchetti, F., Quiroga, F., Estrebou, C. A., Lanzarini, L. C., & Rosete, A. (2016). LSA64: An argentinian sign language dataset. In *Proceedings of the XXII Congreso Argentino de Ciencias de la Computación*.

Saha, H. N., Tapadar, S., Ray, S., Chatterjee, S. K., & Saha, S. (2018). A machine learning based approach for hand gesture recognition using distinctive feature extraction. In *Computing and communication workshop and conference (CCWC), 2018 IEEE 8th annual* (pp. 91–98). IEEE.

Sahoo, J. P., Prakash, A. J., Pławiak, P., & Samantray, S. (2022). Real-time hand gesture recognition using fine-tuned convolutional neural network. *Sensors*, *22*(3), 706.

Saxena, S., Paygude, A., Jain, P., Memon, A., & Naik, V. (2022). Hand gesture recognition using YOLO models for hearing and speech impaired people. In *2022 IEEE students conference on engineering and systems* (pp. 1–6).

Shanableh, T., Assaleh, K., & Al-Rousan, M. (2007). Spatio-temporal feature-extraction techniques for isolated gesture recognition in Arabic sign language. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, *37*(3), 641–650.

Sharma, S., & Kumar, K. (2021). ASL-3DCNN: American sign language recognition technique using 3-d convolutional neural networks. *Multimedia Tools and Applications*, *80*(17), 26319–26331.

Sharma, K., Kumar, B., Sehgal, D., & Kaushik, A. (2022). A comprehensive analysis on technological approaches in sign language recognition. In *Emergent converging technologies and biomedical systems: Select proceedings of ETBS 2021* (pp. 349–361). Springer.

Shin, J., Matsuoka, A., Hasan, M. A. M., & Srizon, A. Y. (2021). American sign language alphabet recognition by extracting feature from hand pose estimation. *Sensors*, *21*(17), 5856.

Shubhangi, D., & Shinde, G. (2017). Gesture to speech conversion for sign language recognition. *International Journal of Innovations & Advancement in Computer Science*, *6*.

Stoll, S., Camgoz, N. C., Hadfield, S., & Bowden, R. (2018). Sign language production using neural machine translation and generative adversarial networks. In *Proceedings of the 29th British machine vision conference*. British Machine Vision Association.

Stoll, S., Camgoz, N. C., Hadfield, S., & Bowden, R. (2020). Text2Sign: towards sign language production using neural machine translation and generative adversarial networks. *International Journal of Computer Vision*, *128*(4), 891–908.

Sung, J., Ponce, C., Selman, B., & Saxena, A. (2012). Unstructured human activity detection from rgbd images. In *2012 IEEE international conference on robotics and automation* (pp. 842–849). IEEE.

Tan, Y. S., Lim, K. M., & Lee, C. P. (2021). Hand gesture recognition via enhanced densely connected convolutional neural network. *Expert Systems with Applications*, *175*, Article 114797.

Tao, W., Leu, M. C., & Yin, Z. (2018). American sign language alphabet recognition using convolutional neural networks with multiview augmentation and inference fusion. *Engineering Applications of Artificial Intelligence*, *76*, 202–213.

Tolba, M., & Elons, A. (2013). Recent developments in sign language recognition systems. In *2013 8th international conference on computer engineering & systems* (pp. xxxvi–xlii). IEEE.

Tyagi, A., & Bansal, S. (2022). Sign language recognition using hand mark analysis for vision-based system (HMASL). In *Emergent converging technologies and biomedical systems: Select proceedings of ETBS 2021* (pp. 431–445). Springer.

Vasani, N., Autee, P., Kalyani, S., & Karani, R. (2020). Generation of indian sign language by sentence processing and generative adversarial networks. In *2020 3rd international conference on intelligent sustainable systems* (pp. 1250–1255). IEEE.

Von Agris, U., Knorr, M., & Kraiss, K.-F. (2008). The significance of facial features for automatic sign language recognition. In *Automatic face & gesture recognition, 2008. FG'08. 8th IEEE international conference on* (pp. 1–6). IEEE.

Wadhawan, A., & Kumar, P. (2021). Sign language recognition systems: A decade systematic literature review. *Archives of Computational Methods in Engineering*, *28*, 785–813.

Wan, J., Zhao, Y., Zhou, S., Guyon, I., Escalera, S., & Li, S. Z. (2016). Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 56–64).

Wang, J., Liu, Z., Wu, Y., & Yuan, J. (2012). Mining actionlet ensemble for action recognition with depth cameras. In *2012 IEEE conference on computer vision and pattern recognition* (pp. 1290–1297). IEEE.

Wang, R. Y., & Popović, J. (2009). Real-time hand-tracking with a color glove. *ACM transactions on graphics (TOG)*, *28*(3), 1–8.

Wang, & Wang, K.-C. (2007). Hand posture recognition using adaboost with sift for human robot interaction. In *Recent progress in robotics: Viable robotic service to human: An edition of the selected papers from the 13th international conference on advanced robotics* (pp. 317–329).

Wang, F., Zeng, Z., Sun, S., & Liu, Y. (2020). Diversity amplification and data generation of Chinese sign language based on generative adversarial network. In *2020 10th Institute of Electrical and Electronics Engineers international conference on cyber technology in automation, control, and intelligent systems* (pp. 139–145). IEEE.

Warrier, K. S., Sahu, J. K., Halder, H., Koradiya, R., & Raj, V. K. (2016). Software based sign language converter. In *Communication and signal processing (ICCSP), 2016 international conference on* (pp. 1777–1780). IEEE.

Wen, F., Zhang, Z., He, T., & Lee, C. (2021). AI enabled sign language recognition and VR space bidirectional communication using triboelectric smart glove. *Nature Communications*, *12*(1), 5378.

Wilbur, R., & Kak, A. C. (2006). *Purdue RVL-SLLL American sign language database*: *ECE technical reports paper 338*, Purdue University.

Wong, W., Juwono, F. H., & Khoo, B. T. T. (2021). Multi-features capacitive hand gesture recognition sensor: a machine learning approach. *IEEE Sensors Journal*, *21*(6), 8441–8450.

World Health Organization (2023). Deafness and hearing loss. https://www.who.int/health-topics/hearing-loss.

Yang, H.-D., & Lee, S.-W. (2011). Combination of manual and non-manual features for sign language recognition based on conditional random field and active appearance model. In *2011 international conference on machine learning and cybernetics, vol. 4* (pp. 1726–1731). IEEE.

Yang, H.-D., & Lee, S.-W. (2013). Robust sign language recognition by combining manual and non-manual features based on conditional random field and support vector machine. *Pattern Recognition Letters*, *34*(16), 2051–2056.

Zheng, X., Guo, Y., Huang, H., Li, Y., & He, R. (2020). A survey of deep facial attribute analysis. *International Journal of Computer Vision*, 1–33.

Zhou, Z., Chen, K., Li, X., Zhang, S., Wu, Y., Zhou, Y., et al. (2020). Sign-to-speech translation using machine-learning-assisted stretchable sensor arrays. *Nature Electronics*, *3*(9), 571–578.

Zhou, H., Zhou, W., Qi, W., Pu, J., & Li, H. (2021). Improving sign language translation with monolingual data by sign back-translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1316–1325).

Zhou, H., Zhou, W., Zhou, Y., & Li, H. (2020). Spatial-temporal multi-cue network for continuous sign language recognition. In *AAAI* (pp. 13009–13016).