Video Sign Language Recognition - A Survey

1st Dr. R Shreemathy
Department of Electronics and Telecommunication
Pune Institute of Computer Technology
Pune, India
rsreemathy@pict.edu

2nd Dr. Jayshree Jagdale
Department of Information Technology
Pune Institute of Computer Technology
Pune, India
jbjagdale@pict.edu

3rd Vedant Dattatray Kulkarni
Department of Information Technology
Pune Institute of Computer Technology
Pune, India
vedantk60@gmail.com

4th Anushree Shrigopal Bajaj
Department of Information Technology
Pune Institute of Computer Technology
Pune, India
anushreebajaj01@gmail.com

5th Himanshu Marathe
Department of Information Technology
Pune Institute of Computer Technology
Pune, India
himanshumarathe09@gmail.com

6th Mandar Deshmukh
Department of Information Technology
Pune Institute of Computer Technology
Pune, India
mandar.md30@gmail.com

***Abstract*—Sign language is a crucial mode of communication for the Deaf and Hard of Hearing (DHH) community. Automated systems that can recognize and comprehend various sign languages can greatly increase accessibility and inclusion. Despite their significant applications, this community often faces substantial barriers to communication, education, employment, and overall participation in society. Traditional approaches on sign language recognition mostly use superficial models such as rule-based or machine learning models, which struggle to capture the intricate and dynamic nature of sign language. Advanced CNNs are useful in these scenarios as they provide a more optimized path for sign language recognition due to their ability to model long-range dependencies and temporal sequences effectively and efficiently. In recent years, the task of sign language recognition has been handled with the help of deep learning models and architectures such as CNN, RNN, MTCNN, etc. to name a few. Additionally, the novel transformer architecture has also been found effective for sign language recognition when trained properly with huge volumes of data. This paper aims to explore the available methods for video sign language recognition, compare them across various factors such as accuracy, dataset size, architecture, training time, etc. and identify gaps between them.**

***Index Terms*—Sign Language Recognition, Deep Learning, Video Sign Language Recognition, Sign Language Translation, Transformer Architecture, CNN**

I. Introduction

Sign Language Recognition plays a pivotal role in enhancing communication accessibility for the Deaf and Hard of Hearing (DHH) community. For individuals within this community, sign language serves as a fundamental means of expression and connection. The development of automated systems capable of recognizing and comprehending various sign languages holds the promise of significantly reducing communication barriers and increasing inclusion. Such systems have far-reaching applications, including education, employment, and social participation, making them an essential component of fostering a more inclusive society.

Traditional approaches to sign language recognition have predominantly relied on rule-based or conventional machine learning models. However, these methods often struggle to capture the intricate and dynamic nature of sign language, which involves nuanced gestures and complex grammatical structures. The advent of advanced Convolutional Neural Networks (CNNs) offers a more optimized pathway for sign language recognition. These deep learning models excel in modeling long-range dependencies and temporal sequences effectively and efficiently, which are essential for understanding the temporal and spatial characteristics of sign language.

In recent years, the field of sign language recognition has witnessed a transformative shift towards the adoption of deep learning models and architectures. These include not only CNNs but also Recurrent Neural Networks (RNNs), Multi-Task Convolutional Neural Networks (MTCNN), and notably, the novel Transformer architecture. When appropriately trained on extensive datasets, these models have demonstrated their effectiveness in recognizing sign language, marking a paradigm shift in this domain.

This paper embarks on the journey to explore and assess the existing methods for video sign language recognition. It undertakes a comprehensive comparison across various factors, including recognition accuracy, dataset size, architectural choices, training time, and other relevant aspects. By doing so, it seeks to identify gaps and potential areas for improvement within the field of sign language recognition, with the ultimate

Identify applicable funding agency here. If none, delete this.

goal of advancing the state of the art and promoting greater accessibility and inclusivity for the DHH community.

II. Literature Survey

Jie Huang et al. [1] addresses the key challenges in Sign Language Recognition (SLR), focusing on both isolated word recognition and continuous sentence translation. It introduces the innovative Hierarchical Attention Network with Latent Space (LS-HAN) framework, eliminating the need for error-prone temporal segmentation and improving the accuracy of continuous SLR. The research also presents a novel two-stream 3D CNN [10] for precise video feature representation and gesture detection. By jointly optimizing relevance and recognition loss, the LS-HAN framework enhances the overall efficiency of SLR. Additionally, the creation of a comprehensive open-source dataset for Modern Chinese Sign Language with sentence-level annotations strengthens the foundation of SLR research.

Hidden Markov Models (HMMs), known for their success in speech and handwriting recognition, prove effective in recognizing complex hand gestures, as in sign language [12]. This paper presents two real-time experiments using HMMs to recognize American Sign Language (ASL) sentences without modeling fingers explicitly. In the first experiment, the system achieves 99 percent word accuracy by tracking hands with colored gloves. In the second experiment, without gloves, it attains 92 percent word accuracy, both using a 40-word lexicon. This adaptable system utilizes a single color camera to track hands and interpret ASL, taking shape, orientation, and trajectory information into an HMM for word recognition. [2]

Prof. Sandeep Samleti et al. [3] in his research paper explores two models: the Hierarchical model combined with steered captioning and the Multi-stream Hierarchical Boundary Model. The Hierarchical model captures clip-level temporal features at fixed intervals in videos, while the Multi-stream Hierarchical Boundary Model combines fixed and soft hierarchies to define video clips. The Steered captioning model employs an attention mechanism to focus on relevant video locations using visual parameters. Additionally, the paper discusses parametric Gaussian attention. Notably, Gaussian attention overcomes the limitation of fixed-length video streams in soft attention techniques.

Dongxu Li et al. [4] discusses the challenges of limited training data in sign recognition and proposes a method to leverage cross-domain knowledge from news sign videos to improve the performance of Weakly Supervised Sign Language Recognition (WSLR) models. The proposed method involves extracting sign words, coarsely aligning news and isolated signs, and using a prototypical memory to learn domain-invariant descriptors. A memory-augmented temporal attention module is used to enhance classification performance [13]. The goal is to address the data insufficiency issue in WSLR and improve models by collecting low-cost data from the internet. The study analyzes the impact of coarse domain alignment and the use of cross-domain knowledge through prototypical memory, highlighting the importance of using news signs for memory to achieve the best performance.

In this research, data collection is emphasized as a crucial aspect of the study's success, with the LSA64 public dataset used, featuring 10 vocabularies performed by 10 signers, each repeating the signs 5 times for a total of 500 videos. The videos were transformed into image sequences, resized, and normalized for processing using a 3D Convolutional Neural Network Architecture known as i3d inception. The dataset was divided into a 6:2:2 ratio for training, validation, and testing with 300 videos for training, 100 for validation, and 100 for testing. Initially, training produced low accuracy, but after training with a single signer for 10 classes, accuracy reached 100 percent. Subsequent training with different dataset structures yielded varying accuracy, with the highest being 100 percent for two signers and 20 classes, but lower for four signers at 20.00

Starner et al. [6] has presented two real-time sign language recognition and translation systems based on Hidden Markov Model(HMM) architecture. The first system captures the signer from a second person perspective which is obtained through a camera placed at desk level in front of the signer. This system achieves a word-accuracy of 92 percent. The second system uses a mounted camera on the signer's cap or head gear. The camera captures the first person view of the signer. The hand signs appear as seen from the signers eyes. It achieves a word-accuracy of 98 percent which is significantly more than the first system.

The study presents a transformative approach that amalgamates Continuous Sign Language Recognition (CSLR) and Sign Language Translation (SLT) within a transformer-based framework. This method obviates the necessity for precise temporal annotations, leading to substantial improvements in both recognition and translation aspects. Notably, the research attains exceptional performance on the demanding PHOENIX14T dataset, surpassing extant models by a significant margin. Emphasizing the intricacies of sign language mapping, the paper introduces two pivotal components, the Sign Language Recognition Transformer (SLRT) and the Sign Language Translation Transformer (SLTT), designed to address the inherent challenges in this domain. The research yields substantial contributions by introducing a multi-task formalization, leveraging transformer technology, and establishing baseline performance benchmarks for future investigations within this field of study. [7], [14]

III. Methodology

In this paper, we have compared the various approaches to video sign language recognition and synthesized the findings to provide a comprehensive overview of the state-of-the-art techniques, datasets, evaluation metrics, and challenges in the field. Our analysis sheds light on the strengths and weaknesses of different methods, highlighting promising avenues for future research and the potential impact of video sign language recognition in applications ranging from assistive technology to human-computer interaction. The current review is categorized

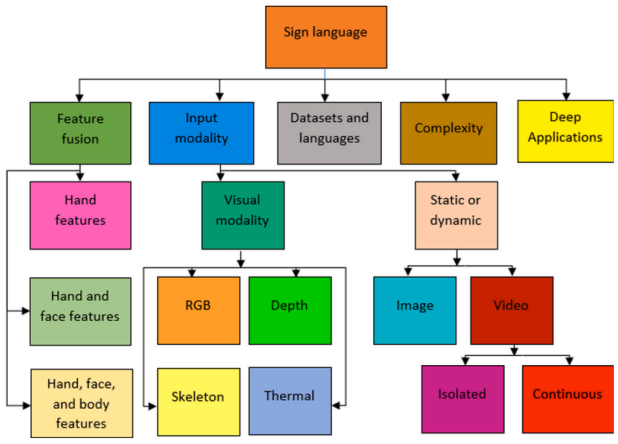

Fig. 1. Taxonomy of deep models for sign language recognition.

into five broad dimensions viz. *Dataset*, *Model Architecture*, *Training Time*, *Level*, *Dataset*,

A. Video-Based Sign Language Recognition without Temporal Segmentation:

Problem Addressed: The research addresses the challenges of isolated Sign Language Recognition (SLR) for word recognition and continuous SLR for sentence translation. It focuses on eliminating the complexities of temporal segmentation and labeling each word separately in sentences, making continuous SLR more efficient and user-friendly.

Exploited Dataset Details: The paper conducts experiments on two large-scale datasets, including a significant open-source Modern Chinese Sign Language (CSL) dataset [9] with sentence-level annotations. These datasets enable a comprehensive evaluation of the proposed LS-HAN framework.

Feature Representation and Selection Method: To tackle continuous SLR, the research introduces the innovative Hierarchical Attention Network with Latent Space (LS-HAN). This approach leverages a two-stream Convolutional Neural Network (CNN) for video feature representation, capturing both global and local information using a two-stream 3D CNN. This combination of features provides a holistic representation of sign language gestures.

Obtained Results: The LS-HAN framework, along with the 3D CNN, significantly improves continuous SLR, offering a more efficient and user-friendly approach to sign language recognition. This is a noteworthy advancement in the field.

Indicated Future Directions: The paper suggests extending LS-HAN for handling longer compound sentences and real-time translation tasks, which can be a promising area of research to make sign language recognition more accessible and practical.

B. Real-Time American Sign Language Recognition from Video Using Hidden Markov Models:

Problem Addressed: This paper aims to provide real-time American Sign Language (ASL) recognition, addressing the challenge of recognizing complex and structured hand gestures that are prevalent in sign language.

Exploited Dataset Details: The paper details two experiments using Hidden Markov Models (HMMs) for ASL sentence recognition [12]. The experiments focus on tracking hands, one using colored gloves and the other based on natural skin tone, making it versatile for different scenarios.

Feature Representation and Selection Method: The paper utilizes HMMs combined with real-time image processing to track hand shapes, orientations, and trajectories without explicit finger modeling. This approach simplifies the recognition process while maintaining high accuracy.

Obtained Results: The system achieves word accuracy of 99 percent with colored gloves and 92 percent without gloves, showcasing the effectiveness of HMMs for ASL recognition in real-time applications.

Indicated Future Directions: The paper suggests that HMMs provide a robust solution for real-time ASL recognition and encourages further exploration of this approach for broader applications, potentially involving wider vocabulary and real-world ASL communication.

C. Real Time Video Captioning Using Deep Learning:

Problem Addressed: This paper addresses the challenge of video captioning by introducing and discussing two distinct models: the Hierarchical model and the Multi-stream Hierarchical Boundary Model. These models aim to capture temporal features and incorporate attention mechanisms, making video captioning more effective.

Exploited Dataset Details: While the paper does not specify the dataset used, it focuses on the development of models for video captioning, which can be applied to various video datasets.

Feature Representation and Selection Method: The paper introduces the Hierarchical model, the Multi-stream Hierarchical Boundary Model, and steered captioning to capture temporal features and guide attention in video content. These models collectively offer a comprehensive approach to video captioning.

Obtained Results: The paper provides insights into models for video captioning but does not present specific results. It primarily lays the groundwork for effective video captioning solutions.

Indicated Future Directions: Future research can explore the integration of these models and attention mechanisms for more effective video captioning systems. This opens the door to developing practical video captioning applications for diverse contexts.

D. Transferring Cross-domain Knowledge for Video Sign Language Recognition:

Problem Addressed: This paper aims to enhance the performance of Word-Level Sign Language Recognition (WSLR) models by transferring knowledge from news sign examples, addressing the challenge of limited training data.

Exploited Dataset Details: The research utilizes news sign examples, which are easily accessible from the web, to transfer cross-domain knowledge to WSLR models. This approach provides a broader and more diverse training dataset.

Feature Representation and Selection Method: The paper develops a method for transferring cross-domain knowledge, focusing on aligning news sign representations with isolated signs. The use of prototypical memory plays a pivotal role in knowledge transfer.

Obtained Results: The approach effectively transfers cross-domain knowledge, significantly improving the performance of WSLR models [13]. This advancement addresses the challenge of limited training data for sign recognition.

Indicated Future Directions: The study highlights the potential of cross-domain knowledge transfer and suggests further research, particularly in optimizing the prototypical memory and domain alignment techniques, which can make WSLR models even more robust and accurate.

E. Sign Language Recognition Using Modified Convolutional Neural Network Model:

Problem Addressed: This paper focuses on improving Sign Language Recognition using a modified Convolutional Neural Network (CNN) model and explores the impact of different dataset structures on recognition accuracy.

Exploited Dataset Details: The dataset is distributed into training, validation, and testing sets, with varying signer and class combinations, making it suitable for evaluating different recognition scenarios.

Feature Representation and Selection Method: The paper utilizes a modified 3D CNN model, specifically the i3d inception architecture, known for its effectiveness in recognizing complex patterns and gestures in video data.

Obtained Results: The paper reports varying recognition accuracy based on dataset structures and signer-class combinations, demonstrating the influence of these factors on recognition performance.

Indicated Future Directions: The study encourages further exploration of dataset structures and model modifications to improve recognition accuracy, particularly in more diverse signer-class combinations, with the potential to make sign recognition systems more adaptable to real-world scenarios.

F. Two Hidden Markov Models:

Problem Addressed: This research introduces real-time sign language recognition systems based on Hidden Markov Models (HMMs), capturing sign language from a first-person perspective and achieving high word accuracy.

Exploited Dataset Details: The paper uses video data to develop two real-time sign language recognition systems, which can be applied to various datasets capturing sign language gestures.

Feature Representation and Selection Method: The paper focuses on capturing the evolution of hand gestures over time, emphasizing coarse hand shape descriptions instead of fine-grained hand shape modeling. This approach simplifies the recognition process, making it more suitable for real-time applications.

Obtained Results: The systems achieve high word accuracy, with the second system outperforming the first due to the first-person perspective, demonstrating the viability of HMMs for real-time sign language recognition.

Indicated Future Directions: The paper opens possibilities for recognizing complex ASL sentences and further user-independent ASL lexicon development using these real-time recognition techniques. This suggests potential advancements in real-time communication and accessibility for the Deaf and Hard of Hearing communities.

G. Transformer-Based Sign Language Recognition and Translation:

Problem Addressed: This paper introduces a transformer-based architecture for Continuous Sign Language Recognition (CSLR) and Sign Language Translation (SLT) in an end-to-end learning approach, addressing the complexity of mapping sign language to spoken language and improving the recognition and translation of sign language [14].

Exploited Dataset Details: The paper evaluates the system on the PHOENIX14T dataset, a challenging dataset for CSLR and SLT. This dataset facilitates comprehensive testing of the proposed approach.

Feature Representation and Selection Method: The paper utilizes transformer networks for CSLR and SLT, eliminating the need for ground-truth timing information and achieving remarkable improvements in the recognition and translation of sign language.

Obtained Results: The Sign Language Transformers outperform existing models, showcasing substantial improvements in recognition and translation, particularly in the challenging task of mapping sign language to spoken language.

Indicated Future Directions: The paper emphasizes the complexities of sign language mapping and outlines the need for mid-level sign gloss representations. It introduces novel multi-task formalizations for CSLR and SLT, offering an exciting avenue for future research and development in the field. The results set a strong baseline for further exploration of text-to-text sign language translation tasks, promising further advancements in the domain of sign language accessibility and communication.

H. Figures and Tables

TABLE I
MODELS PROPOSED AND ACCURACY

Ref.	***Model proposed***	***Accuracy***
[1]	Hierarchical Attention Network with Latent Space (LS-HAN)	0.827
[2]	Hidden Markov models	0.92
[3]	Hierarchical model, MSHB,parametric Gaussian attention	N/A
[4]	WSLR models,RCNN and I3D	N/A
[5]	CNN model	0.917
[6]	real-time hidden Markov model	0.92

Dataset Name	*Used By*	*Opensource*
RWTH-PHOENIX	[1], [7]	Yes
WLASL	[3], [16], [17], [18], [19]	Yes
LSA64	[8], [20], [21]	Yes
Own Dataset	[15]	No
CLAP14	[5]	Yes

IV. Conclusion

This survey paper offers an insightful overview of recent advancements in video sign language recognition (SLR). It systematically categorizes and summarizes twenty-one recently published and highly cited articles in the field. These articles contribute significantly to various aspects of sign language recognition, employing a wide range of techniques for real-time video captioning. Upon a careful analysis of these articles, it becomes evident that there is a continuous need for further development and innovation in algorithmic approaches, making it a dynamic and open field for ongoing research. These surveyed articles serve as valuable reference models against which many newly proposed algorithms are compared, highlighting their importance in shaping the state of the art in SLR. Furthermore, the existence of benchmark datasets, such as WLASL and PHOENIX14T, plays a critical role in evaluating the performance of these algorithms, ensuring rigorous and standardized assessment procedures in the field.

V. Acknowledgment

References

[1] Huang, Jie, et al. "Video-based sign language recognition without temporal segmentation." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32. No. 1. 2018.

[2] Starner, Thad, and Alex Pentland. "Real-time american sign language recognition from video using hidden markov models." Proceedings of International Symposium on Computer Vision-ISCV. IEEE, 1995.

[3] Sahoo, Ashok K., Gouri Sankar Mishra, and Kiran Kumar Ravulakollu. "Sign language recognition: State of the art." ARPN Journal of Engineering and Applied Sciences 9.2 (2014): 116-134.

[4] Li, Dongxu, et al. "Transferring cross-domain knowledge for video sign language recognition." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[5] Gunawan, Herman, Narada Thiracitta, and Ariadi Nugroho. "Sign language recognition using modified convolutional neural network model." 2018 Indonesian Association for Pattern Recognition International Conference (INAPR). IEEE, 2018.

[6] Starner, Thad, Joshua Weaver, and Alex Pentland. "Real-time american sign language recognition using desk and wearable computer based video." IEEE Transactions on pattern analysis and machine intelligence 20.12 (1998): 1371-1375.

[7] Camgoz, Necati Cihan, et al. "Sign language transformers: Joint end-to-end sign language recognition and translation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[8] Boháček, Matyáš, and Marek Hrúz. "Sign pose-based transformer for word-level sign language recognition." Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2022.

[9] Hao Zhou, Wengang Zhou, Weizhen Qi, Junfu Pu, and Houqiang Li, "Improving Sign Language Translation with Monolingual Data by Sign Back-Translation," IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[10] Al-Hammadi, Muneer, et al. "Hand gesture recognition for sign language using 3DCNN." IEEE access 8 (2020): 79491-79509.

[11] Xiao, Qinkun, et al. "Multimodal fusion based on LSTM and a couple conditional hidden Markov model for Chinese sign language recognition." IEEE Access 7 (2019): 112258-112268.

[12] Juang, Biing Hwang, and Laurence R. Rabiner. "Hidden Markov models for speech recognition." Technometrics 33.3 (1991): 251-272.

[13] Sridhar, Advaith, et al. "Include: A large scale dataset for indian sign language recognition." Proceedings of the 28th ACM international conference on multimedia. 2020.

[14] Koller, Oscar, Jens Forster, and Hermann Ney. "Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers." Computer Vision and Image Understanding 141 (2015): 108-125.

[15] Huang, Jie, et al. "Sign language recognition using 3d convolutional neural networks." 2015 IEEE international conference on multimedia and expo (ICME). IEEE, 2015.

[16] Hu, Hezhen, et al. "SignBERT: pre-training of hand-model-aware representation for sign language recognition." Proceedings of the IEEE/CVF international conference on computer vision. 2021.

[17] Tunga, Anirudh, Sai Vidyaranya Nuthalapati, and Juan Wachs. "Pose-based sign language recognition using GCN and BERT." Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2021.

[18] Du, Yao, et al. "Full transformer network with masking future for word-level sign language recognition." Neurocomputing 500 (2022): 115-123.

[19] Eunice, Jennifer, Yuichi Sei, and D. Jude Hemanth. "Sign2Pose: A Pose-Based Approach for Gloss Prediction Using a Transformer Model." Sensors 23.5 (2023): 2853.

[20] Mindlin, Iván, et al. "A Comparison of Neural Networks for Sign Language Recognition with LSA64." Conference on Cloud Computing, Big Data Emerging Topics. Cham: Springer International Publishing, 2021.

[21] Konstantinidis, Dimitrios, Kosmas Dimitropoulos, and Petros Daras. "A deep learning approach for analyzing video and skeletal features in sign language recognition." 2018 IEEE international conference on imaging systems and techniques (IST). IEEE, 2018.