A Systematic Review of Hand Gesture Recognition: An Update From 2018 to 2024

Publisher: IEEE Cite This PDF

Abdirahman Osman Hashi ; Siti Zaiton Mohd Hashim ; Azurah Bte Asamah **All Authors**

Abstract:
Hand gesture is the main method of communication for people who are hearing-impaired, which poses a difficulty for millions of individuals worldwide when engaging with those who do not have hearing impairments. The significance of technology in enhancing accessibility and thereby increasing the quality of life for individuals with hearing impairments is universally recognized. Therefore, this study conducts a systematic review of existing literature review on hand gesture recognition, with a particular focus on existing methods that address the application of vision, sensor, and hybrid-based methods in the context of hand gesture recognition. This systematic review covers the period from 2018 to 2023, making use of prominent databases including IEEE Xplore, Science Direct, Scopus, and Web of Science. The chosen articles were carefully examined according to predetermined criteria for inclusion and disqualification. Our main focus was on evaluating the hand gesture representation, data acquisition, and accuracy of vision, sensor, and hybrid-based methods for recognizing hand gestures. The accuracy of discernment in scenarios that rely on the specific signer varies from 64% to 98%, with an average of 87.9% among the studies that were analyzed. On the other hand, in situations where the signer's identity is not important, the accuracy of recognition ranges from 52% to 98%, with an average of 79% based on the research analyzed. The problems observed in continuous gesture identification highlight the need for more research efforts to improve the practical feasibility of vision-based gesture recognition systems. The findings also indicate that the size of the dataset continues to be a significant obstacle to hand gesture detection. Hence, this study seeks to provide a guide for future research by examining the academic motivations, challenges, and recommendations in the developing field of sign language recognition.

The graphical abstract analyzes hand gesture recognition, showing that 74% focus on double-hand gestures. Key parameters include hand orientation (31%) and shape (25%). F... Show More

Published in: IEEE Access (Volume: 12)

Page(s): 143599 - 143626

DOI: 10.1109/ACCESS.2024.3421992

Date of Publication: 02 July 2024

Publisher: IEEE

Electronic ISSN: 2169-3536

Funding Agency:

SECTION I.
Introduction

In accordance with data from the World-Federation of the Hearing impaired and the World-Health-Organization, an estimated 72 million individuals worldwide grapple with hearing impaired-muteness,

encompassing a total hearing impaired population of 360 million, with 32 million among them being children [1]. A significant portion of the speech- and hearing-impaired populace encounters challenges in conventional literacy [2]. For this, communication for the hearing impaired and mute predominantly occurs through gestures such as sign-language, which relies on manual expressions, encompassing finger shapes, hand gestures, and facial expressions to convey meaning. Despite its critical role in bridging communication gaps, sign language does have its limitations. These include the necessity for broad hand movements, a limited vocabulary, and the complexity inherent in mastering this form of communication [3].

Meanwhile, gesture, defined as deliberate and expressive bodily motions involving the hands, fingers, face, arms, body, or head, serve as a prominent form of non-verbal communication in human interactions [4], [5].

Within the technology domain, tools of authority find heightened preference for leveraging hand gestures over alternative forms of gestural communication. This preference is substantiated by the facilitation of navigation and sustenance within the technology environment through the use of hand gestures [6]. Consequently, an imperative arises for the employment of appropriate approaches in gesture recognition to interpret human hand gestures within machine learning settings. Gesture recognition, in this context, pertains to the identification of class labels from videos or images featuring gestures executed by users. This recognition capability assumes paramount significance in discerning and responding to the nuanced intricacies of hand gestures within the virtual domain [7].

In the domain of virtual reality (VR), the application of gesture recognition models is pervasive across various fields [8]. For instance, author [9] proposed an application for computer mouse control, using an algorithm and specific hand features to optimize performance and enhance user comfort. In a different context, author [10] presented a method for automatic gesture recognition intended to recognize hand gestures during virtual reality (VR) training in crane rigging operations. Similarly. proposing a universal approach, author [11] recommended a standardized set of gestures for VR interaction, drawing inspiration from the versatility observed in desktop computing mouse control across diverse applications.

Hand movements in this context are classified into two distinct categories: “static” and “dynamic.” A static-gesture is similar to a signature, in which the precise hand movements do not contribute significantly to the gesture. Instead, the hand itself is of utmost importance [12]. On the contrary, dynamic hand gestures are contingent upon both the shape and motion of the hand, constituting essential components of the gesture and a critical aspect of human motion perception. Nevertheless, the complexity of this task is further complicated by the greater variety in hand shapes and significant interferences found between fingers, which poses a significant difficulty for accurately capturing dynamic hand movements utilizing single-camera video sensors. The performance of video-based hand gesture detection is greatly limited by these limitations [12].

Recent decades have witnessed the advent of potent depth sensors, including the Microsoft Kinect sensor and LMC, which have significantly enhanced the ability to segment objects and recognize three-dimensional hand movements. The essential elements are recognizing gestures, identifying hand features, acquiring data, and localizing the hand based on the recognized features [13]. Conventional methods for collecting data entail using color cameras, which have proven to be effective in tasks involving recognizing gestures [14]. However, these systems are vulnerable to changes in lighting conditions, sensitivity to clutter, and reliance on skin color. Moreover, video capturing entails a fundamental constraint linked to the velocity of actions.

On the other hand, the aspiration to empower machines to align with human intentions has been a longstanding pursuit since the advent of machinery [15]. During the early stages of machine development, manual interfaces like as buttons and joysticks were used to control the machine’s circuitry and transmit commands to the machine through mechanical transmissions. The evolution of technology has seen a paradigm shift, particularly with the introduction of computers in modern times [16]. The Human-Machine Interface (HMI) has become increasingly user-friendly, allowing individuals to communicate with machines through input devices such as mice and keyboards, while simultaneously monitoring machine activities through display interfaces [17].

In recent years, the diminishing size of computers has corresponded with a noteworthy enhancement in machine efficiency. This evolution signifies that elementary input signals can now suffice to command machines to autonomously execute intricate tasks programmatically. The aforementioned technological advancements have served as the foundation for the creation of diverse human-computer-interaction (HCI) technologies [18]. Voice control has emerged as a prominent HCI modality, enabling users to articulate commands and directives, thereby fostering a more intuitive interaction between humans and machines [19]. Moreover, brain-computer interface (BCI) has become increasingly important, enabling direct connection

between the human brain and machines, surpassing conventional input technologies. Gesture recognition technology represents another significant stride in HCI, wherein machine comprehension of human gestures facilitates a more natural and gestural means of communication. These advancements collectively reflect a trajectory in HCI research and development towards creating interfaces that align seamlessly with human cognitive and physical modalities [18].

Despite these commendable strides in HCI technologies, challenges persist, particularly in the nuanced interpretation of diverse human inputs. The intricate nature of human gestures and the subtleties inherent in linguistic expressions pose ongoing challenges for machine comprehension. Researchers and developers are thus engaged in refining these technologies, delving into sophisticated algorithms and innovative sensor technologies to enhance the precision and versatility of human-machine interactions. As the symbiotic relationship between humans and machines continues to evolve, the pursuit of refining HCI technologies remains an imperative endeavor to foster a more intuitive, efficient, and user-centric interface paradigm [20].

Given the progress made in technology, different approaches have been created to accurately determine the position of hands within collected data. A traditional method involves setting a depth threshold, where depth data is analyzed either through (automated or empirical) [21]. Empirical-solutions define the limits of the potential search area by using trial and error, guiding computational efforts towards locating the hand within this narrowed scope. On the other hand, automatic techniques involve determining the closest point/location to the camera, with the hand being recognized as the nearest object in the scene [22]. These methods can also use additional reference information, such as facial color data and head position, to decide the likely position of the hand inside the gesture.

Nevertheless, previous investigations [23] have identified limitations in the current approaches. A significant issue emerges from the inherent variety and adaptability of human gestures. Even when an individual repeats a gesture, there may be slight differences between each repetition, requiring a complex and flexible technique to recognize motions. Additionally, the limitations manifest in the attempt to encapsulate the entirety of human gesture diversity within a predefined framework, prompting the need for more nuanced and context-sensitive methodologies [22].

The quest for an effective hand localization solution is further complicated by the dynamic and multifaceted nature of human movements. The complex coordination of multiple elements, such as the hand's orientation, its spatial relationship with other objects in the scene, and the temporal dimensions of the gesture, all contribute to the intricacy associated with precise hand localization [24]. Consequently, researchers are compelled to explore sophisticated algorithms and models that can not only adapt to the diversity of gestures but also account for the temporal dynamics and contextual relevance inherent in human interactions. Moreover, addressing the challenge of occlusions in hand localization remains a pivotal aspect of ongoing research. Occlusions, occurring when one object obstructs the view of another, introduce ambiguity and complexity to the hand localization process [25]. Developing algorithms capable of robustly handling occlusions is imperative for ensuring the reliability and accuracy of hand localization systems, particularly in real-world scenarios where occlusions are commonplace and researchers proposed a SLR [26].

SLR represents a critical domain of investigation with the overarching objective of ameliorating communication barriers for the hearing impaired-mute community as mentioned. This area of research seeks to employ computer vision technology to translate sign language gestures into either textual or spoken formats [27]. Its multidisciplinary nature integrates elements of computer science, artificial intelligence, and linguistics to address the intricate challenges posed by the swift and highly coarticulated motions inherent in sign language. The complexity of recognizing gesture sequences within sign language compounds the difficulty of this endeavor [28].

Recent studies in deep learning techniques have shown a good progress in the field of sign language recognition, presenting encouraging outcome. Innovative frameworks leveraging deep learning architectures exhibit promise for achieving signer-independent sign language recognition [29], [30]. Various surveys and reviews have undertaken the task of delineating the technical methodologies, obstacles, and prospective avenues for research within sign language recognition. These comprehensive assessments underscore the pivotal role played by classification algorithms, deep-learning systems, and CNN based methods in advancing the field [30]. Emphasis is also placed on the significance of datasets and the categorization of sign language gestures, elucidating their crucial role in facilitating progress [31].

Furthermore, the purview of sign language recognition extends beyond general applications to encompass specific domains such as fingerspelling recognition and hand pose recognition. These domain-specific

advancements cater to the diverse linguistic needs inherent in different sign languages. Consequently, the focus of research lies in the creation of SL systems that are designed specifically for particular SL, such as Arabic SL and Indigenous People SL [32]. This underscores the imperative of adopting language-specific approaches to account for the nuanced intricacies of different sign languages. These language-specific endeavors not only enhance the inclusivity of sign language recognition but also contribute to a more comprehensive understanding of the diverse linguistic modalities within the hearing impaired-mute community. Moreover, the integration of contextual and contextualized information within sign language recognition systems represents an area of burgeoning interest. The integration of contextual indicators, including body language and facial expressions, is designed to enhance the precision and resilience of the recognition procedure. This holistic approach acknowledges the multimodal nature of sign language communication, recognizing that gestures are embedded within a broader context of non-verbal expression [33].

For that reason, the integration of real-world applications into the research landscape of sign language recognition is gaining prominence. Efforts are focused on creating practical and user-friendly applications to enable smooth communication. These applications extend beyond the theoretical realm, seeking to bridge the communication gap in practical, everyday scenarios. The collaborative synergy of researchers, technologists, and linguists is crucial in translating the advancements in sign language recognition research into tangible tools that empower and enhance the communicative capabilities of the hearing impaired-mute population.

Thus, the aim of this research is to perform a thorough examination of current methods used for hand gesture recognition, more specifically for sign language recognition (SLR), in order to gain a present state of hand gesture recognition systems in this field. Furthermore, this study attempts to present a roadmap for the technological evolution of SLR systems, delineating their features and elucidating the existing limitations of current technology. To achieve this objective, the research framework revolves around addressing six distinct research questions:

(1) How may one categorize studies that pertain to approaches for sign language recognition?

(2) To what extent do preliminary studies elaborate on Sign Language Recognition (SLR) systems?

(3) What types of movements have been identified in prior research efforts focused on gesture recognition?

(4) What sensors have been explored in primary publications focusing on the exploration of sensing technologies?

(5) How have vision-based methodologies been examined in primary publications within the context of the study?

(6) How are hybrid methodologies assessed in primary publications within the scope of the study?

The methodology used in this study adheres to the principles outlined in Noraini's paper [8], Zinah's paper [3] and other pertinent research publications, thereby ensuring a systematic approach to both the Systematic Literature Review and the Systematic Mapping Study conducted herein.

SECTION II.
Problem Background

Communication stands as an inherent and vital facet deeply ingrained in the fabric of human existence, permeating various aspects of daily life in ways often overlooked due to its pervasive nature [34]. Its profound significance extends beyond mere verbal exchange, encompassing nuanced social and emotional dimensions. A deficiency in communication, conversely, can instigate sentiments of isolation, frustration, and solitude [35]. Individuals with hearing impairments, for instance, encounter challenges in perceiving the auditory landscape, rendering them unable to hear ambient sounds and, crucially, their own voices [22]. While these individuals can communicate effectively among themselves using SL, a structured mode of gesture and visual expressions, bridging the communication gap with those unaffected by hearing impairments remains a formidable challenge [36]. Sign language, involving intricate movements of various body parts including

hands, arms, head, shoulders, and facial expressions, encounters limited understanding within the wider hearing community as mentioned before. This limited comprehension contributes to a tangible barrier, obstructing meaningful interaction between individuals with hearing impairments and the broader society.

In light of the existing communication obstacles, numerous researchers have suggested to implement of SLR system, presenting a promising solution to address challenges experienced by individuals with hearing impairments and the broader community. As an example, in a study by [37], an unsupervised learning methodology was suggested as a potential solution to the issue of hand movement in continuous sign language. By implementing vision-based modelling, the establishment of such a system, known as a Sign Language Recognition (SLR) system, is intended to improve the identification and comprehension of sign language gestures. The principal objective is to foster inclusiveness and eradicate enduring communication obstacles that exist between individuals with hearing impairments and the broader community. The SLR system serves as a technological intervention designed to enhance communication accessibility and address the prevailing challenges engendered by the limited understanding of sign language within the broader populace [38].

Hence, the complexity of sign language in the context of hand gesture, stemming from its reliance on intricate visual motions and signs, necessitates a nuanced and technologically sophisticated approach for accurate recognition. For that, the implementation of an effective SLR system requires a comprehensive understanding of the diverse elements encompassed within sign language, ranging from the precise movements of fingers to the subtleties of facial expressions [39]. Recognizing the multifaceted nature of sign language, the SLR system becomes a pivotal tool for overcoming communication disparities by deciphering the rich vocabulary embedded within visual gestures.

Meanwhile, concurrent with technological advancements, ethical considerations emerge in the development and utilization of SLR systems. Ensuring the privacy, consent, and dignity of individuals using sign language is paramount. The responsible design of such systems must encompass robust privacy safeguards to protect the sensitive nature of communication and uphold the autonomy of users [40]. Ethical considerations also extend to the potential cultural biases embedded in the algorithms, necessitating a conscientious approach to system development that accounts for the diverse linguistic and cultural backgrounds associated with sign languages [41]. It represents a technological progress that allows the automatic transformation of SLG into either written or spoken forms. This functionality establishes it as an extremely valuable form of HCI, notably seen in the creation of supportive systems such as sign language interpreters [5].

Within the domain of sign language, every expression consists of two essential parts: the manual and non-manual elements. The manual aspects encompass factors like hand movement, orientation, location, and shape, while the non-manual elements include body-posture, mouth gestures, and facial-expressions [3]. It is noteworthy that the primary means of conveying signs predominantly relies on the manual components, as depicted in Figure 1.

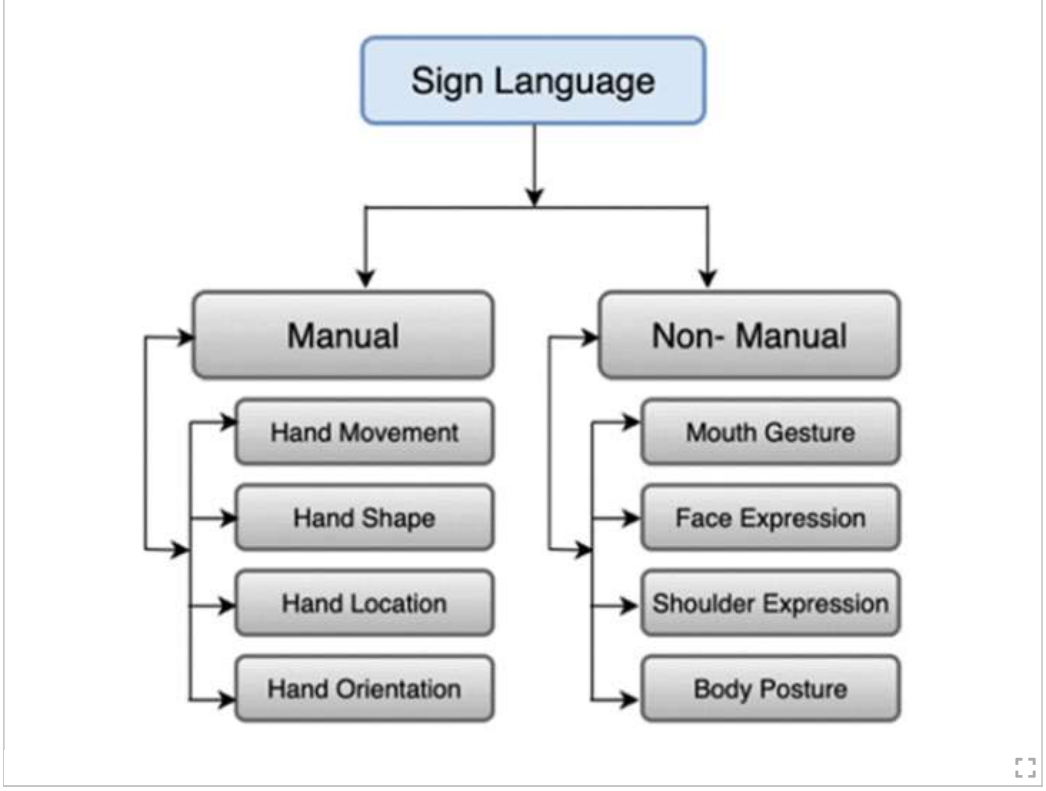


FIGURE 1.
Sign Language comprises two integral parts.

The range of manual signals includes-hand gestures and hand-movements, which can vary from static-gestures to dynamic-gestures depending on the individual application used [25]. Static gestures contain immobile hand positions, while dynamic gestures encompass complex hand movements and quick changes between different positions. The identification of hand motions has been investigated using three separate methodologies: (vision-based), (sensor-based), and (hybrid-based) [3], as illustrated in Figure 2.

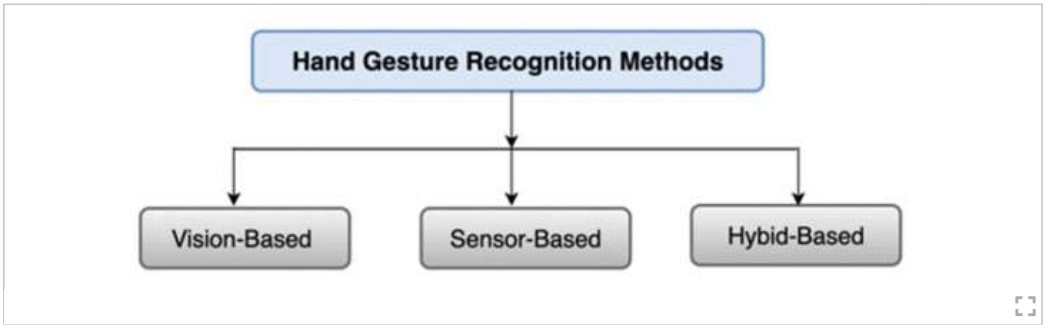


FIGURE 2.
Hand gesture recognition methods.

As evident from Figure 2, the technology's role in facilitating sign language recognition extends beyond mere translation, assuming a pivotal role in fostering more inclusive human-computer interactions. Assistive' systems' advancement, particularly sign-language-interpreters, stands as a testament to the practical applications of this technology in ameliorating communication challenges. This strategic imperative aligns with broader societal goals of inclusivity and equal participation, seeking to empower individuals with hearing impairments in their interactions with the wider community [42]. The nuanced analysis of the manual and dynamic components of sign language serves as a foundational understanding, paving the way for the technological interventions and methodologies used in the recognition of HG. As the field progresses, the ongoing refinement of recognition techniques, coupled with ethical considerations, will contribute to the continual evolution of technology in enhancing communication accessibility and fostering a more inclusive society [43].

The vision-based method entails capturing gesture picture data using a camera and then utilizing image-processing technology to identify motions [43]. This system is designed to be easily used by the user, eliminating the necessity for the user to wear any additional equipment. However, the development of this technology is hard, requiring sophisticated and extensive calculations for the formulation of algorithms in feature and movement recognition [44]. In addition, it is prone to problems related to fluctuations in lighting conditions [13], [45], [46]. In contrast, the sensor-based technique requires the use of a sensory-glove-device to accurately measure finger bending, hand position, and movement. The strategy utilizing a data glove achieves superior precision, rapid reactivity, and improved maneuverability [5]. Nevertheless, this approach places a strict limitation on the structure of the hand, leading to a certain level of discomfort [47]. Significantly, it removes the need for preprocessing and segmentation [41].

The hybrid-based technology, which combines both vision and glove-based-approaches, and it is used to enhance the quality of visual data by integrating sensor readings. Nevertheless, the implementation of this method has been restricted because of the costs and computational burdens involved with the overall system, leading to a scarcity of study in this area [48].

The contrast between the vision-based and sensor-based techniques underscores the trade-offs between user-friendliness and computational complexity [3]. Vision-based systems, while advantageous in their non-intrusiveness, necessitate intricate algorithmic developments and are sensitive to environmental lighting changes [49]. On the other hand, sensor-based methods, particularly those involving data gloves, offer enhanced accuracy and mobility but impose physical constraints and discomfort on users. The hybrid-based approach attempts to exploit the importance of both methods, aiming to improve the overall effectiveness of gesture recognition systems [50]. However, the limited exploration of this hybrid method underscores the challenges posed by economic considerations and computational complexities in its implementation. As technology progresses, addressing these challenges will contribute to the refinement and wider adoption of gesture recognition systems, thereby enhancing their usability and effectiveness in diverse applications [2].

SECTION III.
Overall Systematic Review Protocol

This study is grounded in the methodology of a systematic literature review, a rigorous analytical approach designed to facilitate a comprehensive exploration of SLR in the domain of the hand gesture as mentioned before. The systematic review entails a meticulously structured process encompassing key stages, including the identification of the research domain, screening, formulation of search methodologies, establishment of criteria for study selection, systematic extraction of pertinent data, and the subsequent synthesis of accumulated information. Renowned for its inherent significance and versatility, the systematic literature review method is adept at accommodating various research methodologies. The overarching objective in applying this method is to succinctly encapsulate the focal theme under investigation, thereby illuminating gaps in existing research and providing a foundational basis for the inception of novel research endeavors.

This methodological approach allows for the synthesis of diverse research findings, contributing to a comprehensive comprehension of the state-of-the-art in SLR. Additionally, the systematic literature review serves as a strategic tool to identify lacunae and areas requiring further exploration within the existing body of knowledge. The upcoming figure 3 shows the overall process that we followed from identification to included process.

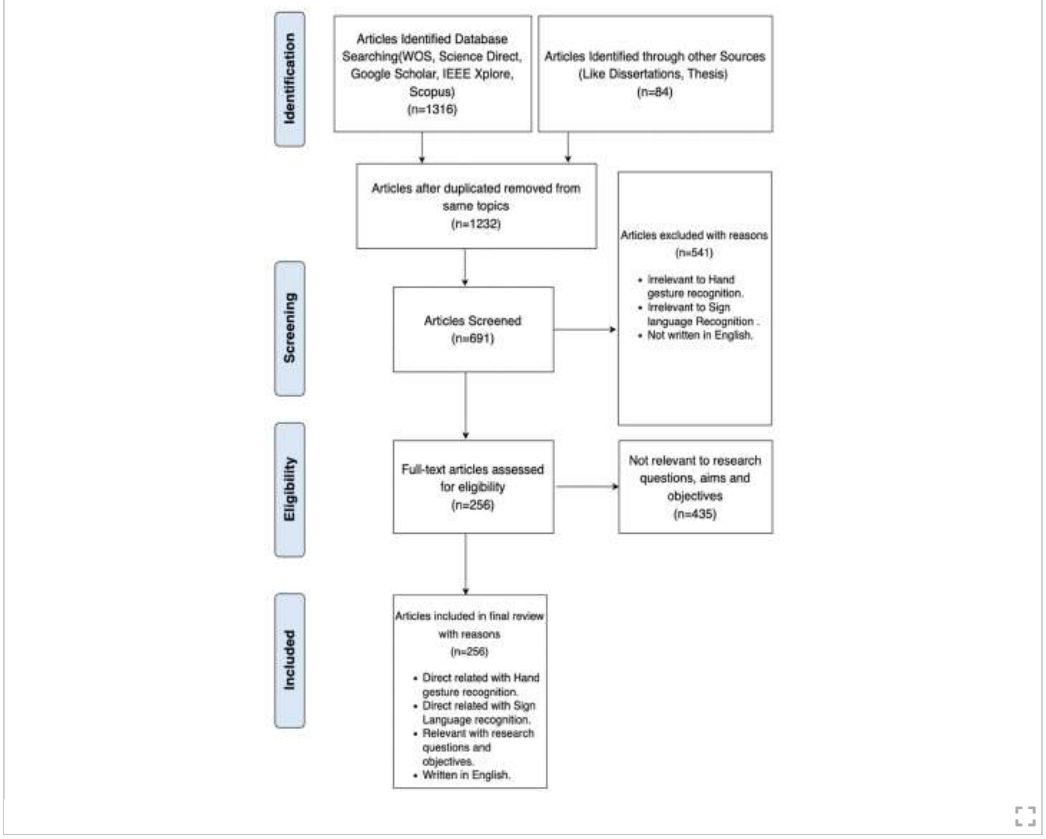


FIGURE 3.
Flowchart of study selection process.

A. Information Source

The selection process was refined by focusing on specific articles and utilizing key digital databases to ensure comprehensive coverage. These databases include:

1) Science Direct Database: This database was chosen for its comprehensive coverage of research papers spanning diverse fields. Its selection is intended to facilitate a thorough examination of scientific endeavors, providing an extensive overview and encompassing pertinent technical literature.

2) Scopus Database: Similar to Science Direct, the Scopus database is a valuable resource offering a plethora of publication materials across different research domains.

3) IEEE Xplore Database: Curated by the Institute of Electrical and Electronics Engineers (IEEE), is a comprehensive collection of scholarly articles and publications that specifically cover technological and engineering advancements across numerous topic areas in technology.

4) Web of Science (WoS): Recognized for its broad spectrum of publications, Web of Science encompasses diverse disciplines.

B. Reseacrh Strategy

The inception of this research project occurred in September 2023. Utilizing the advanced search features of WoS, ScienceDirect, IEEE Xplore, and Scopus databases, a comprehensive examination was conducted considering the distribution of scientific papers spanning the period from 2018 to 2023. In conducting our analysis, a combination of keywords, namely ‘Sign Language,’ ’Hand Gesture,’ ‘sensors,’ and ‘hybrid,’ was employed. These keywords were interconnected using the ‘AND’ operator in our search queries. The specific query text used in this study, depicted in figure 3, is (‘Sign-Language’ and Hand Gesture’) AND (‘Hybrid’) AND (‘Sensors’).

C. Selection of Studies

This study particularly examined two categories of articles: journal papers and conference papers. It intentionally excluded alternative forms, such as book chapters, and also considered the preferences of each search engine in the search process. The search conducted on the four chosen databases produced a combined total of 1,316 articles as primary research findings. These articles were distributed as follows: 92 from Web of Science, 128 from ScienceDirect, 483 from IEEE Xplore, and 613 from Scopus. Out of the total number of 84 theses and dissertations, they were discovered and then eliminated, resulting in 1,232 distinct papers remaining. After carefully reviewing the abstracts and titles, 541 papers were eliminated because they did not meet any of the inclusion criteria. After excluding articles that were not pertinent to the study issues, a total of 256 studies were left for a comprehensive examination of their full texts. A total of 256 studies were included in the qualitative synthesis as a result of this method. The next sections will cover the literature on hand-gesture-recognition, specifically focusing on vision, sensor, and hybrid-based approaches.

SECTION IV.
Review of Literature on Vision-Based

In vision-based, the procedure generally consists of many phases: data gathering, preprocessing, segmentation, feature-extraction, and classification. These stages are categorized as shown in figure 4 [51].

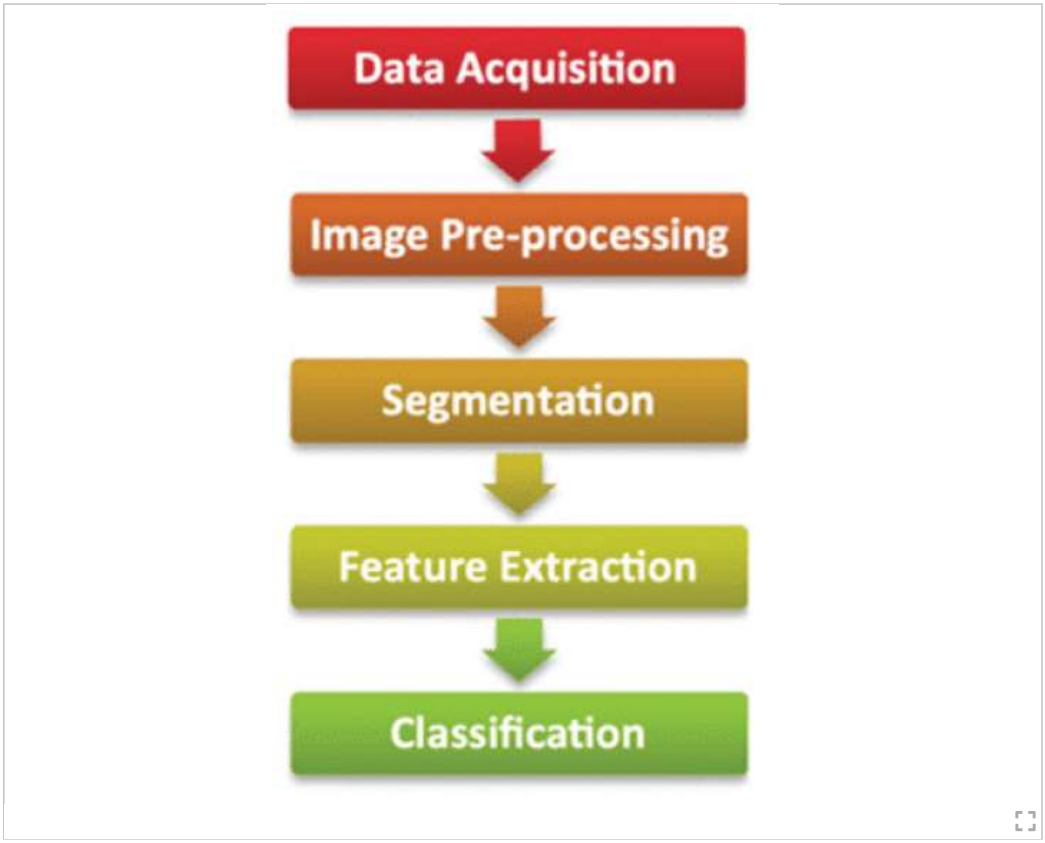


FIGURE 4.
Vision-based recognition process.

Static gesture recognition involves analyzing individual image frames, while dynamic sign languages include analyzing continuous video frames. The main distinction between vision-based approaches and sensor-based approaches lies in their respective techniques for acquiring data [12]. The methods and approaches employed by researchers in the field of vision-based gesture-recognition are examined in the sections that follow.

A. Data Acquisition

Gesture recognition systems that rely on vision-based data acquisition utilize sequences of images. These systems use various image-capturing devices, including video cameras, webcams, stereo cameras, infrared cameras, and more sophisticated active methods such as Kinect and LMC (Light Measuring Camera). Vision-

based approaches analyze the visual information captured by these devices to interpret and recognize gestures [52]. Stereo cameras, Kinect, and LMC are 3D cameras that capture depth and visual data. Figure 5 shows that stereo cameras, Kinect, and LMC use three-dimensional depth information to improve gesture detection. These device cameras improve recognition system robustness and accuracy by recognizing spatial relationships and hand movements [53] The capturing device chosen relies on the application's precision and operating environment [37].

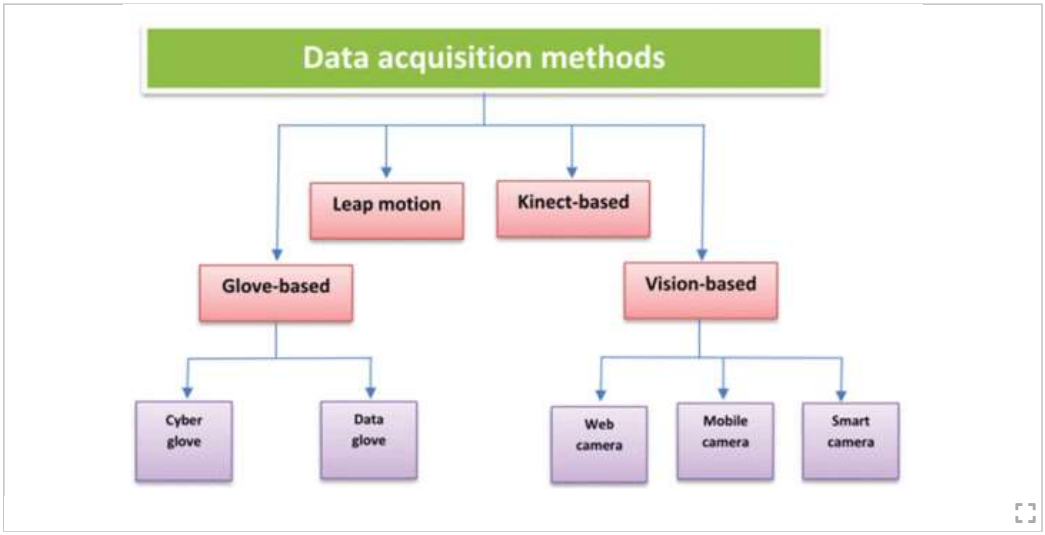


FIGURE 5.
Data acquisition methods.

B. Image Pre-Processing

Image pre-processing is the phase where image/video inputs are altered to improve the system's performance. Commonly used methods for reducing noise in captured photos or videos involve using median filters and Gaussian filters. It is worth mentioning that in certain research cases [41], just median filtering has been used during the pre-processing phase. In addition, morphological techniques are widely used to remove undesirable information. For example, author [37] utilized a sequence in which the input image initially converted into a binary image using a thresholding technique. Then, gaussian and median filters were applied to eliminate any noise present in the image. Following then, morphological operations were utilized as an essential pre-processing phase.

In certain research studies, captured images undergo downsizing image resolution before moving to other stages. As shown in reference [12], decreasing the input image resolution has been shown to improve computational performance. The research [37] presented significant findings, which included a detailed analysis of the time required for processing images at various levels of downsizing in terms of resolution. The study found that dividing the scale by 64 was the most effective approach, leading to a significant 43.8% decrease in processing time.

Moreover, in [50], histogram equalization is applied to enhance the contrast of input images taken in diverse environmental conditions. This technique serves to standardize the brightness and illumination of images, ensuring uniform quality across diverse settings. The incorporation of these pre-processing techniques underscores the meticulous attention given to refining input data, mitigating noise, and optimizing computational efficiency in the pursuit of more accurate and reliable gesture recognition systems.

C. Segmentation

In the domain of image processing, image segmentation pertains to the procedure of partitioning images into discrete and recognizable components [54]. This step focuses on isolating the Region-of-Interest (ROI) from the rest of the image. Segmentation techniques are broadly categorized into two types: contextual and non-contextual. Contextual segmentation involves examining spatial relationships between features, often employing techniques like edge detection. In contrast, non-contextual segmentation ignores spatial connections and groups pixels based on overall characteristics [54].

1) Color Segmentation

Color segmentation is typically performed using different color spaces, including RGB, YCbCr, HSV, and HSI [33]. The task of achieving reliable color segmentation is hindered by issues pertaining to sensitivity to lighting conditions, camera specifications, and variations in skin tones [54]. The preference for using the HSV color space is attributed to the notable variations in hue between the palm and arm, which greatly simplifies the process of separating the palm from the arm [55]. Research [9] specifically examines the segmentation of the face and hand using the HSV color. In contrast, the author [56] perform skin-color segmentation in the RGB color by using the criterion of R > G > B and comparing it with pre-existing skin-color samples to detect different skin tones. Additional inquiries, such as a study [57], confirm that YCbCr demonstrates superior resilience for the segmentation of skin color in comparison to HSV when faced with different lighting situations. Further research [58], [59] demonstrates that the CIE Lab color space provides greater resilience than YCbCr when faced with various lighting conditions. In [60], a normalized RG is introduced to specifically tackle the non-uniformity vulnerability of RGB.

To address the drawbacks associated with fixed skin color thresholds, approaches incorporating skin-color distribution and categorization based on skin-color models have been proposed as a strategy to surpass the limitations imposed by constant thresholds for skin-color. The author conducted skin-color segmentation using the YCbCr color [61]. Gaussian model utilizing the YCbCr color space is utilized in [62] to accurately identify skin pixels from the background. Additionally, the author [63] employs a system that is similar to [62], but with the difference of utilizing a Gaussian model instead of a histogram-model. The authors in [64] propose a dynamic method for modelling skin-color by including measuring elements into both locally trained and globally trained skin models. This approach leads to the development of an adaptable skin color model. The objective of these developments in skin color segmentation approaches is to improve precision and flexibility in various circumstances.

2) Other Segmentation Method

The author [65] proposed a segmentation technique that relies on the disparity between the background picture, which proves to be highly efficient in complex backdrop situations. The method begins with applying the Otsu thresholding technique to the photos. This is then followed by implementing the '3sprincipal' method, which aims to maximize the variance between different classes. In their work, author [41] introduced a framework called Hand-Tracking-Segmentation (HTS) in their publication. They utilized the Continuously-Adaptive-Mean-Shift (CAMShift) algorithm in the HSV color space to generate skin pixels of histogram. The histogram is subsequently employed to ascertain the appropriate threshold value for segmentation. The subsequent stages involve performing Canny edge detection, followed by the application of dilation and erosion processes. In the end, an edge traversal technique is used to differentiate the movement of the hand from the surrounding environment.

In a comparative study carried out by the author [66], the effectiveness of 10 distinct approaches, including Sobel edge detection, low pass filtering, histogram equalization, skin color segmentation in the HSI color space, and desaturation, was assessed. The findings demonstrated that desaturation achieved the highest accuracy. The desaturation method involves converting the image into grayscale by eliminating the chromatic channel and retaining only the intensity channel within the HSI color space. This technique proved superior in terms of accuracy, highlighting its efficacy in image processing applications.

The author [67] investigated the use of entropy measurement to extract hand motion information by subtracting adjacent image frames. This approach encompasses the quantification of entropy, the extraction of the hand-region from images, the monitoring of the hand-region, and the identification of hand gestures. In addition, a method called Entropy Analysis and is suggested in [5], which combines entropy and skin colour data to accurately separate hand motions, even when there are static and intricate backdrops.

D. Feature-Extraction

Feature extraction involves the transformation of significant elements within input data into condensed sets of feature vectors [68]. This process is essential in pattern recognition and data analysis, as it helps distill relevant information from the raw input, facilitating more efficient and effective processing. Within the context of gesture recognition, the extracted features should include pertinent information from the input of hand movements. These features should be presented in a concise way that distinguishes the gesture being classed from other gestures. These can be classified as PCA, LDA, and FEFD.

1) Principal-Component-Analysis

Principal Component Analysis (PCA) is a mathematical technique employing orthogonal transformation to convert a set of potentially correlated variables in observations into a set of values known as principal components [69]. In the context of a training set comprising M images represented by an S-dimensional vector, PCA identifies a subspace of dimension t. Within this subspace, the basis vectors denote the

directions of maximum variance in the original image space. The dimensionality of this new subspace is typically reduced, with t being significantly smaller than s. The mean, denoted as μ , is calculated for each image in the training set, where x represents the ith image, and its columns are concatenated [20].

This process of dimensionality reduction is particularly valuable in simplifying and retaining the essential features of the data, making it more manageable and computationally efficient for subsequent analysis or modeling.

PCA is commonly employed as a method for reducing dimensionality, by converting potentially correlated-variables into a smaller set of uncorrelated main components [69]. PCA is utilized in [70] to extract-features from a dataset consisting of 28 MSL. Locality-Preserving-Projection (LPP) is a kind of Principal Component Analysis (PCA) that modifies the distances between feature vectors by taking into account the known similarities between input features. The performance of PCA and LPP was compared in [71], with PCA obtaining an accuracy of 92.8% and LPP achieving an accuracy of 93.8%. The authors in [72] used PCA features as indicators of hand configuration and orientation. They combine PCA and chain code to improve accuracy. In [72], the authors utilize PCA for reducing dimensionality, specifically by excluding components beyond the 12th using eigenvalue calculations, aiming to decrease computing complexity. In addition, in reference [73], Principal Component Analysis (PCA) features from 27 categories of Very Short Lived (VSL) are identified using Mahalanobis distance, resulting in an accuracy rate of 90.7%.

2) Linear-Discriminant-Analysis

Both LDA and PCA methodologies aim to discover linear combinations of features that effectively capture the intrinsic characteristics of the data. In Equation 1, the expressions denote the between-class scatter matrix (SB) and the within-class scatter matrix (SW), respectively. These matrices are computed by considering all data points across all classes.

$$S_B = \sum^{M} M_i \left(x_i - \mu\cdot\right)\left(x_i - \mu\right)^T$$
$$S_W = \sum^{M} \sum^{M} M_i \left(x_k - \mu\cdot\right)\left(x_k - \mu\right)^T \quad (1)$$

View Source

In this context, M_i indicates the number of training samples in class i, c represents the total number of unique classes, μ i signifies the mean vector of samples belonging to class i, and xk is the kth image of the class.

In this context, M_i denotes the quantity of training samples in class i, c stands for the overall count of distinct classes, μ_i represents the mean vector of samples associated with class i, and x_k denotes the kth image within the class.

The objective of LDA is to find the matrix W, denoted as max SB/SW, which maximizes the between-class scatter while simultaneously minimizing the within-class scatter. The matrix W is a transformation matrix that projects the samples onto a space with reduced dimensions. LDA improves the distinction between different classes of items by identifying a linear combination of features that successfully discriminates among them [42]. PCA, in contrast, specifically aims to identify the primary direction of greatest variability across characteristics and does not consider variations within classes [74].

Linear Discriminant Analysis (LDA) has a dual function, acting as both a linear classifier and a method for reducing dimensionality. The author in [74] obtained PCA and LDA features from five categories of gestures, with PCA obtaining an accuracy of 26% and LDA exhibiting a perfect accuracy of 100%. The underwhelming performance of Principal Component Analysis (PCA) can be ascribed to the issue of overfitting. In a similar vein, a study conducted in [75] examined the precision of Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) by utilising five distinct categories and 50 input photos. The findings revealed that PCA achieved an accuracy rate of 60%, while LDA achieved a slightly higher accuracy rate of 62%.

Author [76] utilized LDA in the recognition of Arabic-sign-language, employing a methodology that closely resembled the one described in reference [72]. In the beginning, SIFT features were recovered from the photos. However, in this work, LDA was later used to improve the distinction between different classes of sign languages.

3) Feature-Extraction in Frequency-Domain

In the frequency domain, feature extraction involves transforming input data from the time domain to the frequency domain through techniques like Cosine Transform, Fourier Transform, and Wavelet Transform. The authors in reference [77] emphasized the beneficial size-invariant characteristics of Fourier Descriptors (FD). Furthermore, FD demonstrates rotation invariance, meaning that alterations in hand motions caused by rotation only lead to a shift in phase. Eliminating high-frequency components is an efficient method for minimizing noise, as noise and quantization mistakes mostly lead to localized fluctuations in high frequency.

On the contrary, according to [69], contour-based features like FD, Wavelet Descriptors (WD), and B-spline may encounter performance issues, especially when fingers are curled inward, leading to the loss of contour properties. In contrast, region-based features, exemplified by the Principal Curvature-Based Region detector (PCBR), leverage semi-local structural information such as curvilinear shapes and edges. This approach demonstrates resilience to variations in intensity, color, and shape, addressing the limitations of contour-based methods. For instance, the 2-D Wavelet Packet Decomposition (WPD-2) in [20] employs Haar basis functions up to level two, effectively utilizing high-frequency channels containing significant information. In a hybrid feature extraction approach discussed in [78], which incorporates PCBR, WPD-2, and Convexity defect, the system successfully recognizes 23 static ISL. This hybrid outperforms combinations involving only two features when employing k-NN and SVM classifiers. Similar positive outcomes are reported in [79], where Discrete Wavelet Transform (DWT) features are extracted for the classification of 23 static PSL. DWT, achieved through iterative filtering operations with rescaling, determines the signal's resolution [80]. This highlights the significance of choosing feature extraction methods that are robust to various hand configurations and can effectively capture essential information for accurate gesture recognition.

4) Other Feature Extraction Methods

Some traits have advantages over others, but they can have problems. As an illustration, SURF is considerably more computationally efficient than SIFT [81]. Nevertheless, it does not possess the rotational and illumination invariance that is present in SIFT [75]. In order to overcome these restrictions, researchers have utilized hybrid feature extraction methods in numerous investigations. The hand motions were analyzed in [82] using Hu moment invariant geometric characteristics, which were then integrated with SURF. The suggested technique was evaluated against SIFT, SURF, and Hu-moment using a hybrid-SVM and k-NN as a classifier. The findings indicated that the combination of SURF-with-Hu moment yielded the maximum level of accuracy.

Author [83] developed an additional hybrid feature fusion method that incorporates Hu moment invariants, finger angle count and non-skin color angle. This hybrid model had a commendable accuracy rate of 90% in correctly identifying and matching ten distinct motions. The study conducted in [84] utilized the Local Binary Pattern (LBP), a texturing operator known for its computational efficiency, to extract features from Chinese and Bangladeshi numeral gesture datasets. The achieved accuracies were 87.13% and 85.10% for the respective datasets. The shape descriptors of HOG and ZIM were utilized in [85] to categories 30 classes of Libras. The rotational invariance of ZIM was utilized as a feature, leading to a total accuracy of 96.77%. In a study, the author [86] conducted a comparison of four different strategies for gesture recognition: Subtraction, Gradient, PCA, and Rotation Invariant. The results showed that the Rotation-Invariant methodology, which is LBP, achieved the highest level of accuracy.

E. Classification

Machine learning classification involves the use of supervised and unsupervised algorithms [87]. In supervised machine learning, the system undergoes training to identify distinct patterns in input data, and this obtained knowledge is then utilized to make predictions about future data. This method entails using a collection of pre-existing labelled training data to deduce a function that helps identify patterns in new, unlabeled data [88]. In contrast, unsupervised machine learning is specifically designed to extract meaningful information from datasets in which the input data does not have a labelled response [89]. Here is the example that scholars have used.

1) Support Vector Machine

Support Vector Machines (SVM) is a supervised machine learning approach designed to identify the optimal hyperplane for categorizing input data points. SVM achieves the maximum margin around the separating hyperplane by employing optimization approaches [90]. Two hyperplanes, which accurately describe the data, are identified. The research conducted on gesture databases provide evidence that linear kernel SVM outperforms non-linear Gaussian kernels. Specifically, the accuracy of linear SVM classification with a set of 14 ESL declined from 99.2% to 82.3% as the number of gestures increased to 25 ESL. Encouraging outcomes have been attained through significant investigations exploring the application of Scale-Invariant Feature Transform (SIFT) in feature extraction. Subsequently, these features undergo quantization via K-means

clustering, and the mapping into a Bag-of-Features (BoF) is performed, followed by classification using Support Vector Machines (SVM).

Meanwhile, Proximal SVM (PSVM) is a different method that replaces the inequality requirement in SVM with an equality constraint. Utilized in diverse research fields, such as [91], PSVM effectively manages multiple categories, with a classification accuracy of 91% in the analysis of 30 TSL. The accuracy of multi-dimensional-classification using non-linear-SVM is higher than that of linear SVM, as shown in [92]. In another study [93], researchers extracted SIFT features from 30 instances of ArSL, resulting in an impressive accuracy of 99% using 7 training photos for each instance.

2) Artificial-Neuralnetwork

Artificial Neural Networks (ANNs) are computational systems designed to process information, mirroring the performance characteristics observed in biological neural networks. In essence, ANNs simulate the way biological neurons interconnect and communicate to enable learning and information processing in a machine context [93]. ANN can be defined by three essential parameters: the arrangement of connections between different layers of neurons, the given weights for these connections, and the activation function that determines the behavior of each neuron. Neurons receive inputs (x_1, x_2, . . . , x_n) that are associated with weights (w_1, w_2, . . . , w_n) indicating their permeability. The function of a neuron is represented as a nonlinear combination of weighted inputs, as depicted in Equation 2, where k denotes the activation-function.

$$y = K \sum i = 1wi \cdot xi \tag{2}$$

View Source

Author [21] utilized Artificial Neural Networks (ANN) to train a system capable of recognizing 15 different gestures. This was achieved by using a dataset consisting of 7392 gesture signals. By utilizing a solitary Artificial Neural Network (ANN) comprising of 45 input nodes, 14 output nodes, and two hidden layers, they were able to attain an average accuracy of 97.01%. The Gesture-Recognition-Fuzzy Neural Network (GRFNN) [59] included fuzzy control to optimize learning parameters, thereby reducing the requirement of preselecting training patterns and enhancing accuracy. The model attained an accuracy of 93.19% in recognizing 36 motions of American Sign Language (ASL). The Time Delay Neural Network (TDNN) is specifically built to handle continuous input, whereas the Multi-Layered-Perceptron Neural Network (MLPNN) is a feedforward neural network that is capable of distinguishing non-linearly separable data.

Meanwhile, in a study conducted by author [94], a Multilayer Perceptron Neural Network (MLPNN) was employed to classify 32 classes of Persian Sign Language (PSL), achieving an accuracy of 93.06%. This was accomplished using 92 input nodes, a hidden layer with 21 neurons, and five linear output neurons. In contrast, a Recurrent Neural Network (RNN) establishes a directed cycle within its connections. The Elman RNN, a specific type of recurrent neural network, features adaptable forward connections and unchanging recurrent connections, allowing the network to retain information from the immediate past. Notably, the application of the Simulated Annealing training approach in conjunction with back-propagation has demonstrated promising outcomes for dynamic sequence training in both [85], [95]. This illustrates the versatility of neural networks, particularly MLPNN and RNN, in effectively handling complex tasks such as sign language classification, and the significance of employing advanced training approaches to enhance their performance.

3) K-Nearest Neighbor (KNN)

The K-Nearest Neighbors (K-NN) method is a non-parametric statistical approach for classifying input data based on the majority vote from its neighboring data points. The class assigned to the data is determined by the most prevalent class among its k closest neighbors. The measure of similarity is often calculated using the Euclidean distance, as defined in Equation 3.

$$Distance = \sum (a_i - b_i)^2. \tag{3}$$

View Source

In this context, the Euclidean distance is calculated for each testing-data-point in relation to the training-data-points. The testing-data is labelled based on the majority classes among the k-th nearest training data-points. A study [96] conducted a comparison between the K-NN algorithm and the parametric Bayes

classifier, demonstrating that the former outperformed the latter. In another study [97], K-NN was employed to classify 30 test photos for each of 26 gestures. The results showed an outstanding overall accuracy of 90%. Nevertheless, other research that have compared the accuracy of SVM and K-NN using equivalent train and test data sizes have consistently demonstrated that K-NN generally exhibits a comparatively lower overall accuracy [98]. However, K-NN is advantageous due to its computational efficiency and ease of implementation.

SECTION V.
Literature Review on Sensor-Based

Sensor-based gesture recognition literature is examined in this section. These systems use sensors on users to track finger and hand placements, movements, and trajectories. This eliminates vision-based gesture detection's pre-processing and segmentation steps. The data, including finger flex angles, orientation, and absolute hand location, is usually shown in 3D. This representation incorporates depth information to estimate gesture distance from sensors. In sensor-based techniques, users wear gloves or use arm probes. Because proper identification requires precise equipment configuration, these procedures are frequently limited to controlled laboratory conditions. Data glove and EMG are the most popular sensor-based gesture detection options due of their effectiveness.

A. Data Glove

The Inertial Measurement Unit (IMU) sensors, which include gyroscopes and accelerometers, are utilized by gesture and sign language recognition systems that make use of data gloves. These sensors facilitate the collection of information regarding orientation, angular movement, and acceleration values [99]. The integration of flex sensors in certain data gloves enables the collection of data pertaining to finger flexion. Figure 6 showcases the VPL-Data glove, a glove that is fitted with flex sensors and fiber optic transducers. This allows for the precise measurement of flex angles and orientation data. The study conducted by [100] utilized 16 unprocessed data streams captured by the VPL-Data glove to categorize the movements of both hands, classifying them into ten fundamental gestures. These movement patterns were employed as inputs for a Fuzzy Min–max Neural Network (FMNN), resulting in an impressive accuracy rate of 85% in recognizing 25 words in Korean Sign Language (KSL). In simpler terms, the study effectively employed a sophisticated neural network, specifically the FMNN, to interpret intricate hand movements captured by the VPL-Data glove. This method demonstrated a high level of accuracy in identifying a considerable number of gestures within the context of Korean Sign Language.

FIGURE 6.
VPL-Data Glove.

In addition to the application of data gloves, another study [101] specifically aimed to identify and interpret 250 words in Taiwanese Sign Language. The retrieved features from the Data-Glove included finger-flexion. These characteristics were further employed as inputs for Hidden Markov Models (HMMs) to identify 51 different body positions, six different orientations, and eight distinct patterns of movement. Notably, the study attained a perfect accuracy rate of 100% in all three categories. The authors also performed experiments including individual hand movements, brief statements, and lengthier statements consisting of 250 words, achieving accuracy rates of 89.5%, 70.4%, and 81.6%, respectively.

In a similar [101], a study by researchers [102] harnessed electromyographic to develop myoelectric prosthetic hands for disabled people. By using HMMs, the system successfully recognized ten dynamic gestures with an impressive accuracy of 91.64%. These findings collectively underscore the efficacy of data gloves in capturing intricate hand movements and gestures, demonstrating their potential for robust sign language recognition.

Beyond the hardware specifications, the choice of recognition algorithms plays an important role in the measurement of performance of gesture recognition systems. The utilization of Fuzzy Min–max Neural Networks (FMNN) and Hidden Markov Models (HMMs) in the mentioned studies reflects the versatility of these algorithms in effectively processing and interpreting the complex data streams obtained from data gloves. Additionally, the studies provide insights into the system's adaptability to different sign languages, showcasing the potential for broader applicability and inclusivity [101].

The continuous advancements in sensor technology, particularly in the domain of data gloves, contribute significantly to the refinement of gesture recognition systems. The integration of IMU sensors, flex sensors, and other sophisticated components enhances the granularity of data capture, enabling a more nuanced analysis of hand movements [103]. This heightened precision is crucial for recognizing intricate sign language gestures, where subtle variations in hand positions and motions convey distinct meanings. Furthermore, the recognition of dynamic gestures, as demonstrated in the studies [104], signifies a step forward in addressing the challenges posed by the coarticulated and context-dependent nature of sign languages.

B. Electromyogrphy (EMG)

Electromyography (EMG) is a crucial method for capturing the electrical signals produced by muscle tissues. This technique involves using electrodes that are either attached to the skin-body or introduced directly into the muscles. The investigation done by the author [105] involved the use of a combination of 6-axis accelerometer input and 10-channel EMG signals linked to the user's hand. The system achieved a remarkable accuracy of 93.1% by utilizing Fuzzy K-means clustering as a classifier to identify 72 dynamic Chinese Sign Language (CSL) movements. The use of electromyography (EMG) with accelerometer data demonstrates the potential of integrating diverse sensor modalities to enhance the precision of gesture identification.

In a parallel attempt, Author [106] explored the application of EMG sensors attached to the user's arm to capture finger-movement. Leveraging a linear combination of Bayes and KNN classifier, the study achieved a commendable accuracy of 94% in classifying 20 different gestures. This research exemplifies the versatility of EMG sensors in capturing intricate muscle movements for gesture recognition applications.

Author [107] delved into the extraction of EMG pattern signatures for various movements, employing Artificial Neural Networks (ANN) for signal classification based on distinctive features. Notably, the study demonstrates the potential of EMG-based classification systems in discerning diverse muscle activities, paving the way for nuanced gesture recognition. The Myo armband, equipped with both Inertial Measurement Unit (IMU) and EMG sensors, has emerged as a comprehensive tool for gesture recognition. In a study by researchers [108], the Myo armband was employed to recognize 20 classes of American Sign Language (Libras). Leveraging the Support Vector Machine classifier, the average accuracy reached an impressive 98.6%, highlighting the efficacy of combining IMU and EMG data for robust gesture recognition.

In pursuit of a hybrid approach, researchers in [27] combined vision input from Leap Motion Controller (LMC) with surface EMG (SEMG). The study achieved an accuracy of 86% using SEMG alone, and with the inclusion of LMC depth-camera-input, the accuracy substantially increased to 96%. This amalgamation of vision-based and EMG data exemplifies the synergistic potential of multimodal sensor integration for improved gesture recognition performance.

Furthermore, the integration of SEMG and Cyberglove was explored in [109] to classify the flexion of all five fingers. ICA and PCA was used to mitigate computational complexity. The study demonstrated the feasibility

of combining surface EMG with glove-based input for accurate finger movement classification, achieving a commendable accuracy of 90% using Linear Discriminant Analysis (LDA).

These studies collectively underscore the diverse applications of EMG in gesture recognition, showcasing its adaptability to capture muscle activities associated with intricate hand and finger movements. The integration of EMG with other sensor modalities, such as accelerometers and depth cameras, further enhances the richness of data input, contributing to more robust and accurate gesture recognition systems. As technology continues to advance, the integration of EMG in multimodal sensor setups holds promise for advancing the field of gesture recognition and fostering enhanced human-computer interaction [110].

SECTION VI.
Literature Review on Hybrid-Based

Meanwhile, on hybrid-based, gesture recognition systems have undergone significant advancements, with researchers exploring hybrid-based approaches that amalgamate multiple techniques to enhance accuracy and robustness [111]. Hybrid models integrate distinct technologies, often combining vision-based and sensor-based methods, to capitalize on the strengths of each. This literature review section explores into the current landscape of hybrid-based gesture recognition, summarizing key findings, methodologies, and technological innovations.

The integration of vision based and sensor-based methods in gesture-recognition systems represents a pivotal advancement, addressing the limitations of standalone approaches. This hybridization is grounded in the principle of combining the strengths of both modalities to achieve more accurate and versatile recognition of gestures. In the literature, several studies have shown the effectiveness of such integrative approaches [112].

According to a study by [51] the integration of vision-based and sensor-based methods enhances the overall robustness of gesture recognition systems. Author [51] used a combination of depth-sensing vision technology and inertial measurement unit (IMU) sensors, showcasing how vision data can provide context while IMU sensors capture fine-grained hand and finger movements. This synergy resulted in improved accuracy, especially in scenarios where one modality alone might face challenges.

In a similar vein, the work of [113] explored the fusion of visual and tactile sensing for gesture recognition. Author [113] used a vision-based system in conjunction with tactile sensors embedded in a glove. The tactile sensors provided detailed information about the pressure and contact points during hand movements. This hybrid model demonstrated increased sensitivity to subtle gestures and a reduction in false positives compared to using either modality in isolation.

The importance of integrating vision and sensor data is highlighted in a study by [35]. This author [35] proposed a gesture recognition system that combined computer vision with data from wearable sensors. By utilizing vision data for gross movement detection and sensor data for precise hand and finger positioning, the hybrid system achieved a more comprehensive understanding of gestures, particularly in complex scenarios.

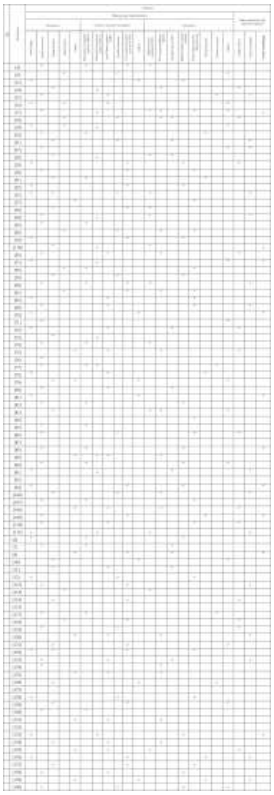
One of the key advantages of integrating vision and sensor-based techniques is the adaptability to diverse usage scenarios. The study by [114] demonstrated a hybrid system that seamlessly transitioned between vision-based and sensor-based recognition depending on environmental conditions. In low-light situations, the system automatically switched to rely more on sensor data, showcasing the flexibility afforded by hybrid models.

However, challenges persist in achieving seamless integration. For example, addressing temporal misalignments between vision and sensor data remains an active area of research [115]. Additionally, the computational demands of processing data from multiple modalities in real-time require careful optimization strategies [116].

In Table 1, a detailed literature review on sign language recognition for hand gesture recognition is presented. This table functions as both a historical review of academic contributions and an analytical instrument that analyses and examines the development and improvement of techniques throughout time. Each reference in the table represents a specific study, including information on the different parameters used, the complex

techniques used for feature extraction, the many classifiers implemented, and the methods used by researchers to get the data.

TABLE 1 Literature Review of Sign Language Recognition



Furthermore, the choice of classifier in each research is an essential element of the identification process. The table displays a range of methods, including traditional machine learning algorithms such as Support Vector Machines (SVM) and more innovative deep learning networks. The data gathering section provides detailed information on the necessary hardware and software for each research, including high-resolution cameras for vision-based detection and advanced sensors for catching subtle motions. This thorough review provides an up-to-date summary of the present state of the art and also serves as a starting point for future study, by identifying effective strategies and pointing out areas that need additional examination.

SECTION VII.
Discussion

As we can see from table 1, several scholars have suggested various techniques and conducted numerous investigations to enhance the recognition of sign language in the field of hand gesture recognition. The process of analyzing hand gesture recognition involves categorizing obstacles into several groups based on common characteristics. This helps academics and individuals understand the difficulties involved in this field. The obstacles can be categorized into four main groups: user-related challenges, hardware-related challenges, processing-related challenges, and limits connected to signs.

Firstly, challenges associated with users encompass a spectrum of factors influencing the effectiveness of SLR systems. These encompass user diversity, variability in signing gestures (hand position, hand movement), and the need for individualized adaptations based on the unique characteristics of users. As highlighted by [20] accommodating the diverse signing styles exhibited by individuals remains a persistent challenge, necessitating the development of adaptive models capable of discerning and adapting to idiosyncratic signing patterns.

The second cluster of challenges pertains to hardware-related issues in the realm of SLR systems. Hardware challenges involve the selection and integration of appropriate sensing devices for gesture capture, ensuring accuracy, reliability, and minimization of environmental interference. The work of [18] underscores the need for robust hardware configurations, addressing challenges such as environmental noise and variations in lighting conditions, which can significantly impact the accuracy-of-gesture recognition systems.

Processing-related challenges constitute the third category, encompassing issues related to the extraction, analysis, and interpretation of SLR gestures. Computational complexities, algorithmic efficiency, and real-time processing constraints are integral aspects demanding careful consideration. Author [21] posit that addressing the intricacies of processing demands collaborative efforts to optimize algorithms for swift and accurate recognition, particularly in dynamic signing scenarios.

The final category of challenges emanates from limitations intrinsic to signs themselves. This encompasses variations in SLR across different communities, regional dialects, and evolving linguistic expressions. Author [56] emphasize the necessity of accommodating these sign-related intricacies, emphasizing the importance of devising adaptive SLR models capable of discerning and adapting to the dynamic linguistic landscape of sign language.

Meanwhile, in the upcoming Figure 7 presents a comprehensive breakdown of key elements in sign language recognition systems. The first pie chart illustrates the distribution of single and double hands, while the second pie chart delineates the considered parameters. The third pie chart provides an overview of feature extraction methods, followed by the fourth pie chart detailing the employed classifiers. Finally, the fifth pie chart encapsulates the diversity of languages incorporated in the analyzed systems.

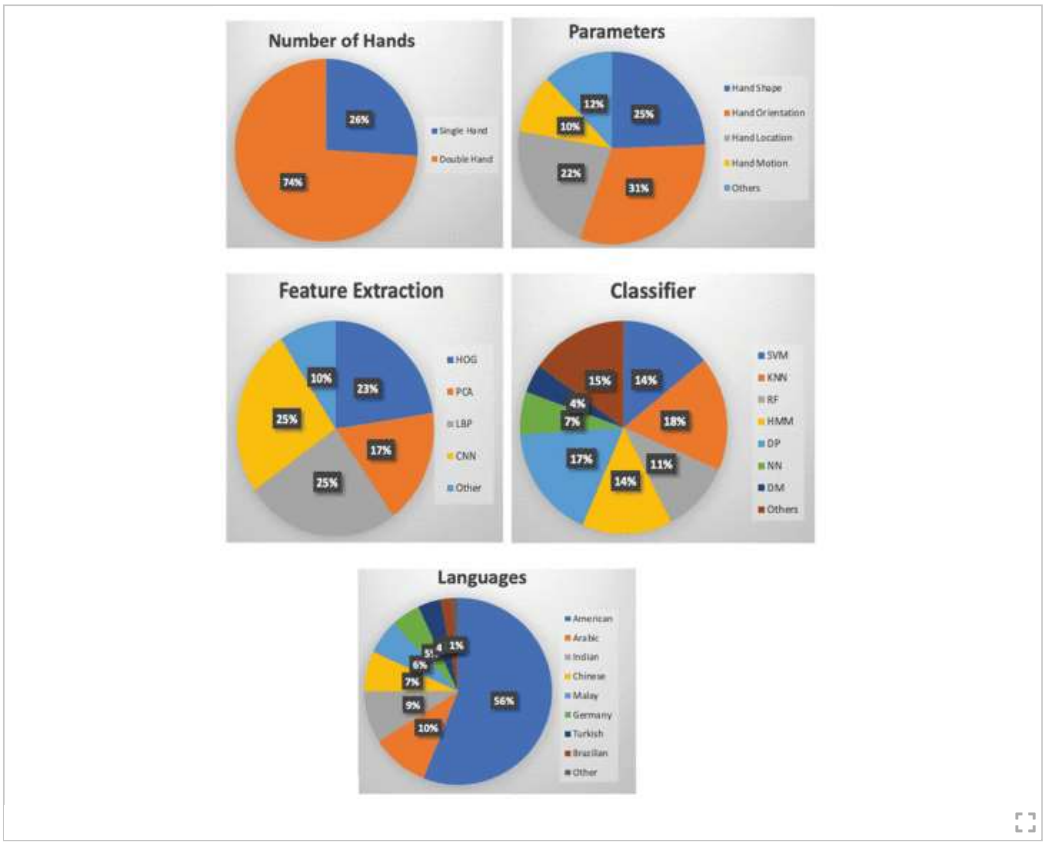


FIGURE 7.
First pie, number of hands single/double. Second pie, considered parameters. Third pie, feature extraction methods. Fourth pie, used classifiers and last pie shows languages.

Through of the analysis, a notable distinction in the prevalence of single-hand and double-hand configurations within the scope of SLR. The ratio of 26 instances of single-hand gestures to 74 instances of double-hand gestures suggests a substantial emphasis on recognizing and interpreting the complexities associated with dual-hand expressions.

Within the corpus of examined literature in Table 1, the classification of gestures based on hand configuration emerges as a significant parameter. The distribution of gestures between single-hand and double-hand configurations serves as a foundational aspect influencing the design and performance of SLR.

The literature study demonstrates a wide range of feature extraction approaches used in sign language recognition. The exploration of techniques such as Histogram of Oriented Gradients (HOG), Principal Component Analysis (PCA), Local Binary Patterns (LBP), Convolutional Neural Networks (CNN), and others, highlights the ongoing effort to capture distinctive features from sign language expressions.

A comprehensive range of classifiers has been investigated in the reviewed literature as well, encompassing (SVM), KNN), (RF), (HMM), (DP), (NN), (DM), and others. The selection of classifiers is indicative of the ongoing quest for optimal algorithms capable of accurately categorizing diverse sign language gestures.

The literature under consideration reflects a global perspective on sign language recognition, encompassing various languages. American Sign Language (ASL) takes precedence with 56 instances, followed by Arabic, Indigenous People, Chinese, Malay, German, Turkish, Brazilian, and other languages. This highlights the cross-cultural applicability and importance of creating recognition systems tailored to different linguistic nuances.

The upcoming Table 2 will demonstrate the literature review of gesture recognition. It will discuss the focus of each proposed methodology, followed by discussions.

TABLE 2 Literature Review of Gesture Recognition

Ref	Authors	Year	Focus
[141]	Gao et al.	2022	Dynamic Hand Gesture Recognition
[142]	Ding & Zheng	2022	Depth-sensor-based Hand Gesture Recognition
[143]	Wang et al.	2020	Continuous Hand Gesture Recognition Method
[144]	Jhansi*	2020	Hand Calculator System using CNN
[145]	J et al.	2020	Hand Gesture Recognition with ML Algorithms
[146]	Du et al.	2018	Gesture Recognition Based on Depth Info
[147]	Nanani et al	2018	Real-time Hand Gesture Recognition System
[148]	Prabhu & Sasikala	2018	Survey on Hand Gesture Recognition Systems
[149]	Kolhe et al.	2017	Part-Based Hand Gesture Recognition
[150]	Gao et al.	2017	Static Hand Gesture Recognition
[151]	Cheng et al.	2016	3D Hand Gesture Recognition
[152]	Rahman & Afrin	2013	Multiclass Support Vector Machine
[153]	Sarkar et al.	2013	Hand Gesture Recognition Systems Survey
[154]	Ibraheem & Khan	2012	Various Gesture Recognition Technologies

As it can be seen from this table, over the past few decades, the field of hand gesture recognition has evolved substantially, transitioning from rudimentary vision-based techniques to sophisticated methods that integrate depth sensing and machine learning. Early works, such as those by Ibraheem 2012, laid the groundwork for understanding how gestures could be interpreted through visual means, a foundation upon which subsequent research has built more complex systems. By 2012, Ibraheem & Khan's exploration of various gesture recognition technologies indicated a burgeoning interest in diversifying approaches to improve recognition efficacy. The timeline reveals a clear trajectory toward increasing complexity and nuance in the recognition process, a trend that is emblematic of broader shifts in the field of computer vision and interactive systems.

In the mid-2010s, advances in 3D modeling and depth-sensing technologies, exemplified by Cheng et al. in 2016, permitted a more detailed analysis of hand movements, facilitating the recognition of gestures with greater precision. The advent of convolutional neural networks (CNNs) and other machine learning algorithms around this period marked a significant turn, as seen in the works of Gao et al. in 2017 and 2022, where these techniques began to outpace more traditional methods due to their ability to learn from vast amounts of data and to recognize subtle patterns not easily discernible by rule-based systems. The incorporation of machine learning not only improved accuracy but also allowed systems to better adapt to a variety of real-world environments, a critical aspect of usability.

As we approach the latest in the timeline, the emphasis on real-time processing and the application of gesture recognition to more diverse domains, such as media players by Rajbonshi et al. in 2022, highlight the maturing of the technology. Real-time recognition is vital for user-friendly interfaces, allowing for more seamless interaction between humans and computers. The scope of research has also expanded, aiming to integrate gesture recognition into everyday devices, thereby enhancing the Internet of Things and making technology more accessible and intuitive. As we look forward, the challenge will be not only to continue refining accuracy and responsiveness but also to address issues of user privacy and data security, ensuring that the benefits of gesture recognition technologies are enjoyed without compromising individual rights. The table reflects a narrative of technological innovation that underscores the vital interplay between theoretical development and practical application, a dynamic that will undoubtedly propel the field into new realms of possibility.

Meanwhile, the upcoming Table 3 explores the fundamental areas of static gesture recognition methodologies. The table is divided into four main columns: classification, feature extraction, segmentation, and scope. This structure will aid in assessing the approach of each technique towards gesture recognition, providing a comprehensive analysis of the unique attributes and wide range of applications within the area.

TABLE 3 Literature Review of Vision Based Static Gesture Recognition

References	Authors	Year	Classification	Feature Extraction	Segmentation	Scope
[155]	Gao et al.	2022	Dynamic Hand Gesture	3D Pose Estimation	Human-Robot Interaction	Human-Robot Interaction
[156]	Rajbonshi et al.	2022	Real-Time Hand Gesture	Static & Dynamic Gestures	Media Players	Media Players
[157]	Ding & Zheng	2022	RGB-D Depth-sensor-based	Deep Learning	Shadow Effect Removal	Smart Gesture Communication
[159]	Neethu et al.	2021	SVM-based	Distance Transform	Autonomous Vehicles	Autonomous Vehicle Applications
[160]	Krish et al.	2020	Machine Learning	Data Acquisition	Pre-processing	Human-Computer Interaction
[161]	Shah et al.	2019	Vision-based	Hand normalization	Manual placement	Hand Gesture Recognition
[162]	Jiang et al.	2018	Gesture Recognition	Depth Information	CNN	Natural Communication
[163]	Liu et al	2018	Real-Time Hand Gesture	Data Gloves	Vision-based	Hand Gesture Recognition System
[164]	Cheng et al.	2016	3D Hand Gesture	Hand modeling	Gesture recognition	Gesture Recognition
[165]	Gao et al.	2017	Static Hand Gesture	CNNs	Space HRI	Human-Computer Interaction
[166]	Xu et al.	2013	Sparse Representation	Kinect-based	Hand Gesture Recognition	Human-Computer Interaction
[167]	Raj et al.	2012	FPGA-based	Gesture recognition	Artificial Intelligence	Computing and Image Processing
[168]	Zhang & Yun	2010	Robust Gesture Recognition	Distance Distribution	Skin-color Segmentation	Gesture Recognition
[169]	Wu & Huang	1999	Vision-Based	Gesture recognition	Gesture recognition	Human-Computer Interaction
[170]	Pavlović et al.	1997	Visual interpretation	Gesture representation	Feature extraction	Human-Computer Interaction

TABLE 4 Literature Review of Vision Based Dynamic Gesture Recognition

Authors	Year	Classification	Feature Extraction	Segmentation	Scope
[171]	2023	Improved Gesture Segmentation Method for Gesture Recognition	CNN, YCbCr	Color Segmentation	Pattern recognition, signal processing
[172]	2022	Dynamic Gesture Contour Feature Extraction Method Using Residual Network Transfer Learning	Residual Network Transfer Learning	Contour Detection	Dynamic gesture category recognition
[173]	2021	Real-Time Recognition of Dynamic Finger Gestures Using a Data Glove	Deep Learning, Data Glove	Glove Sensor Data	Sign language translation
[174]	2021	Two-Stream CNN Framework for American Sign Language Recognition	CNN, Multimodal Data Fusion	Skin Detection & Depth Thresholding	ASL recognition
[175]	2020	Featured Based Segmentation Method for Building Millimeter Wave Radar Gesture Recognition Data Sets	Feature Extraction, Neural Networks	Radar Signal Processing	Effective dynamic gesture data sets
[176]	2020	Gesture Recognition Based on Depth Information and Convolutional Neural Network	Depth Information, Convolutional Neural Network	Depth Segmentation	Long-distance and non-contact interactions
[177]	2020	HAGR-D: A Novel Approach for Gesture Recognition with Depth Maps	Depth Maps, Image Processing	Depth-based Segmentation	Medical, games, sign language
[178]	2019	MEMS Accelerometer Based Nonspecific-User Hand Gesture Recognition	Sign Sequence, Template Matching	Motion Tracking	Gesture recognition models
[179]	2019	Fast and Robust Method for Dynamic Gesture Recognition Using Hermite Neural Network	Machine Vision, Hermite Neural Network	Motion Analysis	Gesture semantics determination
[180]	2018	Open Source Framework for Real-Time Hand Gesture Learning and Recognition	Hidden Markov Models	Shape Context Matching	Static and dynamic hand gesture recognition
[181]	2017	Real-time Recognition for Automotive Interfaces Using Multimodal Vision-Based Approach	Multimodal Vision-Based	Multi-sensor Fusion	Automotive Interfaces
[182]	2017	Visual interpretation of hand gestures for human-computer interaction: a review	Gesture representation	Manual Feature Labeling	Human-Computer Interaction
[183]	2016	Improved Gesture Segmentation Method for Gesture Recognition	CNN, YCbCr	Color Segmentation	Pattern recognition, signal processing
[184]	2015	Dynamic Gesture Contour Feature Extraction Method Using Residual Network Transfer Learning	Residual Network Transfer Learning	Contour Detection	Dynamic gesture category recognition
[185]	2015	Real-Time Recognition of Dynamic Finger Gestures Using a Data Glove	Deep Learning, Data Glove	Glove Sensor Data	Sign language translation
[186]	2015	Two-Stream CNN Framework for American Sign Language Recognition	CNN, Multimodal Data Fusion	Skin Detection & Depth Thresholding	ASL recognition
[187]	2013	Featured Based Segmentation Method for Building Millimeter Wave Radar Gesture Recognition Data Sets	Feature Extraction, Neural Networks	Radar Signal Processing	Effective dynamic gesture data sets

As it can be seen from this table, the compilation of vision-based hand gesture recognition research traces a fascinating evolution of the field, reflecting a shift from foundational principles to advanced computational methods. In the earliest work by Pavlović et al. (1997), the focus was on visual interpretation of hand gestures for human-computer interaction, laying the groundwork for future studies. This work concentrated on understanding how hand gestures could be translated into commands that computers could recognize, thus making technology more accessible and interactive. At this stage, the research was primarily about defining the problems and creating a vocabulary of gestures that could be universally understood in computational terms.

Progressing through the timeline, the field began to explore the three-dimensional aspects of hand gestures with Cheng et al. (2016), moving beyond static two-dimensional recognition to models that could understand and interpret gestures in the space around the user. This advancement was significant because it allowed for a more natural mode of interaction, where users could communicate with computers in a way that was more aligned with natural human behavior. It also presented new challenges, such as the need for more complex algorithms and processing power to manage the additional data from 3D space. With increased computational demands, the research began to lean more heavily on machine learning techniques to improve accuracy and efficiency.

In the most recent studies, like those by Gao et al. [150] and Ding and Zheng [151], the technology has advanced to dynamic hand gesture recognition using deep learning and 3D pose estimation, which marks a stark contrast from earlier approaches. These modern methods can provide more nuanced and sophisticated recognition capabilities, crucial for applications in human-robot interaction and smart gesture communication. Moreover, with the incorporation of depth-sensor-based technologies and the removal of shadows and other visual noise, the accuracy of gesture recognition systems has improved significantly. This progress highlights the rapid development of computational techniques and hardware that can support more advanced gesture recognition applications. The journey from hand-crafted features and gesture libraries to deep learning models that learn from data represents a pivotal shift in the field, where systems are now capable of learning and adapting to new gestures without explicit programming.

The discussion of this body of work illuminates a clear trajectory from simple recognition in controlled environments to sophisticated, real-time interaction in complex settings. The integration of multimodal data, advancements in sensor technology, and the application of cutting-edge machine learning algorithms have transformed vision-based hand gesture recognition into a dynamic field with significant implications for how humans interact with machines. As the technology continues to mature, future research will likely tackle the

remaining challenges of generalizability across diverse user populations and environments, as well as the seamless integration of these systems into a broader range of applications, from assistive technologies to immersive virtual reality experiences.

On the other hand, Table 3 is set to review the literature on vision-based dynamic gesture recognition. It categorizes the body of work into four principal sections: classification, feature extraction, segmentation, and scope as static table. The organization of the table aims to facilitate an in-depth examination of the different strategies utilized in the identification of static gestures, offering an insightful look at their distinct features and the extent of their implications in various applications.

As it can be seen from this table, the array of research from 2013 to 2023 on vision-based dynamic gesture recognition showcases remarkable progress and diversity in methodologies, reflecting the interdisciplinary nature of the field. Initially, methods were rooted in manual feature labeling and basic machine vision techniques, as depicted in Pavlović et al.'s 1997 seminal work, which focused on the visual interpretation of hand gestures for human-computer interaction. These efforts laid the groundwork for the integration of more sophisticated machine learning algorithms and sensor technologies.

By 2013, the field had advanced to incorporate featured-based segmentation methods using neural networks, indicative of the growing trend towards artificial intelligence in gesture recognition. The subsequent years saw a significant tilt towards multimodal and sensor-based approaches, especially with the use of CNNs for gesture recognition, such as the two-stream CNN framework for American Sign Language recognition in 2015 and 2019. These models leveraged multimodal data fusion and deep learning to interpret complex gesture dynamics more effectively.

The inclusion of technologies like millimeter-wave radar gesture recognition and depth cameras signified another leap, providing the means to detect gestures with greater precision and across various environments and applications. In particular, the use of FMCW radar by Wang et al. in 2020 and the novel gesture recognition techniques with depth maps by Santos et al. in 2015 underscore the push towards more seamless, real-time interaction capabilities. By 2021, the shift towards deep learning was more pronounced with the application of data gloves for real-time recognition of dynamic finger movements by Lee & Bae and the use of residual network transfer learning by Ma & Li for contour feature extraction, highlighting the field's trajectory towards greater integration of neural networks and wearable technology.

Future research will likely continue to push the boundaries of what's possible with vision-based hand gesture recognition, delving into areas such as three-dimensional modeling, gesture prediction, and the integration of haptic feedback to provide a more tactile experience. As these technologies mature, we can expect them to become increasingly integrated into our everyday lives, changing the way we interact with our digital environments.

SECTION VIII.
Comprehensive Finding from Research Questions

In this section, we discuss the core findings from our systematic review of gesture recognition technologies. Each research question posed at the outset of this study has been meticulously explored, drawing on a diverse array of studies to paint a holistic picture of the current state and future directions of gesture recognition. The following insights encapsulate the advancements in technology, the exploration of new methodologies, and the practical implications of our findings.

A. Categorization of Studies on Sign Language Recognition Findings

The field of sign language recognition has been explored through various technological approaches, primarily divided into three categories: computer vision techniques, sensor-based methods, and hybrid approaches that combine elements of the first two. Each category leverages different technologies and has its own strengths and challenges, addressing different aspects of sign language recognition.

Computer Vision Techniques: This category utilizes cameras to capture the motion and positioning of the hands and body. Techniques within this category have evolved significantly, with early studies focusing on simple gesture recognition using traditional image processing methods, and more recent work employing advanced machine learning algorithms such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). These methods are highly effective in environments with controlled lighting and background, but their accuracy can decrease in more dynamic settings.

Sensor-Based Methods: These involve the use of wearable devices equipped with various sensors such as accelerometers, gyroscopes, and even electromyography sensors to detect muscle activity. Sensor-based methods are less affected by external visual factors such as lighting and background but can be intrusive, as they require the user to wear specific equipment. Moreover, they provide a high degree of precision in gesture detection and are particularly useful in capturing fine motor movements that might be missed by visual-only methods.

Hybrid Methods: Hybrid approaches aim to combine the strengths of both computer vision and sensor-based technologies to enhance overall accuracy and robustness. For instance, a hybrid system might use a wearable device to capture precise movement data while also employing computer vision to contextualize gestures within a broader range of actions and interactions. This dual approach can significantly improve the system's ability to function reliably in a variety of real-world environments.

B. Elaboration in Preliminary Studies on Sign Language Recognition (SLR) Systems Findings

Preliminary studies in SLR systems predominantly focus on the technical aspects of sign language recognition, such as algorithm development, sensor accuracy, and machine learning techniques. These studies often delve into the specifics of system design, detailing the types of sensors used, the algorithms developed for gesture interpretation, and the overall system architecture as it can be seen from table 1. This technical focus is crucial as it lays the foundational knowledge required for advancing SLR technologies. For example, many papers detail the application of deep learning models like CNNs to interpret raw video data, achieving higher accuracy in gesture recognition through complex computational models.

However, while the technical depth of these studies is profound, there is often a gap when it comes to the application of these systems in real-world scenarios. Many preliminary studies do not extend their findings to practical applications or user interaction studies, which are essential for understanding the usability and effectiveness of these systems in everyday environments. This gap indicates a somewhat isolated approach to technology development, where the emphasis is on solving technical challenges without fully considering the end-user's needs and contexts.

Additionally, there is a notable lack of interdisciplinary integration in many of these studies. While they excel in their technical domain, incorporating insights from fields such as human-computer interaction, cognitive science, and linguistics could greatly enhance the development of more holistic and user-centered SLR systems. Such integration could lead to better understanding of the nuances of sign language, which is not only a series of hand movements but also includes facial expressions and body language, aspects that are often overlooked in purely technical studies.

C. Types of Movements Identified in Gesture Recognition Findings

The exploration of types of movements in gesture recognition systems is fundamental for enhancing the accuracy and utility of these technologies. Schoolers in this domain has identified a broad spectrum of movements, ranging from simple, static gestures to complex, dynamic sequences that involve multiple parts of the body. Initial studies often focused on basic hand gestures, such as swipes and taps, which are relatively easy to detect and interpret. These movements form the backbone of many early gesture recognition systems and are typically used in user interfaces for simple commands.

As the technology has progressed, more complex gestures have been incorporated into research studies. These include intricate finger movements and combinations of gestures that mimic natural human actions, such as sign language or even non-verbal communication cues like nods and shakes. The ability to accurately recognize these complex movements requires advanced algorithms and often the integration of multiple data sources, such as combining visual input from cameras with sensory data from wearable devices. This level of complexity not only challenges the computational models but also tests the limits of current sensing technologies.

Moreover, the most recent studies have ventured into full-body gesture recognition, which includes the detection of postural and locomotor elements such as walking patterns or body orientations. These movements are particularly relevant in fields like augmented reality and immersive gaming, where the user's entire body interacts with the system. Recognizing these types of movements involves a sophisticated understanding of human biomechanics as well as advanced technologies in computer vision and machine learning. The integration of these complex movements into recognition systems marks a significant evolution in the field, expanding the applicability and functionality of gesture-based interfaces.

D. Explored Sensors in Gesture Recognition Findings

The exploration of sensors in gesture recognition has been pivotal in advancing the field, with a diverse range of sensor technologies being implemented to enhance accuracy and responsiveness. Initially, research predominantly utilized basic optical sensors, which relied on visual input from cameras to detect and interpret gestures. These sensors were straightforward in capturing gross motor movements but often struggled with finer gestures or in low-light conditions. Over time, the focus shifted towards more sophisticated sensor technologies, including infrared, ultrasonic, and depth sensors, which could provide greater detail and accuracy, particularly in varying environmental conditions.

The introduction of wearable sensor technologies marked a significant evolution in gesture recognition. These devices, equipped with accelerometers, gyroscopes, and magnetometers, allowed for the capture of precise movement data directly from the user's body. This direct data collection not only improved the accuracy of gesture recognition but also enabled the detection of subtle movements and orientations that could not be easily captured by standalone cameras. Wearables have become increasingly popular in consumer electronics, such as smartwatches and fitness trackers, which utilize gesture recognition for user interface control and activity tracking.

In more recent studies, researchers have explored the use of hybrid sensor systems that combine multiple types of sensors to leverage the strengths of each. For example, combining data from wearable sensors with that from fixed environmental sensors allows systems to gain a more comprehensive understanding of the user's gestures within a specific context. This approach can significantly enhance system robustness, providing reliable performance across a variety of scenarios and conditions. Hybrid systems are particularly effective in applications where environmental factors play a significant role, such as in outdoor sports or in automotive contexts where the user's gestures must be recognized within a dynamically changing environment.

E. Examination of Vision-Based Methodologies in Gesture Recognition Findings

Vision-based methodologies have been at the forefront of gesture recognition technology, primarily due to their non-intrusive nature and the rich data they can capture. These methods utilize cameras to record and analyze movements, employing sophisticated image processing and machine learning algorithms to interpret gestures. Initially, these techniques focused on simpler, static gestures using traditional image processing methods like background subtraction and edge detection. As technology progressed, more advanced machine learning models, particularly deep learning techniques such as Convolutional Neural Networks (CNNs), have been employed to enhance recognition accuracy and adaptability.

The shift towards deep learning in vision-based gesture recognition has been transformative. CNNs, for instance, have enabled systems to learn and generalize from a vast amount of visual data, significantly improving their ability to recognize complex gestures under varied conditions. This is crucial because the effectiveness of vision-based methods often depends on their capability to handle diverse environments and lighting conditions. Deep learning models are particularly adept at this, as they can extract nuanced features from visual data that are not readily apparent to traditional algorithms. Moreover, the integration of Recurrent Neural Networks (RNNs) has furthered advancements, allowing systems to effectively handle sequences of movements, which are common in natural gestures and sign language.

Despite these advancements, vision-based methods face inherent challenges, such as occlusions (where the hand or part of the gesture is blocked from view), variability in human anatomy, and the complexities introduced by diverse backgrounds and lighting conditions. These issues necessitate ongoing research to refine algorithms that can more robustly handle such variations. Furthermore, the computational intensity of processing high-resolution video in real-time presents another significant challenge, necessitating efficient algorithm design and hardware acceleration techniques to enable smoother and more responsive gesture recognition.

F. Assessment of Hybrid Methodologies in Gesture Recognition Findings

Hybrid methodologies in gesture recognition represent an innovative approach that combines the strengths of various sensing and processing technologies to overcome the limitations inherent in single-method systems. By integrating data from both sensor-based and vision-based technologies, hybrid systems aim to achieve higher accuracy and reliability under diverse operating conditions. Sensor-based inputs, such as those from accelerometers, gyroscopes, or even EMG (electromyography) sensors, provide precise data on user movement dynamics, which is often less susceptible to environmental disturbances like poor lighting or occlusions that typically affect vision-based systems.

The application of computer vision in conjunction with these sensors adds a layer of contextual understanding that sensor-only systems might miss. For example, while sensors can precisely track the speed and direction of a hand movement, vision-based systems can interpret the gesture within the broader context of body language or the environment. This synergy allows hybrid systems to not only recognize gestures more accurately but also to understand their significance within a specific situation. This is particularly useful in complex interaction scenarios such as augmented reality (AR) or advanced driver-assistance systems (ADAS), where the interpretation of gestures can depend heavily on the environment.

Furthermore, hybrid methodologies are pushing the boundaries of what's possible in terms of gesture recognition's scope and scalability. By leveraging multiple types of data, these systems can adapt to a wider range of applications and environments, from indoor settings with controlled lighting to challenging outdoor environments. This adaptability makes hybrid systems particularly attractive for applications in mobile technology, healthcare, and automotive industries, where versatility and robustness are critical.

SECTION IX.
Substantial Analysis

After conducting a thorough analysis of the literature, various deficiencies and insufficiencies have been discovered, revealing areas that need additional focus and further research and these are the limitations of current studies.

A. Limitations of Existing Databases

The existing databases utilized in previous studies were found to be considerably limited, often containing only numbers, alphabets, or a meager set of words. This paucity poses a substantial challenge to researchers, primarily due to the inherent difficulty and expense associated with data collection. A critical need emerges for the development of more comprehensive and diverse databases to facilitate a more thorough investigation of SLR systems.

B. Focus on Isolated Signs

The primary emphasis of research in sign language recognition has primarily been on isolated signs, where users typically utilize one gesture at a time. Nevertheless, there is a significant deficiency in the research on continuous gesture recognition, as current endeavors are confined to only about 11 words. There is an urgent requirement for increased efforts in the development of dependable segmentation techniques and the expansion of datasets to support continuous systems, allowing performers to express whole phrases without interruption.

C. Dynamic Gesture Recognition Technology

A significant focus of both early and current studies is mostly on stationary gestures. Therefore, there is a clear and urgent requirement for the development of advanced gesture recognition technology to improve the efficiency of translation systems. The advancement of technology has the capacity to greatly enhance the precision and efficiency of sign language recognition in real-world scenarios.

D. Integration of Non-Manual Signs

Most sign language recognition systems now in use have a significant limitation in that they only concentrate on analyzing hand movements and ignore non-manual indications. To overcome this constraint, it is necessary to do pioneering research that integrates the identification of both physical and non-physical gestures, enabling the expression of supplementary levels of significance, such as exclamations, inquiries, and emotions. This offers a new and innovative opportunity for investigation in the field.

E. Affordability and User-Friendliness

The affordability of gesture recognition hardware, particularly gloves, emerges as a significant consideration. Efforts should be directed towards reducing the cost of such devices, making them accessible to the hearing impaired. Additionally, the design of gloves should not only align with their functional purpose but also prioritize user-friendliness, comfort, and flexibility to ensure widespread acceptance and use.

F. Mobile Gesture Recognition

As mobile phones are increasingly used as personal computing devices, improving the reliability of mobile gesture detection shows potential for addressing communication obstacles between the hearing impaired and the general public. Enhancing the ability to convert speech and text into gestures on smartphones through ongoing improvements could greatly enhance inclusivity and accessibility.

G. Hybrid Feature Extraction

Feature extraction methods play a crucial role in reducing dimensionality in raw data. However, there is a need for hybrid feature extraction methods to provide more robust features for recognition. This innovation could lead to improved accuracy and reliability in the recognition of sign language gestures.

H. Underutilization of Deep Learning

Despite the pivotal role of deep learning methods, particularly in handling large datasets, their application in sign language recognition has received limited attention in previous studies. There is a compelling need to explore and classify overall signs such as hand, head, facial expression by using deep learning techniques, which offer immense potential for advancing gesture recognition. The automatic learning and extraction of features, coupled with automated network hyperparameter fine-tuning algorithms, make deep learning an attractive avenue for exploration.

I. Scalability Challenges

A noteworthy observation from the systematic literature review is that most research papers, while achieving commendable accuracy rates exceeding 90%, primarily focus on a small number of gestures. To broaden the scope and applicability of sign language recognition systems, a rigorous investigation into their ability to recognize extensive datasets, surpassing 200 dynamic gestures, is warranted. To date, there is a notable absence of research examining accuracy rates at this scale, highlighting a critical research gap in the current landscape.

SECTION X.
Conclusion

The current landscape highlights a notable lack in persons' proficiency in using sign language to communicate with the hearing impaired. It is essential to rectify this deficiency in order to promote social engagement with this community. This research primarily investigates sign language recognition (SLR) within the field of hand gesture recognition. This study conducts a systematic analysis of the literature to thoroughly examine the current status, difficulties, reasons, and suggestions related to the recognition of sign language in the field of hand gestures.

The focus of our systematic literature review was to investigate several methods for recognizing sign language using vision-based, sensor-based, and hybrid-based approaches specifically for hand motions. Vision-based methods leverage visual information, sensor-based approaches acquire data from sensors embedded in gloves, capturing parameters like bend, hand orientation, and rotation. This sensor-based method demonstrates resilience to environmental conditions, ensuring more accurate data by mitigating factors such as performer location and background conditions. However, it is acknowledged that the sensor-based approach has its drawbacks, being perceived as cumbersome and bulky due to the requirement of wearing multiple boards and sensors for precise sign capture.

Nevertheless, the authors suggests that the trajectory of future hand gesture recognition studies should encompass the translation of non-manual signs, an aspect often overlooked in existing works. Moreover, expanding datasets to encompass a more extensive lexicon, particularly dynamic words, is identified as a

critical imperative. While prevailing models predominantly focus on isolated sign language recognition, a forthcoming change in perspective requires tackling the difficulties of ongoing SLR. To this end, the study underscores the significance of employing deep learning algorithms for sign classification, positing their potential to enhance the efficacy of gesture recognition.

A crucial recommendation put forth is the reduction in the size of hardware used in glove systems to enhance their conformability and mobility. This adjustment aligns with the evolving landscape of wearable technology, emphasizing the importance of unobtrusive and user-friendly designs. Lastly, the authors propose a future investigation into the trade-off between device robustness and sensitivity, acknowledging the need for striking an optimal balance to enhance the overall effectiveness and user experience of sign language recognition systems. These recommendations collectively chart a course for future research supports, aiming to bridge existing gaps and elevate the capabilities of sign language recognition technologies.

ACKNOWLEDGMENT

The authotrs extend their deepest appreciation to their research team for their wisdom and insightful feedback with stimulating discussions and critical insights that enriched their work. They acknowledge their commitment of everyone involved to advancing scientific knowledge, including the anonymous reviewers who scrutinized their manuscript, they express their gratitude for your constructive feedback leading to significant improvements.

Authors
Figures
References
Citations
Keywords
Metrics