A Methodological and Structural Review of Hand Gesture Recognition Across Diverse Data Modalities

Publisher: IEEE Cite This PDF

Jungpil Shin ; Abu Saleh Musa Miah ; Md. Humaun Kabir ; Md. Abdur Rahim ; Abdullah Al Shiam **All Authors**

Abstract:
Researchers have been developing Hand Gesture Recognition (HGR) systems to enhance natural, efficient, and authentic human-computer interaction, especially benefiting those who rely solely on hand gestures for communication. Despite significant progress, automatic and precise identification of hand gestures remains a considerable challenge in computer vision. Recent studies have focused on specific modalities like RGB images, skeleton data, and spatiotemporal interest points. This paper comprehensively reviews HGR techniques and data modalities from 2014 to 2024, exploring advancements in sensor technology and computer vision. We highlight accomplishments using various modalities, including RGB, Skeleton, Depth, Audio, Electromyography (EMG), Electroencephalography (EEG), and Multimodal approaches and identify areas needing further research. We reviewed over 250 articles from prominent databases, focusing on data collection, data settings, and gesture representation. Our review assesses the efficacy of HGR systems through their recognition accuracy and identifies a gap in research on continuous gesture recognition, indicating the need for improved vision-based gesture systems. The field has experienced steady research progress, including advancements in hand-crafted features and deep learning (DL) techniques. Additionally, we report on the promising developments in HGR methods and the area of multimodal approaches. We hope this survey will serve as a potential guideline for diverse data modality-based HGR research.

The Basic Flowgraph of the HGR Research Work Across Diverse Data Modalities

Published in: IEEE Access (Volume: 12)

Page(s): 142606 - 142639

DOI: 10.1109/ACCESS.2024.3456436

Date of Publication: 09 September 2024

Publisher: IEEE

Electronic ISSN: 2169-3536

Funding Agency:

SECTION I.
Introduction

In our daily interactions, non-verbal communication plays a crucial role, conveying approximately 65% of human messages, compared to verbal communication, which accounts for only 35% [1], [2]. Nowadays, people increasingly use gestures to control daily devices such as televisions, computers, fans, and air-conditioning systems. Non-verbal communication includes body gestures like head movements, facial

expressions, nodding, shaking the head, mouth movements, winking, eye gaze direction, body movements, arm gestures, and hand gestures. Effective HGR methods ensure robust human-computer interaction (HCI), offering alternatives to traditional tools like mice and keyboards [3]. Hand gestures are essential in daily activities, and automatic HGR is vital for natural nonverbal communication. SLR is particularly important in bridging the gap between hearing impaired and general communities by translating hand movements into speech or text, which is invaluable for communication, education, and rehabilitation, especially without a human interpreter. Despite extensive research on static and dynamic HGR systems for various applications, several challenges remain. These challenges include adapting to diverse inputs like environmental noise, signer variations, and language differences [4].

Constraints during development often involve the signer's environment to mitigate segmentation and tracking issues. This helps manage gesture transitions in continuous sign language [5], [6], but it is difficult and can lead to incorrect recognition results. This complexity limits the practicality of vision-based gesture recognition in real-world settings [7], [8]. Creating robust signer-independent HGR systems is another significant challenge. Such systems should work with users who were not part of the training phase, enabling broad applicability without requiring individual training for each new user [9]. Previous research has extensively covered both device-based and vision-based HGR systems. While vision-based systems are ideal for diverse real-world applications, existing reviews often provide broad overviews without focusing on specific advancements or future directions. To address this gap, this paper thoroughly reviews current and historical literature, analyzing advancements in vision-based HGR systems and exploring potential future research directions.

A. Background

HGR is a technology that translates hand movements in sign language into text or speech. It can be divided into vision-based and device-based systems based on how they capture hand gestures [10], [11]. Vision-based HGR systems offer a more natural interaction experience as users do not need to wear any cumbersome devices. These systems find wider applications in outdoor settings due to their ease of use. However, challenges arise in handling dynamic sign language datasets containing both isolated and continuous gestures. Existing research focuses on recognizing isolated gestures, limiting their practical applicability. More robust feature extraction and discrimination methods are necessary to enhance vision-based systems. A temporal modelling-based HGR system is also needed. Due to numerous applications, HGR has sparked significant research interest, as highlighted in numerous review papers [12], [13], [14], [15].

In 2013, Chen et al. surveyed the HGR methodology, vision-based, depth-based and glove-based approach [12]. Check et al. surveyed the state-of-the-art approach used in the hand gesture-based recognition system in 2019. They summarized the hand feature recognition system's preprocessing, segmentation, augmentation and classification technique [16]. Aloysius et al. surveyed only vision-based video or continuous sign language recognition (CSLR system) in 2020 [14]. By including the dynamic dataset-based HGR, Wadhawan et al. surveyed academic literature spanning from 2007 to 2017, where they included six key dimensions such as dataset collection approach, different types of signs based on time, mode of sign, one-hand or two-hand sign, classification approach and rates of recognition in 2021 [13]. Ratsgoo et al. surveyed vision-based HGR for SLR in 2021 [15]. Recent surveys have further expanded the field, such as Jain et al. provided a comprehensive review of DL approaches for HGR in 2022, focusing on advancements in model architectures and applications in human-computer interaction [17]. In 2024, Jing et al. reviewed vision-based data modalities for HGR, focusing mainly on the monocular and RGB-D cameras, discussing data acquisition, hand gesture detection and segmentation, feature extraction, and gesture classification. It also reviews experimental evaluations and highlights advances required to improve current HGR systems for more effective and efficient HRI. [18]. Taw et al. provided a comprehensive review of DL-based gesture recognition, highlighting the roles of Convolutional Neural Network (CNN), Long long-term memory (LSTM) networks, and transfer learning. The paper discusses the strengths and limitations of each technique and their applications in various fields, including human-computer interaction, virtual reality, healthcare, and smart homes [19]. By integrating these recent surveys, we found that no existing paper has studied the various modality-based HGR research work based on recent publication articles.

B. Article Search and Survey Methodology

To identify relevant articles on multimodality-based HGR, we employed a targeted search strategy using specific keywords. The focus was on the following terms:

- Vision and sensor-based recognition of static and dynamic hand gestures and sign language;

- Skeleton-based HGR and SLR;
- Multimodal dataset fusion-based HGR and SLR.

We sourced articles from esteemed databases to ensure a comprehensive review of pertinent literature. The databases included IEEE Xplore Digital Library, MDPI, ScienceDirect, Springer Link, ResearchGate, and Google Scholar. To refine and ensure relevance in our initial search results, we applied the following criteria:

- Publication date between 2014 and 2024;
- Inclusion of journals, proceedings, and book chapters;
- Focus on various data modality-based HGR systems, including RGB image, video, skeleton, depth, EMG, EEG, audio, and multimodal fusion modalities;
- Exclusion of studies where HGR in SLR are only mentioned tangentially and lacking in-depth information about their experimental procedures;
- Exclusion of research articles where the complete text isn't accessible, both in physical and digital formats;
- Exclusion of research articles that include opinions, keynote speeches, discussions, editorials, tutorials, remarks, introductions, viewpoints, and slide presentations.

Using the search keywords outlined in this methodology, we identified over 250 articles that met our inclusion and exclusion criteria. Figure 1 demonstrates the article selection procedure.

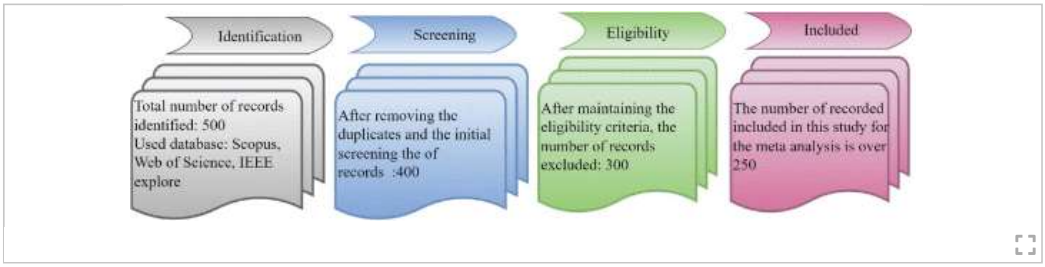
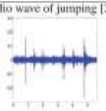


FIGURE 1.
Article selection process procedure.

Figures 2, 3, and 4 show the types of references, year-wise references, and data modalities based on the number of references, respectively. In the survey methodology, we reviewed each article through a process involving abstract review, methodology analysis, discussion, and result evaluations. Different modalities used in HGR (HGR) have unique features, each with its own set of advantages and disadvantages, as summarized in Table 1.

TABLE 1 Characteristics, Advantages and Disadvantages of Different Modalities

Modality	Example	Pros	Cons
RGB Single image	Hand Gesture still image [2]. 	Provide static data from static images. Provide color distribution, color dominant, overall texture	Only be able to look at static information
RGB Video	Hand Gesture [20], [21]. 	Provide motion information. Give details about the rich appearance. Simple to use and obtain numerous uses	Receptive to opinions. Sensitive to context. Light sensitivity.
3D Skelton	Hand palm skeleton [22], [23]. 	Provide the subject's pose's 3D structural information. straightforward but instructive. Inconsiderate of the perspective. Disregard for the context.	Absence of details regarding appearance. Absence of comprehensive shape data.
Depth	Mopping floor [24]. 	Provide details on the 3D structure. Give details about geometric shapes.	Insufficient details about texture and colour. Restricted practical distance.
EMG Signal	EMG Hand Gesture [25]. 	EMG provides muscle contractor information. No need for light color or background.	Difficult to control muscle power. Sometimes appropriate information is costly
Audio	Audio wave of jumping [26]. 	Easy to locate actions in temporal sequence	Lack of appearance information
EEG Signal	EEG Hand gesture [27], [28]. 	EEG provide muscle contractor information. No need for light color or background.	Difficult to control muscle power. No need for light color or background

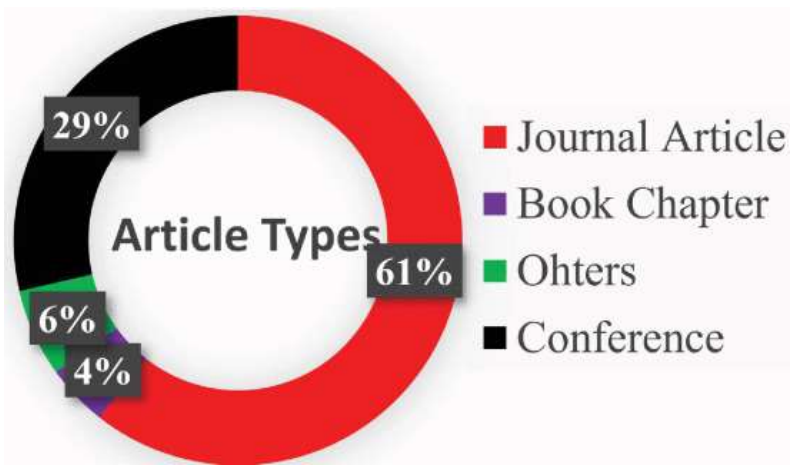


FIGURE 2.
Articles type journal, conference, book chapters, and others.

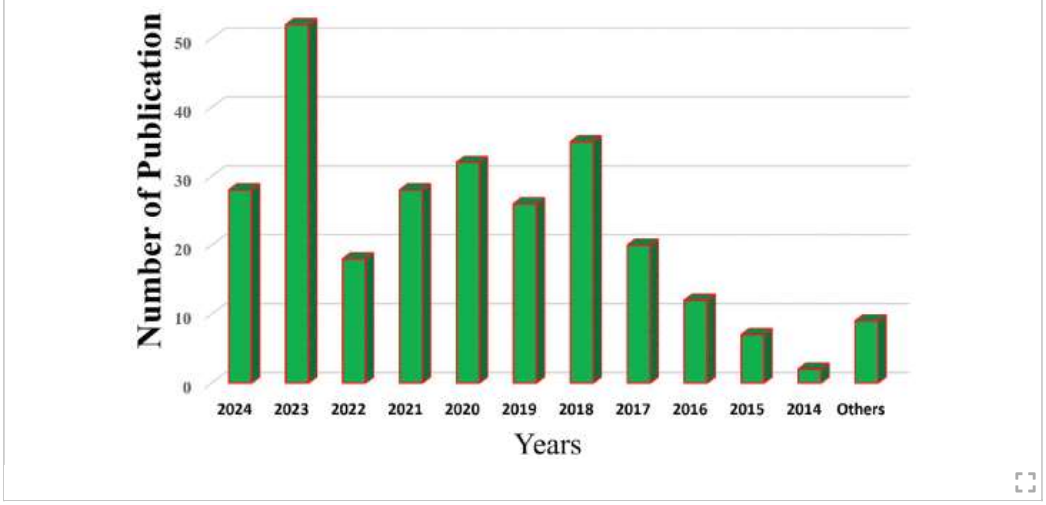


FIGURE 3.
Year-wise peer-reviewed publications used in the study.

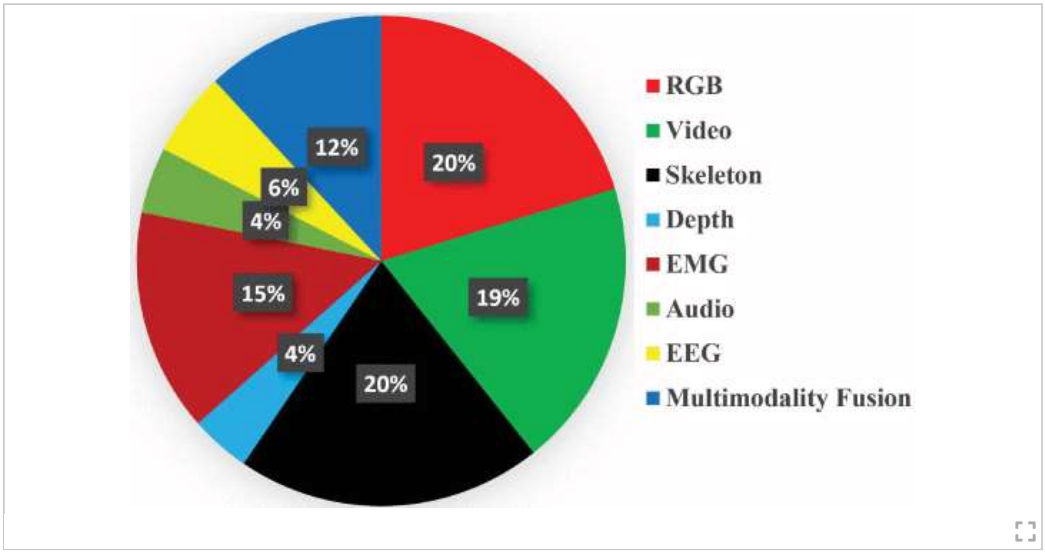


FIGURE 4.
Various data modalities based references.

C. Research Gap and New Research Challenges

While existing reviews have provided comprehensive overviews of HGR research, there remains a significant gap in the literature, focusing specifically on the advancements and potential avenues for multimodality-based HGR systems. Prior studies have largely overlooked the unique challenges and future directions necessary for developing robust and efficient multimodal HGR systems. Our research aims to address this gap by thoroughly reviewing the existing literature to identify the progress made in multimodal-based HGR systems. By exploring the trajectory of multimodal HGR research, we aim to provide valuable insights and guidance for future research directions.

D. Contribution

Figure 5 demonstrates the proposed methodology flowgraph, which was basically followed by HGR researchers. The main contributions of this research are given below:

- Comprehensive Review: We provide an extensive review of HGR systems, focusing on the evolution of data acquisition, data environments, and hand gesture portrayals from 2014 to 2024;
- Multimodal Analysis: Our study is the first to systematically examine the advancements in various data modalities based on HGR systems, which cover RGB, skeleton, depth, audio, EMG, EEG, and multimodal fusion;

- Identification of Gaps and Future Directions: We created the datasets table for each modality with the latest performance beside the performance table. Then, we identify significant gaps in current research and propose potential future research directions;

- Evaluation of System Efficacy: We assess the effectiveness of existing HGR systems by analyzing their recognition accuracy, providing a benchmark for future developments;

- Guidance for Practitioners: The insights and findings from our review offer practical guidance for researchers and practitioners aiming to develop more robust and accurate HGR systems. Our comprehensive survey study will serve as a valuable resource for researchers and practitioners in the field. It will offer insights into the current state-of-the-art techniques, highlight the challenges, and suggest potential future research directions. This will facilitate informed decision-making and advance the development of innovative HGR systems.

By addressing these contributions, our paper not only fills existing gaps in the literature but also paves the way for future advancements in the field of HGR.

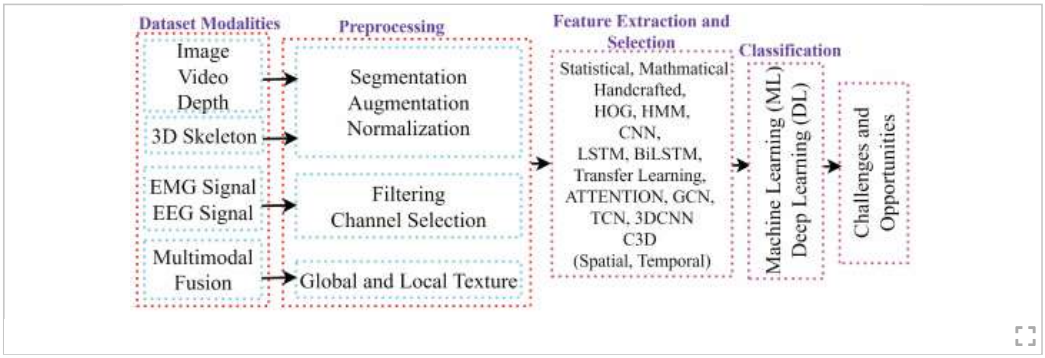


FIGURE 5.
The basic flowgraph of the HGR research work.

E. Research Question

To achieve this, we have formulated two primary research questions:

- **Research Question 1 (RQ1):** How have vision, sensor, and multimodality-based HGR systems evolved in areas such as data acquisition, data environment, and hand gesture portrayals from 2014 to 2024?

- **Research Question 2 (RQ2):** How effective have the current various data modality-based HGR systems been, and what could be the prospective trajectories in this domain?

F. Organization of the Study

The rest of the paper is organized as follows: Section II RGB-modality-based static and dynamic HGR work. Section III described the skeleton-modality based HGR. Section IV described the depth data modalities based on HGR. Section V describes the EMG-modality-based HGR. Section VI demonstrates the EEG signal-modality-based HGR, Section VII describes the audio signal-modality-based HGR, and Section VIII demonstrates the multimodal fusion-based HGR. Conclusion described in the Section XI.

SECTION II.
RGB-Modality Based HGR

RGB modal-based HGR involves processing still images or videos to identify static and dynamic gesture recognition, respectively. RGB dataset, mainly collected using cameras, is relatively simple to gather and provides detailed appearance information about the scene context [29], [30] An example of this is the gesture for the word WELCOME in Chinese Sign Language [31]. This modality can be either static or dynamic, where a single image can express the full meaning of the gesture and sequence of frames needed to completely express the dynamic gesture [14], [31], [32], [33]. Mobile cameras [34], [35], web cameras [36], and

specialized cameras such as HP Pavilion dv6 [37] are mainly used to record the dataset [38]. RGB data modality-based SL datasets have become popular due to their portability and cost-effectiveness over sensor-based systems [39]. Below, we included static and dynamic HGR using vision-based data modalities.

A. RGB Still Image

Static images belong to the RGB modality, where a single image can express the full meaning of the gesture, and this data modality is usually recorded by the camera. Figure 5 demonstrates the overview of the still image-based HGR steps, which mainly included preprocessing, feature extraction, and classification.

1) Dataset

Benchmark dataset is crucial for developing and evaluating the static HGR systems as demonstrated in Table 2. Table 2 included various benchmark datasets' names of this modality, creation year, number of classes, dataset types, sample size and latest performance accuracy. The key datasets include Korean sign language (KSL)-77, which includes 77 classes containing 112564 frames from 20 individuals [40], and American sign language (ASL)-10, featuring ten gestures from 14 individuals, each with ten instances [41]. The ASL-20 dataset comprises 20 ASL words with 18,000 frames [41], while the dataset created by Islam et al. consists of 26 signs from three persons, totalling 9,360 images [36]. Bengali sign language (BSL)-38 includes 38 classes with 320 images per class, created with both hearing impaired and general students [42]. The Irish Sign Language (ISL) dataset contains 58,114 images for 23 common hand shapes from ISL [43], [44]. The Japanese sign language (JSL) word Dataset features videos of JSL word motions, with frames converted to grayscale and analyzed [45]. Lastly, the Large-scale Argentine (LSA)-64 dataset, designed for Argentinian sign language research, comprises recordings of 64 distinct signs performed by 10 individuals, standardized to 48 frames per video [46]. These datasets provide a comprehensive foundation for advancing HGR technology.

TABLE 2 RGB Still Image Modality Based Dataset Description

Dataset Names	Year	Dataset Types	Classes	Subject	Total Sample	Sample Sign	Latest Performance Accuracy
OUHAND	2016	ASL	10	23	3000	-	98.56 [47]
NTU Dataset	2018	Digits	10	10	1000	100	99.00 [2]
ASL-10	2020	SL	10	22	2800	120	95.00 [2]
MUD	2021	ASL	36	-	2520	70	99.23 [48]
ASLAD	2021	ASL	29	-	87000	3000	99.00 [48]
NUS II	2021	ASL	10	40	2000	-	96.5 [49]
Marcel	2021	ASL	6	-	5531	-	96.57 [49]
HGR1	2021	ASL	25	12	899	-	93.36 [47]
ArSL [50]	2020	SL	23	-	-	-	95.00 [51]
ASL-20 [52]	2023	SL	20	5	18000	900	99.00
KSL-77 [40]	2022	SL	77	22	112,564	1461	99.00
KSL-20 [20]	2023	SL	20	25	96200	4800	99.00
BSL [42]	2019	SL	38	35	12160	320	96.00
JSL [51]	2023	SL	41	20	7380	1800	91.00 [51]
Creative Senz3d	2023	SL	11	10	1320	300	98.51 [53]
FMCW-SAR [54]	2023	HGR	5	Simulated	-	-	97.00 [54]
Custom ASL [55]	2023	HGR	10	7	1400	-	99.76

2) Methodology of the RGB Still Image Modality

Figure 5 demonstrated a common workflow diagram of the RGB still-image-based HGR. Table 3 demonstrates the performance of this data modality, including year, feature extraction and classification method performance. The preprocessing, feature extraction, machine learning (ML) and diverse explanations of the DL system are included below:

TABLE 3 Databases for RGB Still Image Modality

Authors	Year	Supported Language	Classes	Sample	Feature Extraction	Classifier	Performance
Tao et. al [38]	2018	American SL	24	500-600	CNN	CNN	84.80
Ibrahim et. al [57]	2018	Arabic SL	30	450	N/A	HMM	97.00
Ito et. al [45]	2020	Japanese SL	20	13200	CNN	MSVM	94.30
Sekda et. al [30]	2022	Turkish SL	32	19200	CNN	LSTM	99.75
Musa et. al [39]	2022	Bangla SL	36	19200	BenSignNet	Softmax	99.00
Wang et. al [59]	2023	China SL	15	720	CRNN	BiLSTM	92.40
Jain et. al [29]	2023	Danish SL	36	252	YOLOv5	CSP Darknet	92.00
Musa et. al [60]	2023	Hand Gesture/ASL	11	11000	Multi-stage CNN	CNN	99.00
Shin et. al [20]	2023	KSL	77	20000	Transformer	CNN	89.00
Noble et al. [61]	2023	Custom capacitive sensing dataset	5	5000	-	Decision Tree, Naive Bayes, MLP, CNN	MLP: 96.87 CNN: 95.94 DT: 91.18 NB: 88.34
Sahoo et al. [62]	2023	I: Custom, II: ASL-FS	6, 24	15,000 (Dataset I), 60,000+ (Dataset II)	DRCAM	Softmax	95.09 93.44
Baptista et al. [63]	2023	HGR	4	Tr:24611 T:17051	SMLT	Softmax	SLT: 84.23 MSMLT: 79.24 SMLT: 90.03
Mohyuddin et al. [64]	2023	Leap motion dataset	8	264 per	SSC-DNN	Softmax	98.7%
Garg et al. [65]	2023	NVGesture, Briareo	-	-	MsMHA-VTN	Softmax	88.22 99.10
Bhartiet al. [66]	2023	MINDS Libras	-	-	Key frame	3D CNN	98.00
Eid et al. [49]	2023	NUS II Marcel	10 6	2000 5531	Skin Segmentation Data Augmentation	CNN	NUS II: 96.5, Marcel: 96.57
Zhang et al. [47]	2023	HGR1, OUHANDS	25 10	899 3,000	BaseNet, MSS_LAS	LHGR-Net	93.36, 98.57
Hao et al. [54]	2023	FMCW-SAR Imaging	5	-	HOG-PCA	Random Forest	Unob:97 Ob:93
Byberi et al. [55]	2023	Custom ASL	10	1400	Inductive Sensing	Random Forest	CV: 99.76
Musa et. al [2]	2024	SL	10-77	1100-20000	GmTC	CNN	95.00
Aurangzeb et al. [48]	2024	MUD ASLAD	36 29	2520 87,000	Image-based features	HVCNNM	99.23, 99.00
Sharma et al. [53]	2024	Ego hand ASL Senz3D	-	-	-	ResNet18 DMD	97.85 98.49 98.51
Bose et al. [67]	2024	NUSHP-II, SENZ-3D, MITI-HD	-	-	-	Yolo-v2 (DarkNet-19), Yolo-v3 (DarkNet-53)	99.10(MI-HD) 99.18(MI-HD)
Alonazi [68]	2024	Egogesture, Jester	15, 15	2081 v, 148092	CNN, Neural gas, Thermal mapping	DBN	90.73 89.33
Montazerin et al. [69]	2023	Egogesture, Jester	15, 27	2081 RGB , 148,092	CNN-based detector	DBN	90.73 89.33
Dumaneh et al. [70]	2024	Massey ASL Alphabet, ASL	-	2520 87,000 23,400	-	CNN, Gabor filter, ORB feature descriptor	99.92 99.80 99.80

a: Preprocessing

Image segmentation and augmentation, including rotation, translation, and scaling, are vital preprocessing techniques to enhance the image data. Miah et al. utilized an RGB still-image-based HGR system for BSL, which utilizes segmentation and augmentation techniques [39]. While skin-color models are commonly used to differentiate hand motions from the background, they may struggle with objects of similar skin tones, such as faces. Segmentation is further improved by training classifiers to distinguish hand regions from non-hand areas using attributes from extensive datasets. CNN-based segmentation methods, particularly Fully Convolutional Networks (FCNs), are increasingly popular. FCNs optimize motion segmentation by employing deconvolution layers and upsampling images to their original size through pixel prediction, handling images of any size without requiring uniform dimensions. Despite challenges like occlusion and varying light conditions, CNN-based segmentation effectively addresses these issues, enabling robust gesture segmentation.

b: Hand Crafted Feature and ML Approach

Researchers have extensively employed hand-crafted feature extraction coupled with ML algorithms for SLR systems [38]. Various techniques, such as the HMM and Pattern Trees (SP-Tree), have been utilized by researchers to develop static HGR recognition. For instance, Ren et al. achieved 93.00% accuracy for Greek Sign Language (GSL) using the Hidden Markov Models (HMM) approach, and Ong et al. reported 88.00% accuracy for German Sign Language (GSL) by using SP-Tree method [56]. Additionally, Linear Discriminant Analysis (LDA), k-nearest Neighbors (KNN), and Random Decision Forest (RDF) have shown efficiency across various SL datasets. Moreover, several prominent techniques have been utilized for this purpose, including histogram of oriented gradient (HOG), CNN, and PCA. An interesting approach described in [44] creates a unique feature vector by combining the Hu invariant moment with structural shape descriptors to extract features from image data.

Takayama and Takahashi [34] utilized HMM for word classification, effectively managing time and amplitude variances in time series signals. They designed left-to-right HMM models with 5 states for pauses and 20 states for each sign word. Athira et al. [35] employed an SVM for SLR, developing three models for Zernike moments, trajectory-based recognition, and shape-based recognition. The SVM used a multi-class C-SVC with a radial basis function kernel and a one-against-all strategy. Oliveira et al. [43] used Principal Component Analysis (PCA) for Irish Sign Language (ISL) recognition. Ibrahim et al. [57] employed the Euclidean distance classifier for Arabic Sign Language (ArSL), achieving a 97.00% recognition rate. Dixit and Jalal [44] used a multi-class Support Vector Machine (MSVM) for ISL recognition, reporting a 96.00% accuracy. Ito et al. [45] used SVMs for JSL classification, achieving mean accuracy rates of 99.2%, 94.3%, and 86.2% for different numbers of JSL words. Islam et al. [58] reported a training model accuracy of approximately 95.00% for recognizing BSL digits.

c: CNN-Based Methods

Researchers focus on DL models for effective, generalized HGR with large-scale datasets that face difficulties in ML algorithms. Miah et al. utilized a CNN-based model, BenSignNet, after preprocessing with segmentation and augmentation, achieving impressive accuracy rates of 93.00% for the BdSL38 dataset and 99.00% for the ASL dataset [39]. Similarly, DL models have shown significant improvements in recognition accuracy for Chinese, Arabic and Japanese sign languages [45], [50], [71]. Wang et al. [59] discuss the CRNN (Convolutional Recurrent Neural Network) architecture, which combines convolutional layers for feature extraction and recurrent layers for sequence modelling. Then, the Bi-LSTM network captures semantic dependencies in both forward and backward directions and reported 98.80% for single gestures. Tao et al. [38] applied a DL algorithm for RGB-based HGR, achieving an exceptional accuracy of 99.9% with leave-one-out evaluations for the 24 indicators. The YOLOv5 model proposed by Jain et al. [29] achieved a commendable accuracy of 92.00% for Danish SLR. In addition, [30], [68] also used a CNN-based system for hand gesture recognition. To enhance feature effectiveness, Miah et al. [60] applied a multi-stage deep neural network, showing good performance accuracy for various hand gesture datasets.

d: Transformer and GCN-Based Methods

Recent advancements in vision-based HGR have seen the emergence of techniques leveraging GNNs or GCNs. The Vision Transformer (ViT) has gained prominence for SL applications [72]. However, concerns about potential information loss with ViT have led to the development of transformer models like CNN meets Transformer (CMT). Guo et al. introduced CMT, incorporating self-attention with CNN layers to extract multi-scale features efficiently [73]. Subsequent optimizations by Shin et al. further improved CMT's performance, achieving impressive accuracy rates for KSL datasets [20]. This system may encounter difficulties with BdSL, JSL, or ASL datasets and vice versa. To address these challenges, researchers have been working to develop automatic multi-cultural SLR systems using ML and DL approaches [74]. More recently, Miah et al. employed a Graph meet with CNN and Transformer model (GmTC) module to enhance multi-culture SLR, and they achieved good performance accuracy [2].

3) Challenges and Future Direction

Current challenges in RGB still image-based HGR involve limited model effectiveness in handling orientation variations, partial occlusion, and the inability to capture depth and spatial information accurately. The lack of diverse datasets and the need for computationally efficient models are additional hurdles. Future research directions can address these challenges by exploring multiview augmentation techniques, integrating 3D approaches, incorporating supplementary features like temporal information and DL, expanding datasets, and developing signer-independent systems for SLR tasks. Improved image-collecting techniques are also crucial for better motion capture. Benchmarking against state-of-the-art methods, exploring 3D approaches, and incorporating additional features or techniques for various SLRs are also important.

B. RGB Video Modality Based Dynamic Gesture Recognition

Figure 5 demonstrates the overview of HGR recognition architecture, including video modality. In dynamic HGR, signers articulate signs in a sequence of frames which need to express the full meaning of a single hand gesture [75]. The main challenge of dynamic HGR is to extract temporal dependencies and relationships among the consecutive frames. An illustration of this can be seen in the continuous gestures used to convey the sentence IT IS CLOSED TODAY in ISL as highlighted in [7], [76].

1) Dataset

Dynamic HGR is crucial to express the meaning of the sign and hand gesture, as in the RGB still image. Table 4 summarizes these datasets, listing details such as name, release year, number of gesture classes, subjects involved, total samples, samples per class, modality (RGB or RGB+D), and primary performance metrics, typically accuracy. Key datasets include Montalbano (v2), 20BN-jester, ChaLearn LAP IsoGD, DVS Gesture, and SKIG, spanning from 2013 to 2022 and covering a wide range of classes and sample sizes. These datasets are essential for advancing gesture recognition research and providing benchmarks for model development and performance evaluation. Notable Ego-Gesture [77], Jester with 148,092 samples [78], LSFB-CONT from Belgium with 85,000 samples [79], and LSA64 capturing 3200 samples [80], Kurdisha Sign Language [37]. The JSL video dataset includes 92 JSL words signed by ten native signers. The videos, recorded in office environments using smartphone cameras, consist of 3,900 usable videos for 78 words [34]. The ISL dataset comprises approximately 900 static images and 700 videos of alphabets and dynamic words collected from seven participants using an external webcam [35]. Moreover, German Sign Langauge (GSL) dataset RKS-PERSIANSIGN [81], includes three combined datasets, while RWTH-PHOENIX-Weather-2014 [82] consists of RGB videos with approximately 1 million frames covering 6,841 sentences in German Sign Language. Annoted Danish Sign Language (DDSL) [29], Arabic Sign Language (ArSL) videos, consisting of 450 videos with daily school life evaluations [57]. Additionally, a large-scale dataset of 25,000 annotated videos was obtained from public ASL videos on video-sharing platforms, captured by ASL students and teachers [83].

TABLE 4 RGB Video Modality-Based Public Dataset That is Commonly Used for Dynamic HGR

Dataset Names Dataset Name	Year	Lang.	Classes	Subject	Total Sample	Samples/C	Input Device	Latest Performance Accuracy
SKIG [84]	2013	Hand Gesture	10	6	2160	-	RGB/D	98.60 [85]
Montalbano (v2) [86]	2015	Hand Gesture	20	27	14000	-	RGB,D	81.62 [87]
RWTH-PHOENIX-Weather [88]	2015	Germany	1200	-	45760	-	RGB	36.7/wer [89]
ChaLearn LAP IsoGD [90]	2016	Hand Gesture	249	21	47933	-	RGB,D	57.40 [91]
DVS Gesture [92]	2017	Hand Gesture	11	29	1342	-	RGB	-
NVGesture [93]	2016	Hand Gesture	25	-	1532	-	RGB	88.22 [65]
IsoGD	2017	Various	13	-	47933	-	RGB	67.14
PHOENIX14 [94]	2018	Germany	1081	9	6841	-	RGB	-
PHOENIX14-T [88]	2015	Germany	1085	9	8227	9	RGB	36.7/wer
20BN-jester [95]	2019	Hand Gesture	20	27	13858	-	RGB,D	-
Jester [78]	2018	Various	27	-	148092	-	RGB	90.73 [69]
IPN [96]	2021	Various	13	-	4000	-	RGB	-
CSL-Daily [97]	2021	Chinese	2000	10	20654	-	RGB	67.00 [89]
SIGNUM [98]	2008	German	1230	25	15075	-	RGB	-
BOBSL [99]	2021	British	395	85	47551	-	RGB	50.5 [99]
LSFB [79]	2021	French, Belgian	6883	100	85132	-	RGB	51.5 [79]
KSL-77 [21]	2022	Hand Gesture	77	25	112564	1461	RGB	-
EgoGesture [77]	2022	Hand Gesture	83	50	2081	-	RGB	91.64 [100]
MSRGesture [77]	2022	Hand Gesture	9	-	-	-	RGB	99.41 [100]
LSA-T [101]	2022	Argentinian	103	-	14880	-	RGB	-
27 ASL [102]	2022	American	27	173	-	-	RGB	-
LSE-Sign [103]	2016	Spanish	5100	-	-	-	RGB	-
ASLG-PC12 [104]	2020	American	-	-	87709	-	RGB	-
ASL-LEX 2.0 [105]	2021	American	-	-	87709	-	RGB	-
LSA64 [80]	2023	Argentinian	64	-	3200	-	RGB	-

2) Methodology of the RGB Video Modality for Dynamic HGR

Table 5 demonstrates the summary of the existing Video modality HGR system including year, dataset information, feature extraction and classification method performance. Dynamic gestures involve both temporal and spatial information, requiring each dynamic HGR system to extract these features to enhance performance, accuracy, and efficiency. Below, we describe the methodology for dynamic HGR, encompassing both ML and DL approaches. In DL, approaches can be categorized based on feature types, including single-stream CNN, two-stream CNN (which includes spatial and temporal streams), 3D CNN, LSTM, and transfer learning. These methods leverage various network architectures to capture and process the intricate details in dynamic gestures.

TABLE 5 Methodological Summary of the Dynamic HGR Using Video Modality

Author	Year	Dataset Name	No Class	No Sample	Feature Model	Classifier	Accuracy/LL(%)
Devi et al. [91]	2015	ChaLearn LAP IsoGD	249	21	C3D	SoftMax	57.40
Szegedy et al. [106]	2016	Montalbano	20	27	Two-stream+RNN	SoftMax	91.70
Cheng et al. [107]	2016	ChaLearn LAP IsoGD	249	21	C3D+LSTM	Softmax	68.14
Oliveira et al. [43]	2017	Iris SL	23	58,114	PCA	PCA	95.00
Ibrahimet et al. [57]	2018	Arabic SL	222	450	Geometric features	Euclidean distance	97.00
Sun et al. [85]	2018	SKIG	10	6	C3D+LSTM	SoftMax	98.60
Zhao et al. X [87]	2018	Montalbano	20	27	DNN+DCNN	SoftMax	81.62
Pigou et al. [108]	2018	Montalbano	20	27	RNN	SoftMax	67.71
Islam et al. [36]	2018	American SL	26	N/A	DCNN	MCSVM	94.57
Mahmood et al. [37]	2018	Kurdish SL	10	N/A	N/A	ANN	98.00
Takayama et al. [34]	2018	Japanese SL	92	N/A	Z-Score, PCA	HMM	93.35
Pu et al. [82]	2018	German SL	9	N/A	3D-ResNet	3D-ResNet	-
Pu et al. [89]	2019	PHOENIX-Weather CSL	-	N/A	3D-ResNet+BiLSTM	Attention Soft DTW	36.7/wer 67.00
Guo et al. [109]	2019	SKIG	10	6	ResC3D+Attention	SoftMax	90.60
Jiang et al [110]	2019	ChaLearn LAP IsoGD	249	21	ResC3D	SoftMax	50.93
Gao et al. [4]	2020	ASL	-	-	2S-CNN	SoftMax	92.00
Sharma et al. [111]	2020	ChaLearn LAP IsoGD	249	21	C3D+Pyramid	SoftMax	49.20
Gao et al. [112]	2020	SKIG	10	6	RSDCNN+RNN	SoftMax	100.0
Rastgoo et al. [81]	2020	Persian SL	100	N/A	3DCNN, ResNet50	LSTM	99.80
Cheng et al. [113]	2020	Arabic	120	-	CNN	Bi-LSTM	97.3
Li et al. [114]	2020	PHOENIX14-T (RPWT)	120	3000	TSPNet	Encoder/decoder	BLEU-13.41 ROUGE 34.9
Camgoz et al. [115]	2020	PHOENIX14 PHOENIX14-T	1081 6841	1085 8227	CNN	CTC	-
Niu et al. [116]	2020	PHOENIX14 PHOENIX14-T	1081 6841	1085 8227	CNN	CTC	26.8(WER)
Yin et al. [104]	2020	PHOENIX14-T ASLG-PC12	6841	8227 87709	STMC-Transformer	Bi-LSTM, CTC	96.60
Hoao et al. [117]	2021	PHOENIX14 PHOENIX14-T	1081 6841	1085 8227	3D CNN	Bi-LSTM+CTC	22.00,22.40
Zhou et al. [97]	2021	Kinetics-400 HSL	8	-	HOG	(3+2+1)D CNN	94.60
Hu et al. [118]	2023	PHOENIX14 PHOENIX14-T CSL-Daily CSL	1081 6841	1085 8227	self emphasizing network (SEN)	CNN	21.0 (WER) 20.7 (WER) 0.8 (WER) 0.8 (WER)
Athira et al. [35]	2022	Indian SL	N/A	N/A	HOG	SVM	91
Jain et al. [29]	2023	Danish SL	36	N/A	Deep CNN	YOLOv5	92
Shanableh et al. [119]	2023	Arabic dataset	120	-	CNN	Bi-LSTM	97.3
Guo et al. [120]	2023	PHOENIX14-T CSL	100	8227 25000	CNN + LSTM + HMM	Transformer	21.10 (WEN) 98.25
DU et al. [121]	2023	WLASL NMFs-CSL	2000 1067	21083 32,010	Transformer-spatial	temporal-softmax	57.13 72.4
Zhou et al. [122]	2023	PHOENIX14-T CSL HKSL	8227 100 50	8227 25000 2400	(3+2+1)D ResNet + BiLSTM + BERT	CTC	20.2 23.30 12.45 WEN
Karsh et al. [123]	2024	MUGD ISL, ArSL NUS-I, NUS-II	-	-	inception V3	mIV3Net	97.14, 99.3 , 97.4 99, 99.8
Karsh et al. [100]	2024	EgoGesture MSR Gesture	-	-	Xception	CNN	91.64 99.41
Hax et al. [124]	2024	Depth_Camera_Dataset	6	662	Inception-v3 (CNN)	LSTM (RNN)	83.66
Zhou et al. [125]	2024	OUHANDS HGR1 EgoHands NUS-II	10 25 4 10	-	FGDSNet	Softmax	89.57 96.89 97.69 99.80
Zerrouki et al. [126]	2024	Interactive Museum Maramotti	7	700	Image-based feature extraction	Bi-LSTM	99.86
Farid et al. [127]	2024	SKIG DCOG	10	1080 -	-	SSD-CNN with deep dilated masks	90.61 88.56

a: ML Based Feature Extraction Classification Approaches

Many researchers have been working to develop dynamic HGR systems using various methodologies, such as the Dynamic Time Warping (DTW) algorithm and the HMM [128], [129]. Dynamic HGR involves analyzing temporal and spatial information, presenting unique challenges. Two classical methods for recognizing dynamic gestures are the Dynamic Time Warping (DTW) algorithm and the HMM. The DTW approach

calculates the similarity between two different-length time series signals. Dynamic HGR can measure the similarity between two videos of hand gestures. The concept is two video modalities need to be aligned before calculating the similarities. Corradini [130] employed DTW for dynamic HGR by extracting features by normalizing video into a sequence template. The resulting sequences were matched with templates in the training set, and the gesture recognition result corresponded to the category with the smallest difference from the template. However, the DTW approach does not utilize statistical formulas during training and faces challenges with large data volumes and complex gestures. HMM used forward and backward algorithms to train each gesture category to overcome these challenges. During testing, all samples traverse HMM models using the forward algorithm to compute the probability of each trial.

Saha et al. developed a dynamic HGR system using HMM, achieving 90% accuracy with 12 dynamic gesture datasets [131]. Similarly, Yang et al. employed HMM for dynamic HGR by converting images to HSV color models, segmenting with threshold values, and using a state-based positioning algorithm to determine gesture start and end positions in continuous video [132]. They extracted hand-crafted features such as size, speed, and shape, as well as trained and tested their model with HMM. Later, Kevin P. Murphy expanded Yang's algorithm using the HMM toolset and an additional library for dynamic HGR. Takayama and Takahashi [34] employed HMM for JSL word classification, significantly reducing annotation workload and achieving 38.01% accuracy. Skin colour and hand region were extracted with preprocessing, and then edge changes, wrist angle, finger angle, cross-section area, movement centre, number of fingers, and radius area were extracted as features.

b: Single Stream and Two Stream CNN

DL-based CNN has proven its excellence in various fields, including computer vision and HGR, achieving significant improvements. Researchers have explored diverse applications using various data. For instance, Xu applied a CNN to recognize gestures using a complex dataset recorded with a monocular camera, reporting good performance accuracy [133].

Islam et al. [36] use Deep Convolutional Neural Networks (DCNNs) for efficient feature extraction in ASL recognition. Athira et al. [35] employed a Histogram of Orientation Gradient (HOG) and Histogram of Edge Frequency (HOEF) with CNNs for static and dynamic gesture recognition in ISL, achieving 91% and 89% accuracy for finger spelling alphabets and single-handed dynamic words, respectively. Traditional 2D CNNs struggle with video data due to their inability to capture temporal features. To address this, researchers use two-stream CNN networks, integrating spatial and temporal feature extraction streams, often constructed with 2D CNNs like VGG, AlexNet, InceptionNet, and ResNet.

The spatial-temporal-based two-stream network, first introduced by Simonyan and Zisserman in 2014, demonstrated impressive results in video classification tasks using Inception-based DL [134]. This network uses five sequences of frames for the spatial stream input and the remaining five frames for the temporal stream. The final features are produced by fusing these streams to determine the gesture category. However, this method struggles to extract long-range information, which is crucial for dynamic HGR. To address this, Wang et al. applied a temporal segment network (TSN) as a sparse sampling-based method to capture long-range dependencies [135]. The TSN method, constructed with Inception v2, divides the video into segments, produces probability vectors for each, and averages these vectors for the final long-term video prediction. To further improve performance accuracy, Feichtenhofer et al. fused spatial and temporal information in TSN, aiming for higher layer fusion to generate category score fusion [136]. Zhu et al. applied a CNN-based motion net to enhance spatial-temporal features, achieving good performance accuracy with a dynamic hand gesture dataset [137].

Kopuklu et al. applied a two-stream approach using frames and a single RGB image in the TSN's temporal network [78]. Wang et al. developed a two-stream RNN for skeleton-based action recognition, effectively modeling temporal dynamics and spatial configurations with skeletal data [138]. Gao et al. proposed a two-stream-based ASL recognition system using CNN, named 2S-CNN, achieving 92% accuracy with an ASL dataset [4]. While 2D CNNs are effective for extracting spatial contextual features, they often overlook temporal contextual information, which is crucial for dynamic HGR involving both images and videos. To address this, it is essential to switch to 3D CNNs, which can capture timing information from frame sequences more effectively.

c: Recurrent Neural Networks(RNN) Based Method

GPT For dynamic gesture recognition, capturing the chronological sequence is crucial, and traditional networks struggle with this task. Recurrent Neural Networks (RNNs) address this challenge by processing sequential data and generating effective features. They link lower and upper frames to represent the time dimension, with LSTM being a type of RNN that enhances this capability [108].

Pigou et al. [108] applied an end-to-end neural network incorporating bidirectional recurrence and temporal convolution to enhance performance. More recently, Molchanov et al. proposed the R3DCNN method for gesture recognition [93]. This method combines a 3D CNN for short-term video sequences with an RNN for long-term video sequences, effectively capturing both spatial and temporal features. The main drawback of RNN models is their susceptibility to the vanishing gradient problem. This issue occurs because RNNs use backpropagation through time, which can cause gradients to become very small, making it difficult for the network to learn long-term dependencies effectively. Consequently, RNNs often struggle to maintain the history effect over extended sequences, leading to challenges in accurately recognizing dynamic gestures that depend on long-term contextual information. To address this problem, Zhu et al. developed an LSTM-based DL system [137]. Additionally, they utilized 2D CNNs to further enhance the effectiveness of the feature extraction process.

Mahmood et al. [37] employed an Artificial Neural Network (ANN) for real-time dynamic HGR in Kurdish Sign Language, achieving a performance accuracy of 98%. Jain [29] utilized sequence modelling techniques with CNNs for Danish SLR, attaining high accuracy with a YOLOv5-based model. Rastgoo et al. [81] demonstrated over 90% recognition accuracy in hand SLR using LSTM on the RKS-PERSIANSIGN dataset.

d: 3D CNN Based Dynamic HGR

The 3D CNN consists of 3D convolutional, activation, and pooling layers, extracting both spatial and temporal information from frame sequences or videos. While the working procedure is similar to 2D CNN, the key difference is that 2D layers operate on a single feature map with height and width dimensions, whereas 3D layers include height, width, and time dimensions. This temporal information helps 3D CNNs capture temporal dynamics in addition to spatial data.

Many researchers have developed 3D CNN-based systems, often referred to as the C3D model [139]. This model consists of eight convolutional layers with ReLU activation, five pooling layers, a softmax layer, and two fully connected layers with a dropout rate of 0.3 to prevent overfitting.

The C3D model [139] is recognized as an efficient spatiotemporal feature extractor and is widely employed in dynamic gesture recognition algorithms. Liu and Shao used C3D [84], while Zhu et al. enhanced its capabilities with pyramid input and fusion strategies [137]. Zhang et al. proposed using C3D and bidirectional convolutional long short-term memory networks (BLCTMNs) to learn 2D temporal feature maps for HGR [31]. However, the simple architecture of C3D limits its feature expression. To address this, Tran et al. introduced residual connections, resulting in the ResC3D model [140], which allows for deeper networks without performance degradation. Li et al. further improved dynamic gesture recognition with ResC3D by incorporating an attention mechanism to focus on relevant video frames and motion regions [141].

Rastgoo et al. [81] utilized a 3D Convolutional Neural Network (3DCNN), specifically ResNet50, for pixel-level feature extraction in hand sign recognition from RGB videos. Pu et al. [82] achieved effective continuous SLR on the RWTH-PHOENIX-Weather dataset, with an inference time of around 1 second for a 140-frame sign video. They employed a stacked dilated convolutional network and the Connectionist Temporal Classification (CTC). Their study evaluated several state-of-the-art approaches as baselines for SLR, including 2D-CNN-LSTM, body key-point, CNN-LSTM-HMM, and 3D-CNN [83]. Joze et al. [83] used I3D and achieved 95.16% and Jain [29] trained with a CSP-DarkNet53 backbone and YOLOv3 outperformed existing methods in Danish SLR with a YOLOv5-based model and reported 92% accuracy 9.02ms average time per image.

3) Current Challenges and Future Direction of the RGB Video Modality

In video-based HGR, several significant challenges persist: **Temporal Dynamics:** Accurately interpreting the temporal dynamics of gestures, especially in longer sequences, remains difficult.

Class Variability: Substantial variability within and between gesture classes complicates the development of robust recognition models.

Real-time Processing: Ensuring efficient real-time processing while maintaining high accuracy is a technical hurdle.

Generalization: Developing models that generalize well across different environments and users is challenging.

To address these challenges, the integration of attention mechanisms and transformer networks can be explored to capture intricate temporal dynamics and long-range dependencies better. Strategies to transfer knowledge from controlled settings to diverse real-world scenarios will be crucial. Combining information

from multiple sources (e.g., RGB, depth, and motion sensors) could enhance recognition accuracy and robustness. Advancements in edge computing will enable efficient real-time processing, while privacy-preserving techniques will ensure user data integrity. Incorporating human pose estimation algorithms to improve gesture recognition accuracy and efficiency.

SECTION III.
3D Skeleton Modality

Skeleton-based HGR addresses challenges image-based SLR systems face, such as background clutter, hand occlusion, lighting variations, and enhancing hand movement representation. Researchers have enhanced system accuracy and efficiency by extracting joint skeleton key points from RGB videos while tackling computational complexity. Advancements in 3D camera technology and tools like Mediapipe, Alphapose, and OpenPose have facilitated precise skeleton point collection, creating a more robust gesture recognition framework. There are many researchers who have been working to develop skeleton-based HGR systems using handcrafted features with ML or end-to-end DL.

A. Dataset

There are few skeleton-based datasets available online. Also, we easily extracted skeleton-based datasets from the RGB video. Table 6 included various benchmark datasets' names of this skeleton-based modality, creation year, number of classes, dataset types, sample size and latest performance accuracy. ASLLVD [142] is the most used skeleton dataset consisting of 2745 ASL signs with 9748 videos, averaging 3/4 videos per sign and 67 joints per frame; the WLASL dataset comprises 2000 ASL signs from 119 subjects across 21089 videos, averaging 10.5 videos per sign with 67 joints per frame [143]. MSL [144], DHG and SHREC dataset covers 14/28 general signs with totalling 2800 videos, with 22 joints per frame [145]. MSRA dataset consists of 17 MSRA signs with 76500 videos (500 videos per sign) and 21 joints per frame. The PSL dataset involves 19 PSL signs with 2700 images (55 images per sign), 67 joints per frame, and so on.

TABLE 6 Skeleton Modality-Based Dataset Description

Dataset Names	Year	Language	Signs	Sub.	Total videos	Videos Per Sign	Joint Per Frame	Latest Performance Accuracy
ASLLVD [142]	2019	ASL	2745	n/a	9748	3/4	67	-
DHG [145]	2017	General	14/28	20	2800	-	22	-
SHREC [145]	2017	General	14/28	27	2800	-	22	-
MSRA [23]	2019	General	17	n/a	76500	500	21	-
WLASL [143]	2020	ASL	2000	119	21089	10.5	67	-
MSL [144]	2022	MSL	30	20	3000	20	67	-
PSL	2020	pakistani	19	n/a	2700 (img)	55 (img)	67	-
AUTSL [146]	2020	Turkey	226	43	38336	-	67	98.00 [147]
Briareo	2020	HGR	12	40	1440	-	-	96.64 [148]
FPHA	2020	HGR	45	6	1175	-	-	91.16 [148]

B. Methodology of the 3D Skeleton Modality

To overcome the redundant background, computational complexity, and partial occlusion issues, joint skeleton-based data modality become popular in computer vision, but it remains a challenging task. Table 7 demonstrates the summary of the existing Video modality HGR system, including year, dataset information, feature extraction and classification method performance. Many researchers employed preprocessing, feature extraction with ML and end-to-end DL, which is given below [170], [171].

TABLE 7 Databases and Performance of Skeleton Modality for HGR

Author	Year	Dataset Name	No Class	No Sample	Feature Model	Classifier	Performance (%)
Smedt et al. [149]	2016	American SL	14	2800	Fisher Vectors	SVM	88.00
Smedt et al. [149]	2016	DHGD	28	2800	ASJT	Softmax	80.11
De et al. [150]	2016	DHGD	28	2800	SoCJ + HoHD + HoWR	Softmax	80.00
Chen et al. [151]	2017	DHGD	28	2800	MARNN	Softmax	80.32
Boulahia [152]	2017	DHGD	28	2800	Boulahia	Softmax	80.48
Liu et al. [153]	2017	American SL	20	13,585	LSTM	LSTM	96.30
Konstantinidis et al. [154]	2018	Argentinian SL	64	N/A	VGG-19 Network	CNN, RNN, LSTM	98.09
Devineau et al. [155]	2018	N/A	14	N/A	DL	DL	91.28
Nunez et al. [156]	2018	American SL	60	14,000	CNN, LSTM	CNN, LSTM	99.00
Ma et al. [157]	2018	DHGD	28	2800	GREN	-	82.03
Ma et al. [157]	2018	DHGD	28	2800	NIUKF-LSTM	Softmax	80.44
CNN+LSTM [156]	2018	DHGD	28	2800	CNN+LSTM	Softmax	74.19
Yan et al [158]	2018	DHGD	28	2800	STA-GCN	Softmax	87.10
Hou et al. [159]	2018	DHGD	28	2800	Res- TCN	Softmax	83.60
Hou et al. [159]	2018	DHGD	28	2800	STA Res- TCN	Softmax	85.00
Si et al. [160]	2019	DHGD	28	2800	CNN+RNN	Softmax	74.19
Res-C3D [160]	2019	Shrec	28	2800	CNN+LSTM	Softmax	89.52
Chen et al. [161]	2019	DHGD	28	2800	MFA-Net	Softmax	81.04
Chen et al. [162]	2019	DHGD	28	2800	DG-STA	Softmax	88.00
Musa et al. [163]	2003	AUSTL	-	-	Multi-Stream GCN	CNN	99.00
Rastgooet al. [15]	2021	Persian SL	100	N/A	3DCNN	2DCNN, 3DCNN, LSTM	99.80
Jiang et al. [147]	2021	-	N/A	N/A	SL-GCN, SSTCN, 3DCNN	GEM	99.81
Han et al. [1]	2021	Chinese SL	22	N/A	ResNet and LSTM	ResNet, LSTM	88.6
Jiang et al. [164]	2021	Turkish SL	226	N/A	SSTCN	CNN, LSTM	98.53
Musa et al [23]	2023	DHGD	28	2800	DG-STA	Softmax	88.00
Shin et al [21]	2023	KSL	77	20000	Dynami GCNN	Softmax	99.00
Himrichs et al. [165]	2023	PHOENIX14 PHOENIX14-T	1081 6841	1085 8227	Transformer	Softmax	18.55, 18.59 (WER)
Slama et al. [148]	2023	SHREC'17 Track Briareo FPHA	14 , 28 12, 45	2800 1440 1175	STr-GCN	Softmax	93.39 96.64 91.16
Peng et al. [166]	2023	SHREC'17 FPHA	14,28 45	2800 1175	EffGCN	ResGCNeXt	95.36 93.45
Musa et al. [22]	2024	WLASL	2000	-	Dynamic SepTCN GCN	Softmax	90.00
Khanna et al. [167]	2024	15 gestures	15		Geometric	multi-view CNN	91.00 (WER)
Mahmud et al. [168]	2024	DHG-14/28, SHREC-2017	14 28		Multimodal fusion CRNN	Softmax	90.82:89.21 93.81:90.24
Jinfu Liu et al. [169]	2024	SHREC'17, DHG-14/28, NTU RGB+D NW-UCLA	14, 28 60, 10	2800, 2800 56880 1494	spatiotemporal	TD-GCN	97.02:95.36 93.9:91.4 96.8 97.4

1) Preprocessing and Pose Estimation

In 2D pose estimation, deformable part models are common but often lack detail and context [172]. For 3D pose estimation, the goal is to create a 3D pose that matches the person's position in the image. Although deep neural networks have improved 3D pose estimation, it remains difficult due to the larger pose space and more ambiguities [173], [174]. Researchers have used media pipe, oppose, alpha pose, and impose to extract the joint key points, which constructed 2D key points or 3D key points based on the settings.

2) Hand Crafted Feature and ML Approach

After extracting skeleton points from RGB-based data modalities, researchers often use joint skeleton-based feature extraction techniques. Shin et al. employed a geometrical approach to extract distance and angular features from 21 hand key points using the MediaPipe system for an ASL dataset [175]. Ohn-Bar and Trivedi proposed a feature generator using the HOG algorithm with a linear SVM for classification [170]. Other methods include using a covariance matrix for joint locations [171] and joint distance and angles to capture intraclass variance [176]. Smedt et al. introduced a hand geometric configuration method for spatial-temporal motion feature extraction, achieving high accuracy on the DHG dataset using SVM [149]. Smedt et al. extracted features based on Fisher vectors and skeleton-based geometric techniques, then applied SVM to the concatenated features, achieving 83.00% accuracy for 14 gestures and 80.00% for 28 gestures in the DHG dataset [150]. Figure 6 (a) illustrates the SoCJ features. They also applied Fisher vectors and connected features for the SHREC'17 dataset with an SVM classifier, achieving 88.24% accuracy for 14 gestures and 81.90% for 28 gestures [145]. Their work highlighted the superiority of 3D skeleton information over depth-based approaches but did not account for gesture amplitude, potentially losing temporal information. Chen et al. used a feature extractor combining articulated finger movements and global hand movement features to extract bone angles, using RNN for classification. Their model achieved 84.68% accuracy for 14 classes and 80.32% for 28 classes on the DHG dataset [151]. Shao et al. proposed a method that integrates shape and motion information using feature descriptors like Motion History Images (MHI) and Predicted Gradients (PCOG) [177]. De Smedt et al. introduced a methodology focused on extracting hand kinematic descriptors from gesture sequences, encoding them via a Fisher kernel and a multi-level temporal pyramid, and using a linear SVM classifier for recognition [178]. Figure 6 (b) demonstrates the feature calculation procedure. Additionally, Rastgoo et al. [81] introduce heatmap images derived from detected key points, offering a new feature representation alongside pixel-level and multi-view hand skeleton features. Table 7 shows various existing systems' methods and performance accuracy.

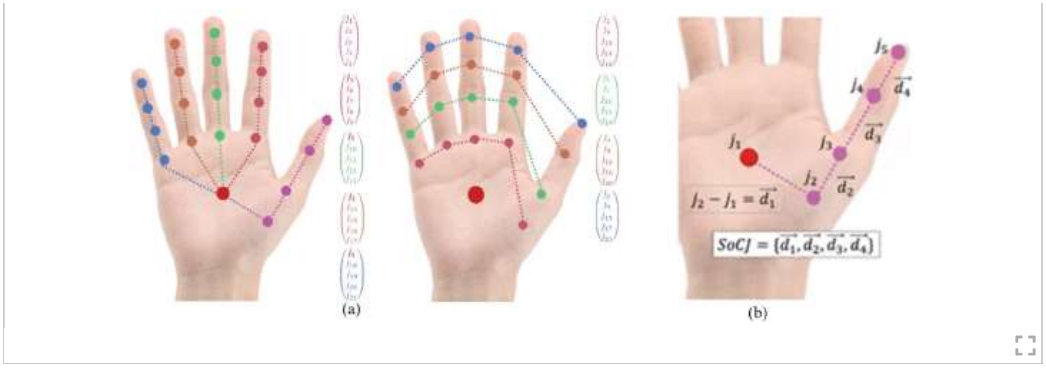


FIGURE 6.
SoCJ handcrafted 9 feature descriptors [145].

3) CNN-Based Methods
Handcrafted feature-based systems face limitations in efficiency and generalization. To address these, researchers have adopted end-to-end DL algorithms for hand gesture classification using raw skeleton data [141], [159]. For instance, Konstantinidis et al. proposed a skeleton-based SLR system using DL, achieving a 2.27% increase in accuracy on the LSA64 dataset when using body features over hand features [46]. Nunez et al. [156] showed extensive experimental study on publicly available data benchmarks, including the MSR Action3D dataset, MSRDailyActivity3D dataset, UTKinect-Action3D dataset, NTU RGBD dataset, and Montalbano V2 dataset. Zhang used Kinect-captured skeletal data for improved gesture recognition [179]. Huynh-The et al. demonstrated efficient action appearance management by pairing skeleton data with CNNs [180]. Devineau et al. showed that DL algorithms significantly improved F1 scores for DHG HGR [155].

4) RNN-LSTM Based Methods
De Smedt et al. [178] leverage the inherent structure of hand topology for skeleton-based HGR. Methods using RNNs, especially LSTM units, combined with CNNs are notably more effective than traditional ML and standalone CNN models. Konstantinidis et al. [46] demonstrates the enhanced precision of SLR by integrating HMM with CNNs and bidirectional RNNs incorporating LSTM units.

Lai et al. integrated a CNN with an RNN DL model for recognizing skeleton-based hand gestures, achieving 85.61% accuracy on the DHG 14 gesture dataset [181]. Ma et al. combined LSTM with an unscented Kalman filter (UKF) [157], while Han et al. proposed a two-stream method combining RGB and skeleton data using KLSTM-3D ResNet to improve recognition rates [1]. Han et al. also highlighted the enhanced spatiotemporal feature extraction by the fusion of ResNet and LSTM networks. Nunez et al. integrated CNN and LSTM features for extracting temporal features from 3D pose estimation, reporting 99.00% accuracy, underscoring the superiority of RNNs, especially LSTM [156].

Si et al. combined a skeleton LSTM and a Res-C3D network for recognizing hand gestures and compared their result with the existing system [160]. To improve the performance, Chen et al. extracted articulated finger movement and hand movement features, then concatenated the features and fed them into the RNN and reported 84.68% accuracy for 14 classes and 80.32% for 28 classes on the DHG dataset [151]. Rastgoo et al. explored DL-based approaches using RNN-LSTM and graph neural networks (GNN or GCN) to enhance gesture recognition performance, offering valuable insights into HGR [81].

5) Attention and GCN-Based Methods
Recent advancements in the various research domains by using self-attention mechanisms and transfer learning to enhance performance [182]. Self-attention networks establish semantic relationships among features [72], and spatial-temporal attention has been integrated with various architectures like CNNs and RNNs [159], [160]. Transfer learning techniques have also been explored to leverage pre-trained models for gesture recognition tasks. Li et al. recently proposed a temporal GCN (TDCN) to develop a skeleton-based dynamic HGR [169]. They mainly combined the different approaches of the temporal contextual enhancement module, namely temporal decoupling learning (TDL), Channel decoupling learning (CDL), etc. The reported SHREC'17 Track: 97.02% (14 classes), 95.36% (28 classes) DHG-14/28: 93.9% (14 classes), 91.4% (28 classes), NTU RGB+D: 96.8% (cross-view), 92.8% (cross-subject), NW-UCLA: 97.4% accuracy.

Existing systems often neglect motion and temporal features, failing to capture intricate joint relationships. Yan et al. used AlphaPose for hand skeleton joint extraction and applied Spatio-temporal GCN for gesture recognition [158]. Recent research combines spatial-temporal attention with architectures like CNNs [160],

RNNs, and soft-attention, as well as memory attention networks (MANS) [183]. To enhance the HGR task, Song et al. utilized RNN and LSTM to extract the spatio-temporal featural features [184]. Hou et al. designed residual connections and a temporal convolutional neural network (STA-Res-TCN) for improving the skeleton-based HGR performance [159]. Various feature levels were extracted based on the time information by using the CNN, and they reported 89.20% and 85.00% accuracy for 14 and 28 gestures of the DHG dataset. Moreover, 93.60% and 90.70% accuracy were reported for 14 and 28 gestures of the shared dataset. More recently, GCN has been used by many researchers to enhance the HGR in terms of accuracy and efficiency [158], [162]. Yan et al. applied a temporal feature enhancement-based ST-GCN model to enhance feature effectiveness, extracting spatial and temporal features, as shown in Figure 8 [158]. More recently, Chen et al. proposed enhancing spatial and temporal features through various stages of GCN for HGR [162]. They mainly applied multi-stage spatial-temporal attention-based feature enhancement approaches.

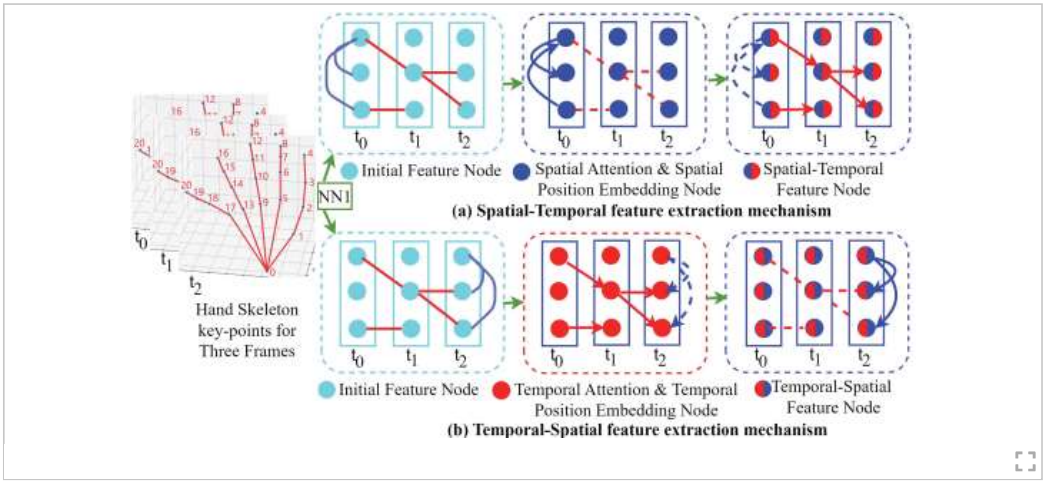


FIGURE 7.
Spatial-temporal graph based dynamic HGR [23].

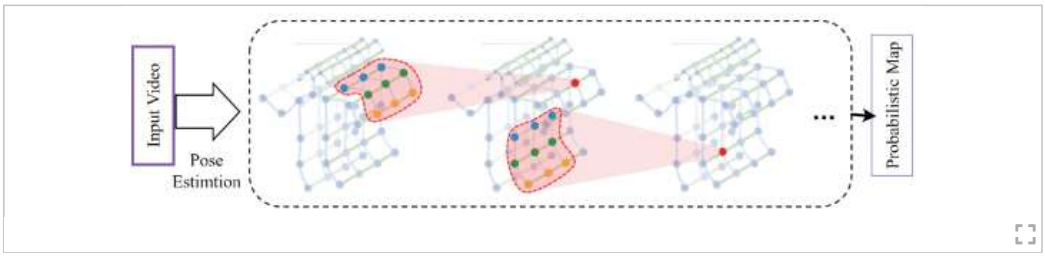


FIGURE 8.
Spatial-temporal graph convolutional neural network (STGCN) [158].

Addressing the drawback, Musa et al. employed a GCN combined with an attention-based spatial-temporal network to extract intricate hand skeleton relationships, achieving good performance on DHGD, SHREC, and MSRA datasets [23]. The working architecture of their method is illustrated in Figure 7. Similarly, Shin et al. utilized a joint skeleton and joint motion-based GCN to recognize KSL, also achieving high accuracy [21]. Ping and Tsai applied a ResGCNeXt model to recognize skeleton-based dynamic HGR [166] and reported 95.36% and 93.45% accuracy for Shrec and FPHA datasets, respectively. They mainly used various stages and parallel combinations of GCN, demonstrating that it can be considered a major advancement use of GCN.

Miah et al. extracted joint and joint motion information from video-based skeleton points using a two-stream GCN, achieving good performance accuracy [22]. To further improve performance, Miah et al. selected 27 key points out of 67 and included joint, bone, joint motion, and bone motion information from the video-based skeleton points. They employed a four-stream graph-based convolutional network, achieving enhanced accuracy.

C. Future Direction

In 3D skeleton-based HGR, several key challenges hinder accurate and robust recognition. Firstly, media pipe, oppose, alpha pose, and impose are still facing challenges in achieving exact points from the RGB video.

Many frames produce 0 value or are skipped, making major issues to efficient work and making a good system. Additionally, the limited availability of large-scale, diverse datasets tailored for 3D HGR impedes developing and evaluating robust models. This scarcity restricts the training and testing of algorithms, limiting their generalizability and performance. Furthermore, the computational complexity associated with processing and analyzing high-dimensional skeletal data is also considered a domain problem.

Future research in 3D skeleton-based HGR can refine hand pose estimation algorithms to prevent skipping frames with no key point values. In addition, strong temporal contextual feature enhancement is achieved by extracting the strong relationship among the consecutive frame joints. The enhanced feature extraction algorithms with an effective classification approach would improve accuracy and reliability for precise gesture recognition in diverse scenarios. Additionally, efforts to enhance model generalization across different environments and user populations are essential, requiring the creation of larger and more diverse datasets tailored for 3D HGR.

SECTION IV.
Depth Information Based Modality

Depth-based HGR utilizes depth data from sensors like Time-of-Flight (ToF) cameras, Microsoft Kinect, Intel RealSense, or Leap Motion Controller [145], [185], [186]. Depth data offers several advantages over RGB images, providing 3D scene representations, robustness to lighting changes, and explicit spatial information for accurate hand tracking and gesture analysis [187], [188], [189]. It captures fine details of hand movements, enhancing gesture precision [190]. As discussed below, researchers employ various ML and DL techniques for depth-based gesture recognition.

A. Dataset

Depth datasets are crucial for advancing HGR and contain depth images or maps captured by sensors such as ToF cameras, structured light sensors, depth cameras, and charge-coupled devices (CCD) [24]. Many existing video RGB or skeleton datasets also contain depth information in Tables 4 and Table 6. Notable depth datasets include MSR Gesture 3D, ChaLearn LAP, DHG, SHREC'17 Track, and REHAP. Specific examples include Kim et al.'s dataset [185] with gestures at different distances, Chang et al.'s BSL alphabets dataset [190], De Smedt et al.'s hand gestures sequences [145], Zengeler et al.'s REHAP dataset [186]. These datasets aid in training and evaluating algorithms. The depth dataset and performance of these datasets are shown in Table 8.

TABLE 8 Methodological Summary of the Depth Modality Based HGR

Author	Year	Lang.	Class	Sample	Feature Extraction	Classifier	Perfromance[%]
Kim et al. [185]	2015	American SL	4	N/A	MLBP	MLBP	95.63
De Smedt et al. [149]	2016	N/A	14	2800	FV	SVM	88.24
Zengeler et al. [186]	2018	N/A	10	600,000	ESF, PFH, VFH	LSTM	91.28
Ding et al. [24]	2022	depth-grayscale	10	N/A	VGG-16 CNN	Softmax	83.88
Chang et al. [190]	2023	British SL	450	N/A	CNN	CNN	58.00

B. Methodology of the Depth Modality

There are many researchers who utilized preprocessing, feature extraction and classification using various ML and DL algorithms, which are described below:

1) Perprocessing Approach

Depth images provide additional spatial information compared to RGB, aiding in gesture classification and recognition. The depth threshold method gauges the proximity of objects to the camera, determining the distance of each pixel. It then extracts an image within a designated range. This technique enhances gesture detection accuracy by pinpointing the depth range of the hands or considering the hand as the closest object. However, this approach imposes constraints on the manner and extent of recognition.

2) Feature Extraction and ML Approach

Many researchers have developed HGR using feature extraction and ML algorithms with depth modality. Depth maps capture spatial information about hand shapes, movements, and configurations, enabling the extraction of key features such as hand contours, finger positions, and hand skeletons [145], [185]. Kim et al.

[185] analyzed depth images to extract hand-shape features using a modified local binary pattern (MLBP). De Smedt et al. [145] utilized 3D depth information to extract hand silhouettes, employing Fisher Vector (FV) for feature extraction. Chang et al. [190] utilized Region of Interest (ROI) segmentation to enhance gesture detection accuracy. Classification methods also varied: Kim et al. [185] used MLBP, De Smedt et al. [145] employed SVM with a linear kernel. Despite successes, these ML approaches have drawbacks, relying on hand-crafted features that are time-consuming to design and may miss details in complex gestures. They often struggle with variations in hand poses, occlusions, and complex backgrounds.

3) Features and End to End DL Based Approaches

Recently, many researchers have been working to develop an in-depth modality-based HGR system. Zengeler et al. [186] explored sensor data fusion within a DL framework for HGR, employing point cloud descriptors like Ensemble of Shape Functions (ESF), Point Feature Histograms (PFH), and Viewpoint Feature Histogram (VFH) implemented in the Point Cloud Library (PCL). Chang et al. [190] used a CNN model for hand gesture categorization, achieving 58% accuracy with a two-phase algorithm involving ROI segmentation and CNN categorization. This approach eliminates the need for hand-crafted feature extraction by automatically learning relevant features from raw data. Zengeler et al. [186] also demonstrated that combining CNNs and LSTMs yields reliable results for HGR using depth data, with improved results when using a second ToF sensor. This combination effectively processes volumetric data and models temporal dynamics, capturing spatial and sequential information in hand gestures. Kim et al. [185] proposed an MLBP hand-tracking algorithm for real-time and accurate hand tracking, outperforming other methods in accuracy and tracking performance. De Smedt et al. [145] showed promising results with a skeleton-based approach, achieving accuracies of 88.24% and 81.90% for 14 and 28 gestures, respectively. Ding et al. [24] integrated CCD RGB-IR and depth-grayscale sensor data using CNNs for HGR, emphasizing the importance of multiple modalities. Li et al. [191] introduced an attentive 3D-Ghost module for dynamic HGR, showing the benefits of multimodal data. Gao et al. [192] developed a two-stream CNN framework for American SLR using RGB and depth data fusion, demonstrating the trend towards multimodal approaches for improved accuracy. Mahmud et al. [193] proposed a DL-based multimodal depth-aware system for dynamic HGR, underscoring the value of combining different input modalities for robust recognition.

C. Future Direction

HGR models need to be robust against variations in lighting conditions, cluttered backgrounds, and different viewpoints. Ensuring these models can generalize well to unseen scenarios while maintaining accurate recognition performance is crucial for practical deployment. Future research should focus on developing a generalized HGR system based on comprehensive depth datasets. Exploring attention mechanisms and transformer architectures could improve the capture of spatial and temporal relationships within the depth data, leading to more effective feature representations.

SECTION V.
EMG Modality Based HGR

Recent advancements in HGR have leveraged electromyography (EMG) signals to overcome limitations associated with RGB, skeleton, or depth datasets. The field of gesture recognition using surface Electromyography (sEMG) data has garnered significant interest across various domains such as medicine, exercise science, engineering, and prosthetic limb control [194]. Surface electromyography (sEMG) signals, which capture electrical activity generated by muscle contractions, offer a promising alternative for gesture recognition [195], [196], [197], [198], [199]. Nevertheless, the ease of implementation remains of the EMG-based HGR a critical factor in ensuring the effectiveness of these assistive technologies [69], [191], [200], [201].

A. Dataset

EMG modality datasets are demonstrated in Table 9. Accordingly, the Ninapro DB series, which consists of 9 datasets (Ninapro DB1 to Ninapro DB9), was recorded using Cybergloves technology with integrated sensors [202]. These datasets are commonly used by researchers to train and evaluate ML models for accurately recognizing hand and finger gestures. Ninapro DB1-DB9 are considered benchmark datasets in this field. In the study by Kim et al. [200], a dataset of 400 EMG data samples from 50 subjects was collected, with each sample consisting of 8 channels and spanning 1 second. The EMG data underwent min-max normalization and was reshaped into a 50×8 two-dimensional format for DL model training. Côté-Allard et al. [203] utilized two datasets: a pretraining dataset with recordings from 18 non-disabled subjects performing seven

gestures for 20 seconds each, and an evaluation dataset involving 17 healthy subjects performing the same seven gestures in three rounds, each lasting 20 seconds. The study by Lee et al. [8] involved ten healthy subjects performing ten hand/finger gestures, including seven individual finger (IF) gestures, resulting in a dataset for their research. Zhang et al. [204] created a specialized dataset for model training, consisting of 30 repetitions of 21 short-term hand gestures by 13 subjects using a Myo armband. Colli Alfaro and Trejos [205] used electromyography (EMG) and inertial measurement unit (IMU) data for user-independent gesture recognition, although specific details about the dataset size are not provided.

TABLE 9 Summary Datasets for EMG Based HGR Modality

Name	Year	Dataset Types	Sensor	No.Gesture	No. of Subjects	No. of Sample	No. Channel	Latest Performance Accuracy
DB1	2014	HGR	EMG,Kinmatics	52	27	-	10 Otto Bock	96.87
DB2	2014	HGR	EMG,Kinmatics	50	40	-	12 Delsys Trigno	92.28 [206]
DB3	2014	HGR	EMG,Kinmatics	50	11	-	12 Delsys Trigno	91.11 [206]
DB4	2017	HGR	EMG,Kinmatics	52	10	-	12 Cometa	93.00 [207]
DB5	2017	HGR	sEMG, inertial	41	10	-	16 electrodes 2 Thalmic Myo Armbands	87.00 [206], [207]
DB6	2017	HGR	EMG,Kinmatics	8	10	-	14 Delsys Trigno	80.62 [207]
DB7	2017	HGR	EMG,Inertial	41	22	-	12 Delsys Trigno	96.76 [207]
DB8	2019	HGR	sEMG, kinematic	9	12	-	16 Delsys Trigno	90.00
DB9	2020	HGR	sEMG, kinematic	40	77	-	sEMG, kinematic	89.47
DB10	2020	HGR	sEMG, Inertial	10	45	-	12 Delsys Trigno	80.00
Mayo Armband	2016	HGR	sEMG	13	21	-	8	88.92
UC2018 [25]	2018	HGR	sEMG	20	8	-	20	85.34
Arm Band	2019	HGR	sEMG	36	6	-	8	90.00
EMG-EPN -612 [28]	2022	HGR	sEMG	5	612	183600	-	94.60 [208]
EMG -5 [209]	2023	HGR	sEMG	5	10	7500	-	89.72
EMG High Density	2021	HGR	sEMG	5	41	-	256	81.74 [207]
DualMyo	2022	HGR	sEMG	8	1	880	–	99.00 [210]
EMG36	2023	HGR	sEMG	8	36	4237907	–	97.00 [210]
ISRMyo-i	2023	HGR	sEMG	4	6	-	–	85.75 [211]
CapgMyo DB-b DB-c	2017	HGR	sEMG	4	10	-	–	82.43 84.75 [211]
SEU	2022	HGR	sEMG	4	20	-	–	83.57 84.75 [211]

B. EMG Based Methodology

Utilizing Electromyography (EMG) signals for HGR involves capturing muscle activity to interpret gestures. This methodology typically includes signal acquisition, feature extraction, and classification algorithms to decode gestures. Analyzing EMG patterns enables real-time and intuitive interaction in applications like prosthetics and human-computer interfaces. Leveraging ML and DL enhances the work of many researchers working to develop EMG-based HGR by enabling complex pattern recognition. Table 10 demonstrates the existing HGR methodology and its performance.

TABLE 10 Methodological Review With the EMG Data Modality

Author	Year	Dataset Name and Types	No. of EMG Channel	Class	Sample	Feature Extraction	Classifier	Performance(%)
Côté-Allard et al. [212]	2019	American SL	8	7	52	CNN	CNN	97.81
Wei et al. [213]	2019	American SL	16	41	N/A	CNN	DL	90.00
Zhang et a. [75]	2019	American SL	8	20		RNN	RNN	89.60
Yang et al. [214]	2023	American SL	16	52	10000	MResLSTM	MResLSTM	93.52
Lee et al. [8]	2021	American SL	3	7	N/A	RMS, VAR MAV, SSC ZC, WL	ANN	94.00
Colli Alfaro et al. [205]	2022	American SL	5	7	N/A	MAV, MAVS WL, AR, ZC	LS-SVM MLP	92.90
Cruz et al. [215]	2023	EMG-IMU-EPN-100+	12	6	300	DQN	Softmax	97.45 88.05
Kim et al. [200]	2023	Myo armband	8	10	37000	CRNN	Softmax	96.57 (training) 95.10 (testing)
Kerdjidj et al. [201]	2023	Public EMG dataset	8	36	40,000	Temporal feature	k-NN	98
Xion et al. [216]	2023	NinaPro DB4, NinaPro DB5, DB1-DB3, Mendeley	Various	-	-	-	GLF-CNN	88.34 (average) 91.40 91.00 88.60
Leelakittisin et al. [211]	2023	CapgMyo ISRMyo-I SEU	-	4	-	Enhanced Lightweight CNN with JCAP	Softmax	CapgMyo DB-b: 82.43, CapgMyo DB-c: 84.75, ISRMyo-I: 85.75, SEU: 83.57
Chen et al. [208]	2023	EMG-EPN-612	6	-	-	Neural Feature Extraction	Supervised learning	94.60
Zhang et al. [47]	2023	HGR1, OUHANDS	25, 10	12, 23	899, 3,000	BaseNet, MSS, LAS	LHGR-Net	HGR1: 93.36, OUHANDS: 98.57
Xu et al. [217]	2023	DB4 DB5	53	12 16	30 tr	SE-CNN	Softmax	89.54 77.61
Zabihi et al. [218]	2023	DB2	12	17	30 tr	HDConv MHSAtten	HDCAM	XXSmall: 81.73 XSmall: 82.61 Small: 82.91
Zabihi et al. [219]	2023	DB2	12	9	30	Tnet,Fnet	TraHGR	93.84
Kim et al. [200]	2023	American SL	8	10	50	CNN	CRNN	96.04
Baroni et al. [28]	2024	HandGesture	5-612		183,600	CNN-LSTM	CRNN	90.45
Wang et al. [209]	2024	EMG Dataset	5	10	7500	BP Neural Network	Softmax	89.72 (4 channels)
Li et al. [220]	2024	NinaPro DB5 Myo	10	-	7500	multi-attention (CNN,channel spatial-temporal)	Softmax	91.64
Wang et al. [207]	2024	High-density sEMG, Ninapro DB4, Ninapro DB5	10	-	7500	RMS WL ZC SSC CDEM	Simplified Cross-Domain Error Minimization (CDEM)	81.74 93.50 , 84.00
Abdelaziz et al. [210]	2024	DualMyo EMG36	8 8	8 8	880 4237907	CNN+LSTM	Softmax	99.00, 97.00
Vasconez et al. [221]	2024	EMG-EPN-612	6	-	-	-	Supervised learning, Reinforcement learning	93.57 (supervised)

1) Preproecssing Filtering and Signal Segmentation
Filtering, Motion detection, frequency controlling, and noise reduction are the most crucial preprocessing in this data modality. Adhering to the preprocessing methodology outlined in earlier investigations [163], [222], a 1st-order low-band pass Butterworth filter was employed to dampen electrical activity data of muscles.

2) ML Based Approach with Mathematical and Statistical Features
Many researchers used statistical and mathematical feature extraction techniques. Among them, Lee et al. [8] utilize six time-domain (TD) features, including root mean square (RMS), variance (VAR), mean absolute value (MAV), slope sign change (SSC), zero crossing (ZC), and waveform length (WL). In the research presented in [205], the mean absolute value (MAV), mean absolute value slope (MAVS), waveform length (WL), 4th-order auto-regressive coefficients (AR), and zero crossings (ZC) are extracted from each EMG channel. Then these mathematical features feed into the Traditional ML techniques like Linear Discriminant Analysis (LDA) [223], [224] and SVM [223] have been utilized for HGR using sEMG signals. Lee et al. [8] experiment with four ML methods, including artificial neural network (ANN), SVM, random forest (RF), and logistic regression (LR), to build personalized classifiers for gestures.

3) CNN Based Approaches
The size of the EMG signal-based dataset has increased because the number of channels and time of the EMG signal is larger than the previous dataset, and it is difficult to get good performance accuracy with the traditional ML algorithm. To overcome the issues, many researchers have been working to apply end-end DL technology. Among them, Wei et al. [225] utilize DL techniques for gesture classification. Kim et al. [200] employ a CNN for time-domain feature extraction. To capture intricate features, the number of filters increases across neural layers, specifically to 16, 32, and 64. Recently, researchers used CNN [203], [204], [225], RNN, multi-stream residual network (MResLSTM) [204]. Côté-Allard et al. [203] employ a CNN architecture for classification, which is further improved using transfer learning techniques to enhance accuracy. Kim et al. [200] adopt a convolutional recurrent neural network (CRNN) structure for training and testing hand gesture classification.

4) RNN-LSTM Based Methods
DNNs trained on subsets of Ninapro data demonstrate promising performance on unseen repetitions [212]. However, accuracy diminishes for less explored repetitions, necessitating extensive training. Domain adaptation techniques like Transfer Learning (TL) and domain alignment aim to enhance model robustness across users and gestures [203], [212]. Zhang et al. [204] make use of an RNN model specifically designed to learn from raw surface sEMG data and predict hand gestures in real time. Zhang et al. [204] introduce a multi-stream residual network (MResLSTM) for dynamic gesture recognition, leveraging surface EMG (sEMG) signals. The MResLSTM model integrates both the residual model and convolutional short-term memory components to classify various types of gestures.

5) Attention, TCN and GNN Based Methods
As the spatial enhancement feature extracted with CNN is not so effective, researchers found that temporal modelling is a crucial feature for EMG signal-based HGR. Recently, Tsinganos et al. applied a TCN module aiming to sequence the classification problem of HGR [226]. After evaluating their model with the Ninapro DB1 dataset, they reported 89.76% top-1 accuracy. Only the temporal feature is not so effective in generalized cases, and it may face difficulties in achieving good performance accuracy due to lacking feature effectiveness. To overcome the problems, Rahimian et al. proposed a few-short learning method by integrating TCN with attention module [23]. They emphasize the challenges faced by DNNs when dealing with limited data and introduce a novel approach known as Few-Shot learning- HGR (FS-HGR) [227]. The FS-HGR model integrates temporal convolutions and attention mechanisms, allowing it to generalize effectively with minimal training instances. By leveraging Few-Shot learning, FS-HGR efficiently infers outputs based on a small number of training observations, making it a practical and promising solution for real-life applications where large datasets are not readily available. Zahihi et al. proposed a Hierarchical Depth-wise Convolution and Attention Mechanism (HDCAM) model to recognize EMG-based HGR where they reported 82.91% accuracy with DB2 dataset [218]. Zahibe et al. again proposed a TraHGR: Transformer for HGR module [219] where they reported 93.84% accuracy for the Ninapro DB2 dataset. TrAGR is mainly composed of Tnet and Fnet, aiming to extract temporal and frequency domain features and make a hybrid framework based on the transformer architecture.

Gestures recognition with sEMG data benefits from domain adaptation techniques [228], [229]. However, challenges persist in adapting to varied signal spaces and postures, and performance with limited training data remains uncertain [228]. Studies incorporating domain adaptation techniques and DL algorithms address the challenge of inter-session classification in sEMG-based HGR [228], [229]. However, their

adaptability to signal variability and performance with limited data and varied subjects remain uncertain [228]. Côté-Allard et al. [203] employ transfer learning techniques to enhance the accuracy of their DL-based approach to classification.

C. Future Direction

Recognizing hand gestures using EMG signals faces significant challenges due to variability between individuals and within the same person over time. In addition, noise in EMG data from movement or environmental interference further hampers accurate gesture recognition, necessitating effective noise reduction techniques. Future research could focus on adaptive, personalized models that continuously learn and adjust to individual EMG signal variations, employing techniques like transfer learning. This can lead to more accurate and user-friendly systems for assistive technologies, virtual reality, and human-robot collaboration.

SECTION VI.
Audio Signal Based Modality

Audio signals, representing sound waves captured via microphones or digital recordings, offer valuable cues for HGR. They provide contextual information and supplement visual data, enhancing system robustness in challenging environments. Many researchers have been working to develop audio signal-based HGR using various ML and DL tools. Below, we discuss the dataset, methodology, current challenges and future direction of these modalities.

A. Dataset

Few researchers have been working to develop an HGR system using audio signals.

Saad et al. [230] comprise reflected ultrasonic signals obtained from one transmitter and four receivers, enabling gesture detection. Siddiqui and Chan [231] collected hand gesture data from three right-handed subjects performing ASL alphabets, numbers, and relaxation gestures. Sang et al. [232] utilized a dataset with hand gesture samples from nine subjects, each performing six gestures. Luo et al. [233] experimented with acoustic signals for device-free gesture recognition, testing various surface materials' impact on recognition. Ling et al. [26] employed Channel Impulse Response (CIR) measurements for gesture recognition, providing a resolution of 7 mm. Wang et al. [234] converted CIR measurements into CIR images for gesture representation. Lastly, Wang et al. [59] gathered data from 12 volunteers performing single motions and sign language movements under different impacting factors.

B. Audio Signal Based Methodology

Table 11 summarised the existing technologies used to implement an audio signal-based HGR system by many researchers, including year, dataset information, feature extraction and classification method besides performance. This section explains details about the ML and DL-based HGR system with Audio data modality.

TABLE 11 Methdological Review of the Audio-Based HGR

Author	Year	Supported Language	Class	Sample	Feature Extraction	Classifier	Performance (%)
Siddique et al. [231]	2017	HGR	-	-	-	LDA	80.00
Sang et al. [232]	2018	HGR	-	-	CNN	Softmax	96.34
Saad et al. [230]	2018	American SL	7	128	Amplitude	SVM	96.00
Wang et al. [188]	2018	American SL	15	312	CNN	LSTM	98.40
Wang et al. [188]	2018	American SL	15	N/A	CNN	Bi-LSTM	98.80
Sang et al. [232]	2018	American SL	6	2700	Moving Trajectories, Precise Ranges, Velocities	SVM	96.32
Siddiqui et al. [235]	2020	American SL	36	N/A	RMS, MAV Skew, Kurtosis	SVM, DT, LDA	80.00
Luo et al. [233]	2020	American SL	7	256	STE, ZCR	SVM	93.20
Ling et al. [26]	2020	American SL	12	480	CNN	CNN	99.60
Luo et al. [233]	2020	HGR	-	-	CNN	Softmax	93.20
Ling et al. [26]	2020	CT-HGR	12	-	CIR	CNN	99.00
Wang et al. [234]	2020	HGR	15	-	CNN	LSTM	93.20
Wang et al. [59]	2023	CT-HGR	15	-	BiLSTM	Softmax	98.80

1) ML-Based Methods

Table 11 demonstrated various existing models for audio-based HGR. Saad et al. [230] developed an ultrasonic gesture recognition system using SVM classification, achieving a gesture detection sensitivity and

specificity of 99% each, with a classification accuracy of 96%. Siddiqui and Chan [231] explored HGR through acoustic measurements, achieving an average classification accuracy exceeding 80% using Linear Discriminant Analysis (LDA). Sang et al. [232] compared the HMM and end-to-end neural network methods, achieving accuracies of 89.38% and 96.34%, respectively. Luo et al. [233] developed an HCI mechanism achieving a gesture recognition accuracy of 93.2% for seven common gestures on smart devices. Wang et al.'s RobuCIR system [234] demonstrated high accuracy and robustness in recognizing 15 gestures, outperforming existing approaches. Wang et al. [59] achieved a combined recognition rate of 98.8% for single gestures and maintained high accuracy for continuous and sign language gestures.

2) CNN, RNN-LSTM Based Methods

Wang et al. [234] integrated CNN and LSTM networks, demonstrating high accuracy and robustness in recognizing 15 gestures. Wang et al. [59] used the CTC algorithm and Bi-LSTM network, achieving a combined recognition rate of 98.8% for single gestures and maintaining high accuracy for continuous and sign language gestures. Ling et al. [26] introduced UltraGesture with an average accuracy exceeding 99% for 12 gestures, leveraging Channel Impulse Response (CIR) measurements and a CNN model.

C. Challenges and Future Direction

Audio-based HGR systems face challenges such as ambient noise interference, varying acoustic conditions, limited gesture vocabulary, user dependency, privacy concerns, and integration difficulties. These systems must handle environmental noise, adapt to different acoustic settings, recognize diverse gestures, accommodate user variations, address privacy issues, and integrate smoothly into existing devices. Improvements in signal processing, noise reduction, ML, and user interface design are essential to enhance robustness, accuracy, and usability for broader adoption. Future work should integrate advanced algorithms like TCN-LSTM networks, develop attention-based temporal modelling, expand recognition to complex gestures, and explore novel data augmentation techniques for improved robustness.

SECTION VII.
EEG Modality Based HGR

EEG-based HGR measures brain activity using scalp electrodes, and they try to use different hand movements to create unique neural patterns. The process includes acquiring EEG data, noise removal, feature extraction, and training algorithms like support vector machines or neural networks to link EEG patterns to gestures. Real-time recognition, crucial for brain-computer interfaces, involves continuous data acquisition and processing. Challenges include low spatial resolution, non-stationary signals, inter-subject variability, and signal artefacts. Research focuses on advanced signal processing, better algorithms, and integrating modalities like EMG or motion tracking. These advancements could enhance human-computer interaction and expand applications in assistive technologies, virtual reality, and human-robot collaboration.

A. Dataset

Table 12 demonstrates thedataset information and performance of the EEG modality. AlQattan and Sepulveda [27] used an Enobio wireless system to acquire EEG data, utilizing sixteen of the twenty available channels due to connectivity issues with channels Pz, O1, and O2. The data was sampled at a rate of 500 samples per second.

TABLE 12 Databases and Performance Accuracy With EEG Modality for Existing Model

Author	Year	Dataset Name and Type	Classes	Subject	Sample	No. of EEG Channel	Feature Extraction	Classifier	Performance (%)
AlQattan et al. [27]	2017	American SL	6	-	500	16	Energy, Entropy, SD	SVM and LDA	75.00
Chaves et al. [236]	2017	American SL	40	-	50	128	LSTM	Deep Learning	90.34
Spampinato et al. [237]	2017	American SL	40	-	40	128	CNN	RNN	89.70
Al-Anbury et al. [238]	2021	American SL	10	-	6487	14	PCA	Deep Learning	95.75
Wang et al. [239]	2022	Hand Movement	4	-	-	61	Decoder-ensemble framework	Ensemble	70.00
Hossain et al. [240]	2022	Hand Movement	-	-	-	63	Phase-locking value (PLV) Multiple linear regression (MLR)	Anova	-
Altameem et al. [241]	2022	NEUROML2020	-	-	-	19	FFT	KNN, XGBoost	88.00
Tao et al. [242]	2022	Self-Created	-	-	-	30	MEMD,CNN	MECN	81.14
Y Ai et al. [243]	2023	American SL	7	-	N/A	3 to 9	Temporal Contrast Coding	Spike Encoding and Spatial Clustering	97.07
Ganesan et al. [244]	2023	Hand Movement	4	-	-	6	Wavelet and FFT	DL	97.20
Kim, et al. [245]	2023	Ultrasonic patterns finger motion	-	-	-	-	3D GB-RRAM neuromorphic sensory system	Softmax	97.9 (motion) 97.40
López et al. [28]	2024	EMG-EPN-612 (EMG)	5	-	183,600	8	Spectrograms, Rectification, Filtering	CNN, CNN-LSTM	90.55

B. EEG Based Methodology

The process of recognizing hand gestures using EEG signals involves three main steps: signal preprocessing, feature extraction, and classification.

1) Preproecssing Filtering and Signal Segmentation

EEG signals contain high noise due to their narrow strategy, necessitating preprocessing for noise removal. Researchers commonly use a bandpass filter to eliminate artefacts from the raw EEG signal, such as eye blinking, sudden sounds, and muscle movements. For Motor Imagery (MI) tasks, the EEG bandwidth is often subdivided into narrower frequency bands like Mu-band (8-13 Hz), low-beta (13-22 Hz), and high-beta (22-35 Hz) to enhance classification accuracy. Studies show that brain activity during MI tasks primarily falls between 7 Hz and 36 Hz, supporting the use of narrowband signals for feature extraction [246], [247]

2) Feature Extraction and ML Based Approaches

In the research presented by AlQattan and Sepulveda [27], EEG signal analysis involved utilizing nine different feature types: cD1, cD2, cD3, cD4, cD5, cA5, Energy, Entropy, and standard deviation. These features were employed to analyze EEG signals and identify hand movements associated with sign language from brain activity. In the research by Al-Anbary and Al-Qaraawi [238], EEG signals were classified into five types (Theta, Delta, Beta, Alpha, and Gamma waves) to capture various brain activities. Preprocessing was performed on these signals, followed by applying PCA for unsupervised feature extraction. PCA effectively reduces data dimensionality while preserving essential information, enhancing the characterization of EEG signals by emphasizing relevant information and reducing noise and non-relevant data. In the work by Spampinato et al. [237], a low-dimensional manifold within the multidimensional and temporally varying EEG signals was extracted, resulting in a 1D representation referred to as EEG features. These features primarily encoded visual data, facilitating the extraction of corresponding image descriptors for automated classification. Additionally, in [243], spike-related features were utilized by applying a temporal contrast coding scheme, translating measured analog signals into spike streams. These spike streams, consisting of positive and negative spikes, were used as features in the classification process. AlQattan and Sepulveda [27] used SVM and Linear Discriminant Analysis (LDA) algorithms for classifying EEG signals, achieving an accuracy rate of approximately 75% with the Entropy feature type. In EEG-based BCI for SLR, the SVM and LDA algorithms achieved 75% accuracy. Chaves [236] developed DL models, achieving an 89.03% accuracy in semantic image classification.

3) CNN-Based Methods

Chaves [236] introduced a universal end-to-end DL model designed to predict the semantic content of images from the ImageNet dataset, achieving an accuracy of 89.03%. In [238], the researchers employed a neural network architecture with three hidden layers and a DL classifier for classifying ten classes of EEG signals, encompassing facial expressions and motor execution processes. In another study by Chaves [236], LSTM, a recurrent neural network (RNN), extracted features from raw EEG signals. LSTM effectively captures sequential information from time-series data like EEG signals, enhancing gesture recognition accuracy. Ai and Rajendran [243] utilized Recurrent Neural Networks (RNNs) to capture discriminative brain activity related to visual categories using EEG data. They integrated spike encoding and spatial clustering techniques to accurately classify gesture signal, capturing rich spatiotemporal patterns in brain signals. The study in [27] demonstrates the potential of EEG-based motor imagery brain-computer interfaces (BCIs) for linguistic communication with paralyzed patients, achieving approximately 75% accuracy in identifying hand movements associated with sign language. In [236], DL models predicted the semantic content of images based on EEG signals, with the universal model achieving 89.03% accuracy and the personalized model

reaching 90.34%. The sign language software model in [238] achieved a high classification accuracy of 95.75% for EEG signal samples, offering a promising approach to assist speechless individuals in communicating their thoughts non-invasively. The brain-driven approach for automated object categorization using EEG signals in [237] achieved an average accuracy of approximately 83%, demonstrating the potential of brain signals for visual classification tasks. In [243], the algorithmic framework achieved superior accuracy in identifying hand gestures and motor imagery tasks, with accuracy ranging from 92.74% to 96.51%, highlighting the effectiveness of the proposed convolutional spiking neural network for BCI classification tasks. Al-Anbary and Al-Qaraawi [238] propose an EEG-based sign language software model, achieving 95.75% accuracy.

Spampinato et al. [237] achieve an average accuracy of 83% in automated object categorization via EEG signals. Wang et al. demonstrate the decoding of hand movement parameters from low-frequency EEG signals, highlighting the correlation between physical hand movements and neural activity [239]. Ganesan et al. explored spectral analysis and validation of parietal signals for different arm movements, emphasizing the differentiation of continuous EEG signals based on finger flexion movement [244]. Additionally, Hosseini Shalchyan et al. investigated the continuous decoding of hand movements from EEG signals using phase-based connectivity features, demonstrating the feasibility of this approach [240]. Furthermore, Altameem et al. analyzed the performance of ML algorithms for classifying hand motion-based EEG brain signals, focusing on controlling prosthetic hands for amputees [241]. Tao et al. proposed a novel algorithm using Multivariate Empirical Mode Decomposition and CNNs to decode multi-class EEG signals of hand movements, demonstrating advancements in decoding methodologies [242]. Zakrzewski et al. utilized EEG recordings from 52 healthy subjects engaged in motor imagery hand movements to classify tasks related to these movements, achieving notable accuracy [248]. Crell et al. demonstrate the classification of hand movements in four directions using EEG signals, with accuracies ranging from 55.9% to 80.2%. This study focused on continuous kinematic decoding of hand movements, showcasing the potential of EEG signals for accurately classifying various hand movement tasks [249]. Fujiwara and Ushiba applied deep residual CNNs to differentiate between rest, left-hand movement, and right-hand movement tasks with high accuracy [250]. They achieved precise decoding of hand movements by utilizing a larger dataset and advanced neural network architectures.

C. Future Direction

EEG-based HGR faces several challenges hindering its widespread adoption and accuracy. Artefact and noise are still considered the major challenges besides narrow problems of the EEG signal. Filtering out these unwanted signals while retaining relevant information for gesture recognition is complex. Additionally, the limited spatial resolution of EEG electrodes makes capturing fine-grained hand movements accurately difficult, which restricts the ability to decode complex gestures with high precision. Furthermore, the temporal dynamics of hand gestures present a challenge in extracting relevant features from EEG signals, as gestures involve intricate patterns that evolve over time.

Future research can explore advanced signal processing techniques to enhance the signal-to-noise ratio and extract discriminative features from EEG data effectively. In addition, it should focus on developing subject-adaptive models and exploring practical applications of EEG-based HGR in domains such as human-computer interaction and assistive technology, emphasizing usability and real-world deployment considerations.

SECTION VIII.
Multimodality Dataset Based HGR

Multimodal data-based HGR integrates various data sources, including RGB, skeleton, EMG, and EEG, to accurately interpret hand movements. This approach offers enhanced accuracy, robustness, flexibility, reduced ambiguity, and improved user experience compared to single-modality methods. Combining information from multiple perspectives provides a comprehensive understanding of gestures, making it valuable for applications in human-computer interaction, virtual reality, and SLR. Multimodal HGR (HGR)is crucial for effective human-robot interaction (HRI) [251], [252], [253].

A. Dataset

According to our study, multimodal datasets combining sEMG signals, RGB images, and depth images enhance HGR precision and reliability [192] demonstrated in Table 13. Wang et al. [254] created a dataset of 3,000 hand gesture samples across ten classes, each sample including an image with intricate backgrounds and strain data from sensors on finger knuckles. Siddiqui and Chan [235] used a dataset of 140 trials per participant, featuring parallel acoustic recordings from ten microphones and motion data from IMUs, each trial lasting three seconds. Sun et al. [255] compiled a database of 20,000 RGB and depth images representing various gestures, with each category containing 2,000 samples for training and testing. Qi et al. [256] utilized multimodal data sources, including depth vision data and sEMG signals from devices like the Leap Motion Controller and the Myo Armband. Table 14 demonstrates the multimodal datasets.

TABLE 13 Databases for Fusion Modality

Dataset Names	Year	Modality Name	Language	Class sign	Sub.	Total Samples	Latest Performance (%)
WLASL [143]	2020	RGB Skeleton	ASL	2000	-	8227	54.69 [257]
AUTSL [146]	2020	RGB Skeleton Depth	Turkey	226	43	38336	98.00 [147]
Somatosensory–visual (SV) [254]	2020	RGB, Depth	Roman numerals	10	-	300	100
IMU Signal. [235]	2020	Motion, Signal	Hand Gesture	14	10	140	75.00
DHGD,SHREC [145]	2017	RGB, Depth	Hand Gesture	14,28	20,27	2800	88.64,90.24 [193]
HRI Hand Gesture. [192]	2021	EMG,RGB	N/A	10	6	37,500	92.45
Honkong Sign Language (HKSL) [122]	2021	RGB Skeleton	Honkong	50	6	2400	94.6 [122]
SLR500	2021	RGB Skeleton	American	200	-	125000	-
NMFCSL	2021	RGB Skeleton	Chinese	500	-	32,010	78.40 [257]
MSASL	2021	RGB Skeleton	ASL	200	-	32,010	71.4 [257]
KArSL [258]	2021	RGB Skeleton Depth	Arabic	502		75,300	95.84 [51]
How2Sign [259]	2021	RGB Skeleton Depth	American	-	11	2,456	91.60 [259]
GSL [260]	2021	RGB Depth	Greece	441	100	51,080	95.68 [260]
LaRED	2023	RGB, Depth	na	1000 81	-	-	93.98 [255]
Qi et al. [256]	2023	RGB, Depth	N/A	N/A	-	-	100
PHOENIX-2014 [94]	2023	RGB, Depth	Germany	1081	9	6841	20.7 (WER)
PHOENIX-2014-T [94]	2023	RGB, Depth	Germany	1085	9	8227	21.0 (WER)

TABLE 14 Fusion Data Modality-Based HGR

Author	Year	Modality Name	Lang.	Class	Sample	Feature Extraction	Classifier	Performance (%)
Cheng et al. [107]	2016	RGB Depth	ChaLearn LI		-	C3D+LSTM	SoftMax	68.14
Sun et al. [85]	2018	RGB Depth	SKIG	10	6	C3D+LSTM	SoftMax	98.60
Wang et al. [188]	2018	RGB Depth	American SL	15	N/A	CNN	Bi-LSTM	98.80
Sang et al. [232]	2018	RGB Depth	American SL	6	2700	Moving Trajectories, Precise Ranges, Velocities	SVM	96.32
Siddiqui et al. [235]	2020	RGB Depth	American SL	36	N/A	RMS, MAV Skew, Kurtosis	SVM, DT, LDA	80.00
Nevorova, et, al. [26]	2020	RGB Depth	Chalearn 2014	20	13,858	ModDrop	CNN	96.81
Zhou et al. [261]	2020	PHOENIX -2014 CSL PHOENIX-2014-T.	RGB Skeleton	1081 - 1085	6841 - 8227	CNN + TCN	SBiLSTM + CTC	20.7 (WER) 28.6(WER) 21.0 (WER)
Gao et al. [4]	2020	RGB Depth	ASL			2S-CNN	SoftMax	92.00
Gao et al. [192]	2020	RGB Depth	ASL	10		CNN	SoftMax	-
Duarte et al. [259]	2021	RGB Depth	Multi-modal ASL	8	16000	How2Sign	DL	91.60
Mahmood et al. [193]	2021	RGB Depth	Depth and Shrec	28	2800	CRNN	Softmax	90.04
Zhou et al. [122]	2021	RGB Skeleton Depth	PHOENIX14-T CSL HKSL GSI	8227 100 50	8227 25000 2400	(3+2+1)D ResNet + BiLSTM + BERT	CTC	18.60 19.80 7.19 19.80 WER
Hu et al. [257]	2021	RGB Skeleton	WLASL MSASL SLR500 NMFCSL STB HANDS17	2000 200 100 500 1067 - -	8227 25000 125,000 32,010 18000 292820	BERT	SoftMax	97.60 59.54 - - 46.39 95.48(AUC) -
Ding et al. [24]	2022	RGB Depth	depth-grayscale	10	N/A	VGG-16 CNN	Softmax	83.88
Yang et et al. [214]	2023	RGB Depth	FMCW	8	-	Multi-modal	ML	93.30
Qi et al. [256]	2023	RGB Depth	Hand gesture	7	128	Amplitude	KNN	96.00
Liu et al. [262]	2023	RGB Signal	HGR	10	2400hr 7200 ist	Two-branch fusion network LSTM	Softmax	94.58 (cr-valid), 94.63 (out usr), 93 (in usr), 93.68 (out dark), 89 (in dark), 83 (bright light)
Duan et al. [206]	2023	sEMG ACC	DB2, DB3 DB5, DB6, DB7	50, 50, 41, 8, 41	-	CNN Transformer	AiFusion	95.28 91.13 87.04 80.62 96.76
Wang et al. [263]	2023	DB5(E+A)	HGR	-	-	HybridCNN	CNN-LSTM -CBAM	92.16
Duan et al. [264]	2023	sEMG,ACC	DB2 DB3 DB7	-	30 tr	HyFusion	Softmax	94.73 89.60 96.44
Shin et al. [51]	2024	RGB Skeleton	JSL	41	-	Handcrafted and CNN	SoftMax	90.00
Balaji et al. [265]	2024	RGB skeleton Depth	Shrec	14 28	2800	MF-HAN.	Softmax	93.93 92.26
Balaji et al. [265]	2024	RGB skeleton Depth	Shrec	14 28	2800	MF-HAN 2-stream.	Softmax	94.17 93.21
Wang et al. [266]	2024	EMG ACC	Sensor Dataset	-	-	Time domain frequency domain	Transfer Learning	Outperforms

B. Fusion Based Methodology

Many researchers have been working to develop fusion-based HGR systems. These systems commonly use ML and DL techniques for feature extraction and classification. Below, we describe the technologies involved in these systems. Figure 5 also demonstrates the multimodal data fusion-based approach outline diagram. Table 14 demonstrates the multimodal fusion-based HGR methodology and its performance.

1) ML-Based Fusion System

Siddiqui and Chan [235] extracted features and selected 25 features with the mRMR algorithm and reported 75% accuracy for the multimodal fusion approach. Qi et al. [256] achieved high accuracy and computational efficiency, with the depth vision-based k-NN classifier attaining a remarkable 100% identification accuracy for gestures, highlighting its effectiveness. Shin et al. proposed a fusion-based Japanese SLR system using a skeleton and RGB information as inputs [51] where extracted handcrafted features for skeleton and deep learning features for RGB stream. Ding et al. integrated CCD RGB-IR and depth-grayscale sensor data using CNN DL for hand gesture intention recognition, emphasizing the importance of multi-modal approaches to improve recognition performance [24].

2) CNN-Based Systems

Gao et al. [192] implemented a multiscale parallel CNN for HGR by fusing sEMG signals, RGB images, and depth images. Parallel CNN subnetworks are used to extract features for each modality and combine them to generate final features. Wang et al. [254] used a CNN to extract hierarchical deep spatial and shift-invariant features from hand gesture images. They applied a sparse neural network for feature-level fusion and recognition of sensor data, achieving high accuracy. They reported 100% accuracy in human gesture recognition, even with noisy or over/under-exposed images, and an error rate of 1.7% under standard illumination and 3.3% in low-light conditions for robot navigation through hand gestures.

Neverova et al. proposed the ModDrop model, an adaptive multi-modal gesture recognition approach utilizing multi-scale and multi-modal DL where they included gradual fusion to learn cross-modality correlations while preserving each modality's unique representation [26]. How2Sign [259] is a multimodal and multiview dataset for continuous ASL, featuring RGB, depth, and 2D key point data from both frontal and side views. The dataset includes up to 80 hours of video content. Jiang et al. [164] employed a multi-modal approach for feature extraction, incorporating RGB-based 3DCNN, hand RGB, key points, RGB frames, and RGB flow. They used the Sign Language Graph Convolution Network (SL-GCN) to capture the embedded dynamics of skeleton key points. Their model achieved the highest performance in both the RGB (98.42%) and RGB-D (98.53%) tracks, demonstrating its effectiveness in SLR. The results indicated that their approach outperforms both ResNet3D and ResNet21D, showcasing its superiority. Recently, Duan et al. proposed an AiFusion: Alignment-Enhanced Interactive Fusion Mode by integrating the CNN with various transformer levels [206]. In the study, they employed two streams, including sEMG and ACC signal extracted individual domain features and finally concatenated them to generate final features, and they reported DB2: 95.28%, DB3: 91.11%, DB5: 87.04%, DB6: 80.62%, DB7: 96.76% accuracy. Dean et al. also proposed a hybrid multimodal fusion model including DL and transformer [264]. They reported DB2: 94.73%, DB3: 89.60%, DB7: 96.44% accuracy with their model.

3) BiLSTM and Hybrid Systems

Mahmud et al. [193] utilized a variational autoencoder, CNNs, and Long-Short Term Memory (LSTM) networks for fusion-based HGR. They reported good performance accuracy and low computation time because of the effective finger motion and global motion features.

4) Transformer-Based Systems

Sun et al. [255] integrated multi-level feature fusion using a two-stream CNN, reporting a 1.08% improvement in average detection accuracy and a 3.56% increase in mean average precision (mAP) compared to a single-channel model. They achieved an average gesture recognition rate of 93.98% under occlusion and varying light conditions, enhancing robustness against challenging scenarios but facing potential limitations with dataset variability. Qi et al. [256] used an ensemble classifier combining the k-nearest neighbour method for finger angle analysis and a deep CNN for gesture identification based on sEMG signals. Despite its adaptive learning mechanism, this approach involves complexities in model training and optimization.

C. Challenges and Future Direction

Challenges in multimodal dataset-based HGR include managing data variability across different sensors, synchronizing multimodal data streams, addressing computational complexity, ensuring robustness in diverse environments, and overcoming limited training datasets. Advanced integration techniques and

efficient algorithms are required to enhance recognition accuracy and system performance. Integrating DL architectures tailored to handle data variability across different sensors and synchronizing data streams, reducing computational complexity, can be fruitful work in this domain in the future.

SECTION IX.
Discussion

In this study, we review the current advancements in HGR (HGR) across diverse data modalities. We present a comprehensive framework that addresses key elements of feature extraction using ML, as well as end-to-end DL modules. Each data modality is explored in detail, including primary datasets, preprocessing techniques, proposed architectures, fusion methodologies, state-of-the-art performance, and the challenges and future trends specific to each modality. For dynamic HGR, we emphasize the importance of managing temporal contextual features. This section aims to guide future research by highlighting the challenges and potential directions involving 3D DL models. We evaluate the associated problems, complexities, and the need for practical solutions within the field, providing a clear pathway for future advancements.

We identify and discuss new research gaps and offer guidance to overcome challenges in next-generation HGR technologies. Multimodal datasets, which integrate hand-crafted features with new CNN features combined with RGB, skeleton, and depth-based information, are examined for their efficacy in improving gesture recognition accuracy. In the skeleton modality section, we focus on GCNs and their application to large-scale, real-time HGR. This has become a focal point of the research community due to its potential to solve significant challenges. We highlight the need for gesture localization within realistic, uncut, and extended videos, predicting that emerging challenges such as early recognition, multi-task learning, gesture captioning, recognition from low-resolution sequences, and life log devices will gain increased attention in the coming years. Our discussion underscores the necessity of addressing these challenges to advance HGR systems, particularly emphasising multimodal approaches and the integration of new technologies. By providing a detailed overview of the current state and future directions, we hope this study serves as a guideline for researchers and practitioners in the field of HGR.

SECTION X.
Challenges and Future Directions in Diverse Data Modality-Based HGR

Currently, research in diverse data modality-based HGR (HGR) faces several significant challenges. While isolated gesture recognition achieves high accuracy in small to medium-sized datasets, it struggles with large-scale datasets. Continuous gesture recognition performs well in simple scenarios but has room for improvement in complex environments. The following are specific challenges and future trends in this field:

A. Short Duration of Videos

One major challenge in HGR (HGR) is the short length of videos in most datasets. Real-world gestures often form long sequences without clear breaks between gestures. Current methods struggle with these long sequences due to limitations like RNNs' long-term dependency issues and the high computational demand of Transformer models. Future research should focus on improving models' ability to process long sequences and accurately segment gestures within continuous streams while also handling shorter sequences effectively.

B. Unseen Gestures and User Independence

Recognizing gestures and unseen gestures from different users is crucial for practical HGR applications. User independence means the model performs well across various individuals despite differences in gesture speed and body dimensions. Many current methods capture specific traits of individuals, leading to performance drops with new users. Recognizing unseen gestures, where test gestures weren't seen during training, is also challenging. Future research should develop methods that allow models to learn generalized features from large datasets and improve algorithms for gesture segmentation to handle unseen gestures effectively.

C. Generalization Beyond Training Data

The wide range of hand gestures makes it impractical to include all possible gestures in a dataset. Few-shot and zero-shot learning approaches help models recognize new gestures with minimal training data. While progress has been made in isolated gesture recognition using these techniques, continuous gesture recognition remains largely unexplored. Advancing these techniques for both isolated and continuous gestures is crucial for creating robust and versatile HGR systems.

D. Multi-Person Recognition

Most current datasets focus on single-person scenarios, but real-world applications require robust recognition in multi-person environments. Models need to distinguish between gestures from the intended user and interference from others. Future research should focus on developing robust models that accurately focus on the target user while ignoring background interference.

E. Model Complexity and Deployment

Many HGR systems achieve high accuracy on servers but are too complex to use on portable devices. The goal is to create HGR systems that aid everyday communication, which requires lightweight models suitable for mobile use. Future research should prioritize developing efficient models that maintain high accuracy and are feasible for real-time use on portable devices.

F. Online Recognition

Current methods often use recorded datasets for validation, allowing models to access future context within sequences. Real-time gesture recognition requires models to predict based solely on past information, as future data isn't available. Developing methods that can perform online recognition and accurately predict gestures from one-directional sequences is crucial. Improving real-time performance will enhance the practicality and usability of HGR systems in dynamic, real-world scenarios.

SECTION XI.
Conclusion

This comprehensive review has provided an in-depth analysis of advancements in vision-based HGR (HGR) for sign language from 2014 to 2024. By examining over 250 articles from reputable online databases, we have highlighted significant achievements and identified critical areas needing further exploration. The findings reveal a dynamic research landscape with a steady stream of publications across various journals and conferences, focusing predominantly on three critical aspects: data collection, data contextualization, and hand gesture representation. The review underscores the efficacy of HGR systems, particularly in terms of recognition accuracy, which remains a crucial benchmark in this domain. Notably, there is a significant gap in research on continuous gesture recognition, emphasizing the need for further efforts to enhance the precision and applicability of vision-based gesture recognition systems. This gap offers valuable perspectives for future research directions and highlights the evolving nature of this important field. The advantages of current HGR systems include improved accuracy, robustness in varied conditions, and the potential for real-time application. Future work should focus on addressing the challenges in continuous gesture recognition, integrating multimodal approaches, and developing more sophisticated models that can handle diverse and complex gesture patterns. These efforts will be crucial in advancing the capabilities and practical applications of HGR systems, ultimately contributing to more natural and efficient human-computer interactions.

Abbreviations

Abbreviation	Expansion
HCI	Human-Computer Interfacing.
BCI	Brain-Computer Interface.
EEG	Electroencephalography.
MEG	Magnetoencephalogram.

RQ	Research Question.
HGR	Hand Gesture Recognition.
ASL	American Sign Language.
ArSL	Arabic Sign Language.
ISL	Indian Sign Language.
KSL	Kurdish Sign Language.
JSL	Japanese sign language.
PSL	Persian Sign Language.
GSL	German Sign Language.
DSL	Danish Sign Language.
SLRS	Sign Language Recognition System.
CNN	Convolutional Neural Network.
ADDSL	Annotated Dataset for Danish Sign Language.
ML	Machine Learning.
DL	Deep Learning.
HMM	Hidden Markov Model.
HOG	Histogram of Oriented Gradient.
PCA	Principal Component Analysis.
CRNN	Convolutional Recurrent Neural Network.
LSTM	Long Short-Term Memory.
Bi-LSTM	Bidirectional Long Short-Term Memory.
SVM	Support Vector Machine.
ArSLRS	Arabic Sign Language Recognition System.
CTC	Connectionist Temporal Classification.
RTDHGRS	Real-Time Dynamic HGR System.
SMKD	Self-Mutual Knowledge Distillation.
CTC	connectionist temporal classification.
MUD	Massey University Dataset.
ASLAD	American Sign Language Alphabet Dataset.
SSC-DNN	Spotted Hyena-based Sine Cosine. Chimp Optimization Algorithm with Deep Neural Network.
DMD	Dynamic Mode Decomposition.

MsMHA-VTN	Multiscaled Multi-Head Attention Video Transformer Network.
HSL	Hongkong Sign Language.
FPHA	First Person Hand Action.
STr-GCN	Spatial Graph Convolutional. Network and Transformer Graph Encoder for 3D HGR.
MF-HAN	Multimodal Fusion Hierarchical Self-Attention Network.
SMLT	Simultaneous Multi-Loss Training.
ResGCNeXt	Efficient Graph Convolution Network.
HDCAM	Hierarchical Depth-wise Convolution and Attention Mechanism.
HOG	Histogram of Orientation Gradient.
HOEF	Histogram of Edge Frequency.
IMU	Measurement Unit.
MLP	Multilayer Perceptron.

Authors
Figures
References
Citations
Keywords
Metrics

Back to Results