Complex & Intelligent Systems (2024) 10:1581–1606
https://doi.org/10.1007/s40747-023-01173-6

SURVEY AND STATE OF THE ART

Computer vision-based hand gesture recognition for human-robot interaction: a review

Jing Qi[1,2] · Li Ma[1,2] · Zhenchao Cui[1,2] · Yushu Yu[3]

Received: 6 March 2023 / Accepted: 29 June 2023 / Published online: 19 July 2023

Abstract

As robots have become more pervasive in our daily life, natural human-robot interaction (HRI) has had a positive impact on the development of robotics. Thus, there has been growing interest in the development of vision-based hand gesture recognition for HRI to bridge human-robot barriers. The aim is for interaction with robots to be as natural as that between individuals. Accordingly, incorporating hand gestures in HRI is a significant research area. Hand gestures can provide natural, intuitive, and creative methods for communicating with robots. This paper provides an analysis of hand gesture recognition using both monocular cameras and RGB-D cameras for this purpose. Specifically, the main process of visual gesture recognition includes data acquisition, hand gesture detection and segmentation, feature extraction and gesture classification, which are discussed in this paper. Experimental evaluations are also reviewed. Furthermore, algorithms of hand gesture recognition for human-robot interaction are examined in this study. In addition, the advances required for improvement in the present hand gesture recognition systems, which can be applied for effective and efficient human-robot interaction, are discussed.

Keywords Hand gesture recognition · RGB-D camera · Human-robot interaction · Robot

Introduction

The exchange of information is the fundamental means through which individuals engage with the outside world. To explore their surroundings and learn about their environment, people rely on their sensory organs. People simultaneously communicate information to others around them through their motions, eyes, faces, gestures, and other physical cues. Human contact is created through the communication of information in various ways.

Gestures are an integral part of everyday human life. Vision-based gesture recognition is a technique that combines sophisticated perception with computer pattern recognition. It is used in many different sectors, including engineering and research, and is essential for enhancing human–machine interaction. Because natural gestures change constantly, the current gesture detection technology is unable to fully achieve genuine human–machine communication.

Human-computer interaction history is the progress of humans adapting to machines and machines continuously adjusting to humans. Each new type of human-computer interaction has resulted in the substantial changes in associated businesses. The advent of the computer mouse made computer operation more natural and contributed to the rapid spread of computers. And the debut of the iPhone based on touch screen technology considerably improved the user experience and influenced the cell phone sector. In recent years, there has been a boom of interest in natural human-computer interaction approaches such as facial recognition, human motion analysis, gesture recognition, etc. Gestures are a natural and intuitive means of human-computer communication that has emerged as a significant technology

✉ Yushu Yu
yushu.yu@bit.edu.cn

Jing Qi
jingqi@hbu.edu.cn

Li Ma
mali@stumail.hbu.edu.cn

Zhenchao Cui
cuizhenchao@gmail.com

[1] School of Cyber Security and Computer, Hebei University, Qiyi East Road, Baoding 071002, Hebei, China

[2] Machine Vision Engineering Research Center of Hebei Province, Hebei University, Qiyi East Road, Baoding 071002, Hebei, China

[3] School of Mechatronical Engineering, Beijing Institute of Technology, Zhongguancun South Street, Beijing 100081, China

in contemporary human-computer interaction. A successful implementation of gesture recognition will provide a revolutionary human-computer interaction experience.

Taxonomy of hand gesture recognition

Gesture recognition methods are divided into sensor-based and vision-based.

Sensor-based methods

Due to the range of sensors, the primary recognition may be grouped into three categories: using data gloves, using EMG signals and using Wi-Fi.

(1) Data gloves

With the widespread use of sensors, wearable sensor-based gesture recognition has advanced quickly. To recognize gestures using wearable technology, a glove with numerous sensors must be worn on a hand, and the data from the glove must be analyzed. In particular, gesture recognition based on data gloves may more intuitively gather three-dimensional spatial information from hand posture by utilizing many sensors and is not limited by the surrounding environment.

A sensing glove was proposed by Komura and Lam [1] for controlling 3D game characters. Kim et al. [2] suggested a data glove-based sign language recognition system and achieved a 99.26% motion detection rate and an approximately 98% finger state recognition rate. Using data gloves, Rodriguez et al. [3] investigated the use of gestures for human-computer interaction (HCI) in virtual reality (VR) applications. Individuals with hearing and speech impairments can readily wear and use the gloves that Helen Jenefa and Gokulakrishnan [4] made for people with speech impairments. Such a glove is equipped with bending sensors, accelerometer sensors, and touch sensors that measure the bending and movement of the user and enable nonmute persons to comprehend the gestures produced by a speech-impaired individual. To recognize several motions, including move-ready, grip, loose, landing, takeoff, and hover motions, as well as to enable remote control of a six-axis vehicle, Huang et al. [5] created a data glove. The researchers achieved an overall recognition rate of 84.3%. Using five fundamental classification algorithms—decision trees, support vector machines, logistic regression, Gaussian naive Bayes, and a multilayer perceptron—Antillon et al. [6] created a smart diving glove that was trained and tested. Additionally, a study was performed underwater to determine whether the environment had any impact on how each algorithm classified objects. Mummadi et al. [7] demonstrated a data glove prototype with a glove-embedded gesture classifier that used information from inertial measurement units on the user's fingertip.

(2) Electromyography (EMG)

Using electrodes affixed to the skin or injected into the muscles, electromyography records the electrical activity of the muscle tissues. This technique primarily uses sensors to gather electrical signals from the skin and muscles on the surface of a human body. After expanding the signal data and processing it further, the method screens the information that might be contained in each gesture before recognizing gestures.

In 2002, Vuskovic and Du [8] used two-channel sEMG signals simultaneously to identify six different gestures with an accuracy of 78%. Nazarpour et al. [9] performed feature extraction using high-order statistics in 2005, and correctly identified 4 forearm movements using a clustering method with an accuracy of 91%. In 2018, Wu et al. [10] considered 15 features, including integral electromyographic data, to recognize five hand motions with an average accuracy of 90% using upgraded k-nearest neighbor algorithms. The maximum accuracy of the random forest neural network and the best accuracy of the support vector machine (SVM) classifier, were 88.7% and 85.9%, respectively, in 2015, when Guo et al. [11] used four different types of data as feature input. To identify four different types of gesture motions, Kim et al. [12] considered power spectral density as feature input and chose the SVM classifier with the accuracy of 91.97%.

(3) Wi-Fi and radar

Due to the growing use of Wi-Fi devices in indoor settings, gesture recognition technology based on Wi-Fi signals has drawn increasing attention. Wi-Fi devices are now present in many different indoor settings.

Wi-Fi signals were first used for sensing in 2000 by Bahl and Padmanabhan [13], who suggested a system for indoor localisation based on the received signal strength of such signals. In 2013, Pu et al. [14] proposed Wisee, which used Doppler shift as a feature for gesture recognition to identify nine gestures with an accuracy of 94% for gestures such as pushing and pulling gestures. Wisee, however, used specialized software defined on radio devices and could not be immediately implemented on the existing Wi-Fi devices. The method required a distance of more than 10 cm between the two antennas to obtain good results, although Mudra [15] developed a Wi-Fi-based finger-level gesture detection method with a 96% accuracy by utilizing the difference in signals between antennas in different locations. WiFall [16] applied the random forest method and an SVM classifier to categorize various human activities and implement fall detection , yielding an average false alarm rate of 18% and a detection accuracy of 87%. To address the issue of signal propagation through walls, Wu et al. [17] suggested a passive human activity identification system based on Wi-Fi signals without any extra equipment.

Users of sensor-based typically must wear gloves that have sensors or probes attached to users' arms. Additionally, the methods used in a laboratory setting are frequently constrained by the instruments that must be set up before recognition.

Computer vision-based methods

Image sensor technology has undergone continuous updating and iteration since its inception. Due to the inability of 2D-based image sensors to provide the additional information required to fulfill the needs of the contemporary society, the interest of academics in such sensors is currently declining, and the AI internet-of-things (IoT) field is shifting toward 3D. Monocular, binocular, and depth (RGB-D) cameras are the three main types of cameras used in vision-based gesture detection systems.

Microsoft's Kinect V1 (the first generation of Kinect) is a depth camera that was first unveiled on June 14, 2010, combining OpenNI and the SDK library to monitor the bones of human joints as a foundation for gesture recognition research. A dynamic Arabic sign language recognition system for Kinect was introduced by Hisham and Hamouda [18]. It combined decision trees with Bayesian classifiers for gesture identification and then used the AdaBoost method to improve the system, resulting in a recognition rate of approximately 93.7%.

Leap Motion, a body controller manufacturer focused on the PC and Mac platforms, introduced its body controller on February 27, 2013, utilizing the stereo vision principle and two cameras to determine coordinates of spatial objects similarly to the human eye.

RealSense cameras from Intel are also depth cameras with gesture recognition capabilities. Extracting valid descriptors from the hand skeleton's connected joints returned by the Intel RealSense depth camera, De Smedt et al. [19] proposed a skeleton-based 3D gesture recognition method. The Fisher vector obtained by using a Gaussian mixture model represented the encoding of each descriptor used to obtain the final feature vector. The original data were not filtered, and SVM was the only classifier used, leading to a comparatively poor recognition rate.

Gesture recognition processes

Static gesture recognition and dynamic gesture recognition are two categories of gesture recognition technologies [20]. The former implies that the hand is fixed for recognition and that aspects such as hand posture, shape, and location do not change [21]. Dynamic gestures are composed of sequential frames of static gestures, implying that the latter are a subset of dynamic gestures [22].

The main process of visual gesture recognition is as follows:

- Data acquisition: acquiring gesture images with video camera and preprocessing images;
- Gesture detection and segmentation: detecting the position of the hand in the gesture image and segmenting the hand region;
- Gesture recognition: extracting image features from the hand region and recognizing the gesture type based on the features. In "Hand gesture recognition process", the discussion will be divided into the two respective parts.

"Hand gesture recognition process" will thoroughly analyze and define the essential tactics involved in gesture recognition, using the basic steps of gesture recognition as the main consideration. "Experimental Evaluation" will present several evaluation metrics for gesture recognition and segmentation. With the emergence of depth cameras, there has been significant growth in studies of gesture recognition based on depth data; "Hand gesture recognition based on RGB-D cameras" will detail the research of various scholars on gesture recognition based on RGB-D cameras. The applications of gesture recognition, particularly in robotics and human-computer interaction, will be described in "Hand gesture recognition applications". "Problems, outlook, and conclusion" will discuss the current difficulties and the future directions in gesture recognition.

Hand gesture recognition process

Designed to enable the information exchange between users and intelligent devices, vision-based gesture recognition technology refers to the acquisition of video images containing operation gestures through acquisition devices such as cameras and the corresponding processing of video images, such as gesture segmentation, gesture feature extraction, gesture feature classification, etc.

Data acquisition

The data collected for vision-based gesture recognition is an image frame. The image needs to be preprocessed because it cannot be recognized directly after being acquired by the camera. To improve the overall performance of the system, the image preprocessing stage modifies the input image or video. The following are some common preprocessed steps.

Image grayscaling

Image grayscaling is the conversion of color images to grayscale images for better processing of image information. Ĉadik et al. [23] and Benedetti et al. [24] reviewed the research on grayscaling of color images. In grayscaling, the value of each pixel in an image is calculated from the values of its red, green, and blue channels by a specific algorithm to obtain a gray value that represents the luminance of that pixel [25]. The purpose of grayscaling is to simplify image processing and reduce computational and storage requirements. Common image grayscaling algorithms include the average method, the weighted average method, the maximum value method, and the minimum value method. In general, the average method and the weighted average method are used more often because they are simpler and more effective.

Image smoothing

Noise is removed from images using image smoothing techniques (e.g., Gaussian filtering, median filtering, etc.) to extract gesture information more accurately [26]. Gaussian filtering is a smoothing method based on a Gaussian function that can effectively smooth an image while preserving edge information. Such filtering involves convolving the image and replacing the value of each pixel with the weighted average of pixels in the region around that pixel. The kernel size and standard deviation of Gaussian filtering determines the degree of image smoothing; in gesture preprocessing, the appropriate Gaussian kernel size and standard deviation are usually chosen to achieve the best smoothing effect. Median filtering, on the other hand, is a smoothing method based on ranking statistics that can eliminate outliers such as pretzel noise while preserving the details of the image. The basic idea is to replace the value of a pixel with the median of its neighbors. Certain objects with hue and saturation characteristics similar to those of skin can produce pretzel noise in images generated by skin region detection, and the noise can be suppressed using median filtering and morphological methods [27]. In [28, 29], researchers used median filtering to process images.

Edge detection

Edge detection algorithms (e.g., the Canny algorithm, the Sobel algorithm, etc.) are used to detect edge information in an image for better segmentation of gestures. It is a process of identifying and locating points of sharp discontinuities in an image that represent the boundary between an object and the background or between adjacent objects [30, 31]. The Canny algorithm detects edges by calculating the gradient and orientation of pixel points in an image, and its main features are high accuracy, low noise sensitivity, and clear details of the detected edges [30]. The advantage of this algorithm is that it can effectively remove noise from an image and extract clear and continuous edges. The disadvantage is that it is computationally intensive, requires multiple filtering and processing operations, and is highly complex. The Sobel algorithm is similar to the Canny algorithm in that it detects edges by calculating the gradients in the x and y directions of each pixel and has the advantage of being computationally simple, fast, and capable of detecting fine edges [32]. The disadvantage is that it is not sufficiently accurate to detect straight edges, and it easily generates noise. Therefore, the image is usually smoothed using methods such as Gaussian filtering in the early stages of edge detection to reduce the effect of noise.

Morphological image processing

Morphological operations (e.g., expansion, erosion, etc.) are used to morphologically process an image to better extract the shape information of a gesture [26]. Commonly used morphological operations include expansion, erosion, open operation, and close operation. The expansion operation can expand the target region in the image to make it more visible, which is suitable for extracting gesture edge information. The erosion operation can shrink the target area in the image, which is facilitates removing noise and small details in the image. The open and close operations can remove burrs and holes in the image, respectively, to make the shape of the gesture clearer.

Optimum thresholding

Threshold segmentation algorithms (such as the Otsu algorithm, the Niblack algorithm, etc.) are used to divide an image into two parts, foreground and background, to better segment the gestures [33]. The Otsu algorithm is a global threshold segmentation algorithm; its basic idea is to divide the pixel gray values of an image into two classes such that the sum of the variances of the two classes is minimized [34]. This algorithm can adaptively determine the threshold and is hence suitable for image segmentation tasks in various scenes. The Niblack algorithm is a local binarization algorithm that divides the image into several small regions and then binarizes each region; it uses gray value thresholding to decide whether each pixel belongs to the foreground or the background [35]. This algorithm is more suitable than the global thresholding algorithm for segmenting images with highly diverse gray value distributions and can effectively handle images with uneven illumination and complex backgrounds.

In general, an input image is first thresholded as a binary image, and then noise is subtracted using the median and Gaussian filters; a preprocessing stage using morphological operations follows.

Gesture detection and segmentation

The gesture must first be separated from the background for the computer to recognize it. The reason is that the computer records the details of both the gesture and the scene in which it occurs. Hand segmentation is the division of the collection of pixel point coordinates obtained in the earlier gesture detection phase, which thus reduces the computation of pixel points and facilitates the subsequent operations. The first crucial stage in the algorithm for recognizing gestures is gesture segmentation, and a successful completion of this stage is necessary for accurate gesture identification. There are many gesture segmentation techniques, but from the practical and application point of view, almost all of them still face great difficulties in terms of accuracy, stability and speed (such as in the case of gesture segmentation in complex backgrounds), the impact of the distance between the camera and the person on gesture segmentation, etc. A comparison of gesture detection and segmentation methods is shown in Table 1.

Skin color segmentation

The most fundamental apparent characteristic of a human hand is skin color. Even though everyone has a unique skin tone, the skin tones of the human body are concentrated in a certain region of a particular color space. In addition, the orientation, size, and perspective of an image itself have very little impact on skin color, which is highly invariant in terms of rotation, translation, and scale reduction. Hence, a significant portion of current studies of gesture recognition rely on skin tone information for gesture segmentation. The three most commonly used color spaces are the RGB, HSV, and YCbCr color systems..

In [36], the hand skin tone was segmented using the threshold method. In [37], a skin color detection method was used to detect hands and faces. A skin tone model was utilized in [38] for segmentation, and the HSV color model was selected after a comparison with the RGB model because of the influence of luminance. A simple segmentation technique based on calculating the maximum and minimum skin probabilities of the input RGB image was used in [39]. In [40], a skin detection model and an approximate median model were applied to segment the image. The approximate median model was utilized for background subtraction, and the skin detection model was used to identify the hands and fingers in the image. At the same time, without relying on any artificial neural network training, Dhule determined the precise sequence of moving hands and fingers by calculating the change in RBG color pixel values in a video and controlled the mouse movement in a window in real time according to the hand and finger movements. A gesture recognition scheme based on the skin color model approach and the threshold approach combined with effective template matching using the principal component analysis was proposed in [41]. Veluchamy et al. [42] used a skin color thresholding model for segmentation; numerous characteristics were extracted using the scale-invariant feature transform and monogenic binary coding algorithms before being identified using an efficient classifier. In [43], the problem of segmentation of the hands involved in gesture production was solved differently, using a ribbon-based segmentation algorithm, the first using special color stickers for the fingers, and the second based on normal skin color segmentation. Wang et al. [44] segmented the hand region by locating the skin tone region in the CbCr plane of the YCbCr color space using the “elliptical boundary model”. Considering the YCbCr color space, Patel [45] noted that the hand region could be cropped from all the images in the dataset by thresholding segmentation. The skin color detection algorithm used in [46] facilitated communication between the user and the computer.

Contour information segmentation

Another crucial component of gesture segmentation is detection of the presence of contours and edges. A new technological advance in gesture segmentation has been provided by target segmentation techniques based on contour information. Edge detection operators, template matching, and active contour models are the three primary categories of conventional gesture segmentation techniques based on contour information.

Traditionally, gesture contours have been extracted from photos by using edge detection operators to identify edges in images. In the context of gesture segmentation, template matching—a traditional target localization technique—has also been applied to some extent. To discover the best match, template matching requires placing a preset template on a point in the image, calculating how well the template matches the image at that point, and then iteratively moving through the entire image.

A segmentation of gestures’ grayscale images was performed in [47] using a histogram thresholding segmentation technique. A morphological filtering technique was created to represent the gesture contours to successfully filter out the background and target noise from the segmented image. The wrist cropping approach in [48] used a width and contour heuristic to estimate the wrist position among the segmented hand patches and then extracted the hand from the estimated wrist position to separate the segmented hand from the rest of the arm. The nodes of the human body were identified in [49] by using background subtraction, edge detection, contour detection, and edge detection using the Sobel operator. Generating a population gesture feature collection for multi-view gesture photos was proposed in [50], along with a new

Table 1 Comparison of gesture detection and segmentation methods

Methods	Representative Algorithms	Advantages	Disadvantages
Based on skin color	Color space	Fast processing, invariance to rotation, partial occlusion, pose change	Susceptible to interference from skin-like areas
	Edge detection operator	Fast and accurate gesture edge information extraction	Edge extraction results may be broken, overlapping, etc., and require subsequent processing
Based on contour information	Template matching	Adaptability to different shapes and sizes in gesture segmentation; for relatively simple gestures, the matching accuracy is high and the segmentation results are more accurate	The need to prepare many templates in advance increases the system complexity
	Active contour model	Adaptive adjustment of contour shape, suitable for gesture segmentation of various shapes. Better for gesture segmentation in the presence of noise, complex backgrounds, etc.	High algorithmic complexity and significant demand for computational resources
Based on a depth sensor	–	High identification efficiency	Reliance on depth cameras and the need for improved accuracy
Based on deep learning	–	No requirement of manual analysis of gesture data for segmentation, which is thus more convenient and robust	Need to improve the real-time performance of detection
Gesture tracking (dynamic gestures)	Frame differential method	The algorithm's simplicity to implement, and the low programming. Lower sensitivity to scene changes such as light, ability to adapt to various dynamic environments, and relative robustness	Inability to extract the complete area of an object, and presence of "holes" inside the object; extraction of only the boundary with an outline that is coarse and often larger than the actual object
	Background subtraction	Ability to extract the complete area of an object, reduced sensitivity to changes in the scene such as light variations, and ability to adapt to various dynamic environments	Required initial background modeling and sensitivity to light changes
	Optical flow	Ability to extract the complete area of an object, and insensitivity to light changes	Inadequate detection effectiveness in the case of fast movement or unclear object surface texture
	MeanShift algorithm	Good real-time performance, fast calculations	Ineffectiveness of tracking in cases of large variations of the shape and size of the target
	CAMShift algorithm	Better tracking for large variations in target shape and size	The calculations is large and takes a long time to calculate
	Particle filtering algorithm	Fast calculations and low storage requirements	Ease of mismatching and losing the detailed features if the target is similar to the background

Pareto optimum frontier-based multiview gesture detection method.

Of course, there are other methods of gesture segmentation based on appearance features apart from skin color and contour. The shape and direction of the hand, taken from an input video stream captured under stable lighting and simple background conditions, were proposed in [51] as a basis for recognition of static gesture images. In [52], an adaptive threshold binarization-based homomorphic filter was used in the construction of a system resistant to changing lighting conditions. Additionally, edge-based grayscale segmentation ensured that the method could be applied to users with a range of skin tones and backgrounds. In [53], a hand detection method that incorporated skin filtering and three-frame differencing was proposed.

Other segmentation approaches

The depth sensor-based gesture segmentation method uses a depth camera to gather hand-depth structure information and then segment the hand region, as detailed in "Hand gesture recognition based on RGB-D cameras".

Gesture segmentation methods mentioned above require manual segmentation by designing features of target images, which can achieve gesture segmentation by creating features in simple contexts; however, it is difficult to design effective gesture features in complex environments, and thus this approach is difficult to apply in natural human-computer interaction systems. Development of deep learning opens up new options for gesture segmentation since a model can be trained using vast amounts of data to automatically learn target gesture attributes, allowing it to complete target gesture recognition and gesture segmentation using the identified target gestures [54]. Paul et al. [55] offered a method for extending the hand segmentation approach based on convolutional neural networks from still photos to video images. The proposed technique was more resistant to distortion and occlusion issues, resulting in improved accuracy and delay tradeoffs. Compared to traditional approaches, the deep learning-based gesture segmentation method eliminates the need for manual analysis of gesture data for segmentation, making segmentation more convenient and being a promising approach to gesture segmentation [56, 57]. However, in terms of the current development status, this method has flaws: first, some network layers are complicated, slowing gesture segmentation; second, edge detection may yield blurred results, and its accuracy still needs to be improved [56, 57].

Multiple methods have indeed been used to acquire hand segmentation images. For example, a single Gaussian model was applied in [58] to describe the hand color in the HSV color space. To achieve reliable hand tracking, Ding and Su [58] combined optical flow and color cues. Zhao and Jia [59] proposed a manual segmentation technique for depth images based on a random decision forest architecture.

Tracking

In this paper, tracking is considered a component of segmentation because the goal of both tracking and segmentation is to separate the hand from the background. The frame-by-frame analysis of temporally continuous images and the determination of the tracked target during the image change interval are the fundamentals of gesture tracking.

Fukunaga proposed the MeanShift algorithm in 1975; its basic idea is to use the gradient ascent of probability density to find the local optimum [60]. It is a straightforward algorithm with excellent real-time performance. If the target size changes, however, it is prone to tracking drift [61]. Khan et al. [62] combined the spatial information of moving targets with the traditional MeanShift algorithm to effectively solve the problem of tracking effectiveness degradation caused by the moving target's occlusion.

The CAMShift algorithm is a modification of the MeanShift algorithm, and its full name is "Continuously Adaptive MeanShift". Its basic idea is that all frames of video images are used for the MeanShift operation, and the result of the previous frame (i.e., the search window's center and size) is used as the initial value of the search window of the next frame of the MeanShift algorithm, etc., iteratively [63]. It can adapt to target deformation by changing the window size adaptively, but if the surrounding environment is complex, the tracking window will diverge, and the target will be lost [64]. Several studies have used the CAMShift method to track the position of gestures; examples are the applications in [65, 66] to detect and track gestures. The CAMShift method detects the position of gestures by continuously resizing the search window.

Motion detection is a popular technique for segmenting dynamic targets. Its fundamental principle is to fully localize and extract currently moving targets by combining the visual information of previous moments in a video. The temporal difference approach [53, 67], the background subtraction method [68, 69] and the optical flow method [70, 71] are the three primary categories of conventional gesture segmentation techniques based on motion information.

The temporal difference method's primary premise is to choose a number of adjacent frames in a video sequence, execute the difference operation, and then extract the moving target by separating it from the backdrop using a predetermined threshold. The two frames before and after the image's pixels are subtracted from one another. If the impact quality difference is negligibly different from the environmental brightness, the item can be assumed to be stationary. However, if a significant change in pixel values occurs anywhere in the image region, the change is assumed to be the result of a moving object. If a moving object was previously stationary or blocked by another object, the temporal difference approach and its upgraded algorithm frequently lose object information. Additionally, because the temporal difference method assumes the image backdrop's invariance, it is inappropriate if the background is moving. In [72], the issue of the frame difference pairs between the stored gestures and the query gestures used for matching was resolved. The hand trajectory following the center was monitored according to the direction between consecutive frames and the distance from the center of the frame.

The background subtraction method's fundamental concept is similar to that of the temporal difference method in that the input image is compared to the background model, and the moving target is segmented by looking for changes in either statistical information such as histograms or features such as a grayscale representation or changes in the histogram. Prior to storing the backdrop image, a background model is constructed. The input picture is subtracted from the background image, and a certain threshold T is used to decide which pixel belongs to the foreground target when the current frame is subtracted from the background image based on exceeding that threshold. After thresholding, the background picture is represented by the point in the image with a value of 0, and the pixel point in motion in the scene is represented by the point with a value of 1. If the entire backdrop image is available, this approach can capture objects more effectively with good real-time performance. However, the method is less reliable, and the outcome is significantly affected by changes in the dynamic scene. Simple background subtraction was avoided in [73] due to the complex background and potential dynamic hand movements. Instead, certain morphological approaches and two-stage skin color identification were applied to reduce noise. The method suggested in [74] used background subtraction techniques and skin color-based schemes to identify palms in video feeds, exploring the potential for a usable computer vision framework for gesture detection. To overcome the restriction on gesture input caused by gesture background subtraction, hand, and rotation invariance, a new effective recognition elimination method was provided in [75].

An optical flow field is a velocity field that depicts three-dimensional movements of object points via a two-dimensional map. Optical flow is defined based on pixel points that indicate picture changes caused by motion over the time interval. The pixel regions that best match the motion model can be found using the optical flow method and the regions can then be combined to create moving objects for object detection. The optical flow method can be used to detect the object independently without the need for additional camera data, but in most cases, this method is more time-consuming than desired, computationally complex, and susceptible to noise. Real-time detection can only be accomplished using specialized hardware support, meaning that the optical flow method has a significant time overhead and poor real-time performance and is impractical. In a related study, the hand was separated from the background by using the inverse projection of the color model and motion cues in [76]. Ganokratanaa and Pumrin [77] proposed a dynamic gesture identification system for older individuals that tracked six dynamic movements and categorized their meanings using optical flow and speckle analysis used in vision-based gesture recognition.

For visual tracking, the particle filter method is commonly applied. It is a sequential Monte Carlo significance-sampling method for estimating the latent state variables of a dynamic system from a series of observations. Particle filtering is commonly used in conjunction with other techniques for gesture tracking. The combination of particle filtering and the Mean-Shift algorithm was shown in [78, 79] to accurately identify hands. In the study of [80], the Kalman particle filter was introduced as an improvement to particle filtering in gesture tracking. The Kalman filter was also applied for gesture tracking in [81, 82].

Feature Extraction

Gesture feature extraction is the key step in gesture recognition. The feature extraction part involves processing the input gesture image and then extracting the features that can represent the gesture from the image. Gesture features are the features that can characterize the gesture form and motion state extracted through the analysis of hand movements and postures [83]. The selection and design of gesture features are related not only to the accuracy of gesture recognition but also to the complexity and real-time performance of the system. Gesture features are mainly divided into global and local features.

(1) Global features

Global features are features that describe the morphology and movement of the entire hand. They include the size, shape, direction, speed, acceleration, rotation angle, etc., of the hand. These features can describe the overall action state of the hand and are applicable to some simple gesture recognition tasks. Global features also usually include color histograms, grayscale histograms, Gabor filters, etc. A color histogram refers to the statistics of the occurrence frequency of various colors in an image, represented in the form of a histogram. A gray histogram is a count of the frequency of occurrence of each gray level in an image, represented in the form of a histogram. A Gabor filter is a filter capable of extracting image texture information; it detects the texture direction and frequency in an image and represents it in the form of a feature vector [84]. Global features are simple to compute and have intuitive representations and good invariance properties. However, such features mostly have pixel point-based feature representations, and hence there are problems such as high feature dimensionality and large computational effort. In addition, such feature descriptions are inapplicable in the case of image blending and occlusion.

(2) Local features

Local features are features extracted by analyzing local areas such as fingers, palms, and wrists. They include the curvature of fingers, the degree of palm protrusion, the rotation angle of the wrist, etc. These features can more accurately characterize the detailed information about the hand and are more applicable to gesture tasks that require high-precision recognition. Commonly used local features include the histogram of the oriented gradient (HOG), the local binary pattern (LBP), the shift-invariant feature transform (SIFT), the sped-up robust features

(SURF), features derived from the principal component analysis (PCA), and linear discriminant analysis (LDA), etc.

HOG is a feature descriptor proposed by Navneet Dalal and Bill Triggs in 2005 [85]. It is used as a feature descriptor for target detection and is a statistical value used to compute the directional information of local image gradients in computer vision and image processing [85, 86].

LBP is a feature extraction method applied in image processing and computer vision. It converts a pixel point into binary code by comparing the magnitude of the pixel's gray value with that of the surrounding pixels to obtain a feature that describes the texture of the image. LBP features have significant advantages such as grayscale invariance and rotation invariance [86, 87]

SIFT is a scale- and rotation-invariant feature extraction technique proposed by Lowe [88], which has been widely used in gesture recognition. It is a feature detection and description algorithm for image processing that can detect and describe feature points in images at different scales and rotation angles. SURF is a descriptor developed from SIFT that improves the speed and robustness of feature detection and description by using techniques such as integral image and fast Hessian matrix computation. SURF has the absolute advantage of being computationally fast compared to SIFT [89]. Sykora et al. [90] applied a support vector machine classifier to classify SIFT and SURF features extracted from 500 test images with recognition rates of 81.2% and 82.8%, respectively.

PCA is a commonly used data dimensionality reduction and feature extraction method that converts high-dimensional data into low-dimensional data while preserving as much information as possible about the original data [83]. LDA is a common classification algorithm and feature extraction method that converts high-dimensional data into low-dimensional data while maximizing the variability between different categories and minimizing the variability within each category [91].

In the study of gesture image features, global features have difficulty extracting the information of interest within the hand region due to the small proportion of the gesture region in the image, and therefore, the image performance is poor. In contrast, local image features are numerous and stable, have low interfeature correlation compared with that of the global features, can avoid occlusion of the hand region to some extent and are robust to image transformations such as illumination, rotation, and viewpoint change. Therefore, to enrich gesture feature information, most researchers tend to fuse local features with global features to achieve higher recognition rates. In this paper, we summarize the gesture features and the accuracy rates mentioned in several recent studies; the results are shown in Table 2.

Table 2 Partial gesture feature extraction techniques

References	Features	Accuracy
[92]	LBP, PCA	99.97
[93]	Harr-like features	95 37
[94]	SIFT	99
[95]	SURF	63
[96]	Fused features consisting of blended Hu moments, finger angle counts, skin tone angles, and nonskin tone angles	90.0
[97]	SIFT, Hu moments, LBP	87.3,85.1
[98]	Distance and angle from the end point of the hand	92.13
[99]	SURF	84.6
[100]	SURF, longest common subsequence	93.0
[101]]	Skin detector	97
[102]	Harris	94.8
[103]	SIFT	90
[104]	HOG, SIFT	91
[105]	PCA	91.5

Feature elimination and selection is an important step in machine learning that aims to select the most useful features for a classification or regression task to improve the accuracy and generalizability of the model.

- Variance filtering: Eliminate features with variance below a certain threshold because they have less impact on the classification or regression task.
- Correlation filtering: Eliminate features that have a low correlation with the target variable.
- Regularization method: Eliminate features by making the weights of some features converge to zero through L1 or L2 regularization.

Features can be selected by the following methods:

- Filtering: Evaluate each feature according to dispersion or relevance, set a threshold or the number of thresholds for features to be selected, and select features.
- Wrapper: Select a number of features or exclude a number of features each time according to the objective function until the best subset is selected.
- Embedding: First, use machine learning algorithms and models for training to obtain the weight coefficients of each feature, according to the coefficient from the largest to the smallest selection of features. This approach is

similar to the Filtering method, but training is used to determine the utility of features.

In traditional methods, after extracting multiple features such as HOG, Harr, skin color, etc., classifiers such as SVM are used to segment or recognize gestures, which involves feature elimination and selection. In deep learning methods, multiple branches can be used to extract different features; for example, one branch can be used to extract the motion trajectory of the gesture, and another can be used to extract the color information of the gesture. Then, these branches can be used accordingly to obtain the final gesture features, which also involves feature elimination and selection.

Gesture classification

Following the acquisition of the segmented image, important information from the image is extracted via feature extraction, and the gesture type is recognized using these features. Gesture classification is the classification of the extracted spatiotemporal features of gestures and is the last stage of gesture recognition. The main methods of gesture classification are listed and compared in Table 3.

Template matching

The first suggested recognition technique was a very simple template matching technique, usually used for static gesture recognition. The approach involves classifying an input image in accordance with how closely it matches a template (a point, a curve, or a shape). To calculate the matching degree, one can use the coordinate distance, the point set distance, contour edge matching, elastic map matching, etc. Although the classification accuracy is not very high and the types of gestures that can be recognized are limited, the template matching method has the advantages of being very quick in the case of small samples, being adaptable to lighting and background changes, and having a wide range of applications.

Similarity of gestures was determined in [47] by assessing the similarity of sequences of local gesture outlines. Bhame et al. [39] used variable distance features and straightforward logic to calculate the active fingers involved in a gesture, which sped up recognition and made the technique suitable for real-time human-computer interaction applications in contrast to the conventional method of extracting input features and comparing them with all database features. In [74], an iterative polygonal shape approximation technique was proposed and combined with a unique chain coding scheme for shape similarity matching to recognize gestures. RGB and depth descriptors were combined to classify the movements in [106]. Chaudhary and Raheja [107] argued that lighting inconsistencies and backdrop irregularities had an impact on image segmentation, and the scholars provided a method for recognizing gestures based on constant light intensity. The system was tested using the Euclidean distance approach and artificial neural networks, and the method relied on a gesture image database to match the test movements.

Methods based on geometric information

Gestures can also be recognized using geometric information, e.g., by fingertip detection, convex packet detection, the circle drawing method, the cross-hatch method, etc.

Depending on the application, fingertip detection can be separated into single-fingertip detection and multiple fingertips' detection. The "distance from the center of gravity" method can be used to detect a single fingertip. This method involves finding the point in the hand area that is furthest from the center of gravity and then determining whether the point is the fingertip. If the distance from the point to the center of gravity is greater than 1.6 times the average distance from the edge to the center of gravity, the point is the fingertip; otherwise, it is not.

Wen and Niu [108] suggested a fingertip angle calculation method to detect the fingers of the hand after discovering that the fingertip angle values of the turning points of the curve were much larger than the fingertip angle values of other points. Shin and Kim [109] detected fingertips based on the coordinates of the hand position derived from the skeletal information. Meng and Wang [110] read and preprocessed gesture sample templates, which included filtering, segmenting using the HSV color space threshold, and extracting contours. The contours were then calculated approximately to produce polygons, enabling the detection of the fingertips. The Hu moment and the number of fingertips were finally determined.

A convex packet, which can hold all the points in a contour, is a convex polygon created by linking the outermost points. Following contour analysis, convex packet identification is frequently utilized. A convex packet can be built for each contour following the contour analysis of a binary image, and the set of points contained in the packet is returned once the building is finished. The convex packet matching the contour can be drawn using the returned collection of points. Convex profiles are often fully convex or at least flat. A convexity defect is having a concavity in at least one location. In a related study, Wang et al. [111] applied the Douglas-Peucker technique for contour approximation to produce polygons throughout the feature recognition procedure. The type of gesture was then determined by bump detection on the polygons.

Table 3 Comparison of gesture classification methods

Methods		Advantages	Disadvantages
Template matching		High speed in the case of small samples, good adaptability to light and background changes, wide range of applications	Low classification accuracy and limited types of gestures that can be recognized
Geometric information-based	Fingertip detection	Quick detection of the location and number of fingers	Recognition effectiveness may be different for different hand types, and in the case of small finger spacing, false or missed detections may occur
	Convex packet detection	Suitability for recognition of various hand types and good ability to detect the position and number of fingers; ability to perform detection correctly if the finger spacing is small	Greater algorithmic complexity and difficulty of implementation; the possibility of impacted recognition effectiveness in the case of hand occlusion or insufficient light
Dynamic time warping		Good algorithmic performance in matching and recognizing gestures with different motion speeds if the gesture sample template library is relatively small	Greatly reduced recognition speed and stability of the algorithm if the gesture sample template library is large, especially if the gestures are complex or in the case of a combination of two-handed gestures
Hidden Markov model		Ability to capture dynamic features and important timing information in gestures	Needing a large amount of data, which may lead to over- or underfitting if the dataset is too small
Machine learning		Simplicity of the algorithm and ease of implementation and debugging; relatively small data requirements for traditional algorithms	Sensitivity to the interference of light, angle and background of gestures, low accuracy, and the need for manual extraction of gesture features
Deep learning		Automatic extraction of gesture features, eliminating the need for manual extraction, and greater robustness to interference such as that of lighting, angle, and background; high accuracy	Greater algorithmic complexity and the need for a large amount of data and computational resources; longer training time and greater computational requirements

Dynamic time warping

Dynamic time warping (DTW) is a nonlinear time-normalized matching a technique that is frequently used in speech recognition, image matching, gesture classification, etc. It overcomes the matching problem of inconsistent lengths of two sequences. To identify the optimal alignment strategy for the two end sequences, a dynamic programming approach is used to allow the point values of the input sequence and the template sequence to achieve one-to-many or one-to-one matching [112]. The DTW algorithm performs well at matching and recognizing gestures with various motion speeds if the gesture sample template library is not too large. However, if there is a vast number of gesture sample templates, especially if the gestures are complicated or in the case of a mix of two-handed gestures, the identification speed and stability of the algorithm decrease significantly.

The DTW technique was utilized in [113] to determine the optimal alignment between the query features and the stored database features for the recognition of Indian sign language. Zhi et al. [114] implemented and trained two classifiers for static and dynamic gesture recognition: an N-dimensional DTW classifier and a multiclass support vector machine classifier. The running time was greatly reduced, and the average recognition rate was 95.5%.

Hidden Markov model

Russian scientist Vladimir V. Markovnikov developed the hidden Markov model (HMM), a statistical model, in the 1970s [115, 116]. HMM offers a broad range of applications and a great learning capacity and is efficient in modeling time-varying and nonstationary time series. Due to the context-sensitive nature of gesture actions, in the domain of gesture recognition problems HMM is better suited for continuous gesture recognition scenarios. However, HMM training and recognition are computationally demanding in continuous signal applications, where the state transition aspect will require a significant amount of probability density computing, and as the number of parameters rises, the pace of model training and target identification will decline. Discrete HMM is frequently utilized in generic gesture recognition systems to overcome this problem.

Numerous studies [37, 48, 117–119] used HMM for gesture classification. HMM in [37] was used to compute the log-likelihood of the symbols and to identify the most likely route through a network work. In [48], the discrete-density and continuous-density HMMs were trained and tested, and it was demonstrated that the continuous HMM performed better than the discrete HMM at classifying data. In [119], the retrieved symbols were classified and identified using HMM, which was commonly used following the testing of alternative techniques such as Independent Bayesian classifier combinations.

Machine learning

With the use of machine learning, computers can create models that follow the fundamental rules of data in big datasets. For example, Tutsoy [120] proposed an artificial intelligence-based multidimensional policy-making algorithm, which was an advanced predictive model developed under a large number of uncertain factors and time-varying dynamics, aimed at controlling epidemic casualties. Also, Tutsoy [121] proposed a new high-order, multidimensional, strongly coupled, parametric suspicious-infected-death model. This model uses machine learning algorithms to learn from large data sets, including epidemiological data, demographic information and environmental factors, to identify complex relationships and make accurate predictions. There are many well-known machine learning classification algorithms, such as support vector machines, neural networks, and conditional random field and k-nearest neighbor algorithms, which can resolve the problem of gesture recognition.

(1) Support vector machines

Support vector machines, first described in [122] in 1995, are a class of generalized linear classifiers for binary classification of data via supervised learning [123], which is mainly governed by the idea of structural risk minimization and Vapnik–Chervonenkis dimension theory. SVM-based gesture recognition is currently an important research method in gesture recognition technology [124]. SVM is a revolutionary learning technique that optimizes both empirical risk and model complexity. The training error is the constraint of the optimization problem, and the goal is to minimize the confidence range. Because increasing the dimensionality of the sample space has no effect on computational complexity, the SVM approach is frequently applied to high-dimensional problems. The main challenges currently faced by research on gesture-driven interaction are how to process the acceleration values of gesture signals, build a multiclassification model using the SVM algorithm, improve the accuracy of gesture recognition, and create a gesture-based interaction model.

Predicted wrist positions in [125] are used to extract the HOG image descriptors, and a multiclass SVM trained offline is used to categorize the hand shapes. In [126], a previously trained SVM is used to normalize and classify the collected feature vectors. In [127], a method for RGB video-based gesture identification using SVM is proposed. In [128], real-time video-based signs are retrieved using a skin tone segmentation method, and appropriate feature vectors are produced from motion sequences. The features are then categorized using an SVM. In [129], two modules, namely an SVM model for static gesture classification and an HMM for dynamic single-stroke gesture detection, are used to understand a user command consisting of a set of static and dynamic gestures. The classifier for hand gesture recognition in this system is a linear SVM described in [130]. Local binary patterns and a binary SVM classifier are utilized as feature vectors in [131] to look for probable hand motions in every frame of a video stream.

(2) Neural networks

Deep learning algorithms rely on neural networks, which are a subset of machine learning. The name and structure of such algorithms are inspired by the human brain, and they is designed to mimic the way biological neurons communicate with one another [132].

A common structure in artificial neural networks, a multilayer perceptron (MLP), is a feedforward artificial neural network that is typically implemented with the backpropagation (BP) algorithm. Artificial neural networks are highly parallel, have a powerful ability to process information, establish a nonlinear mapping from the input space to the output space, have better fault tolerance and memory function, and store memory information in neurons [133]. The MLP, which accepts the feature set of clusters as input, correctly classifies the clusters and outputs their intensity levels, is the learning algorithm used for classification in [49]. A context-aware gesture-based intelligent system architecture is presented in [134].

Connection weights between the hidden layer and the input layer in radial basis function (RBF) neural networks, which can be of both approximate and exact varieties, are determined in a fixed manner rather than at random [135]. A classification of gestures in photographs using the chosen combinatorial characteristics is proposed in [136]based on an upgraded version of the

RBF neural network. In the latter, the estimated weight matrix is iteratively updated for better hand gesture image identification using the least-mean-square algorithm, and the center is automatically determined using the k-means algorithm.

(3) Conditional random fields
The theory of a conditional random field (CRF) model, first proposed in 2001, was quickly applied to a variety of problems because of its extensions and extensions to the structure of an undirected graph model, which could characterize data dependence problems more accurately than other models. CRF, a probabilistic graph model, was first described by Lafferty et al. [137]. This original construction of a conditional random field was based on HMM in terms of model structure and was influenced by the maximum entropy model (MEM) in terms of model probability representation. In [138], a gesture identification approach was suggested to detect forward and backward movement toward and away from the camera, respectively, using CRF as a classifier and a parallax map-based center-of-mass motion and its intensity's fluctuation as features.

(4) K-nearest neighbors
In 1968, Cover et al. proposed the K-nearest neighbors (KNN) algorithm. Although the idea behind this traditional classification algorithm is straightforward and easy to understand, the algorithm is now very mature and stable [139, 140]. The classification effectiveness of this algorithm, which is one of the fundamental classification algorithms, is good, but due to the use of numerous square root operations, the computational efficiency is insufficient compared to that of other classification algorithms when dealing with complex classification scenarios, particularly in regard to image classification [141, 142]. Jasim et al. [143] used the KNN algorithm to classify static hand gestures and the longest common subsequence (LCS) algorithm to classify dynamic hand gestures.

(5) Naive Bayes classifiers
The Naive Bayes classifier [144] is an algorithm for classification based on the Bayes' theorem. First, it is assumed that the extracted features are uncorrelated and are independent of each other; this assumption simplifies the subsequent operations. Hands that match skin color patches were identified using a Bayesian classifier, and spots with the desired color distribution were recognized and modeled in [145].
In addition to discussing the common machine learning algorithms for gesture recognition described above, many researchers compared different classifiers. Cropped images' HOG characteristics were generated in [45] and used to train the classifier. The study discussed the recognition rates of classifiers such as LDA, SVM, and KNN. In [53], the chosen features were inputs to the ANN, SVM, and KNN models, which were then fused to create a classifier fusion model with the accuracy of 92.23%. In [146], three classification techniques—the mean classifier (NMC), KNN, and the Naive Bayes classifier—were applied to categorize and compare data. In [147], five classifiers—SVM, KNN, Naive Bayes, ANN, and extreme learning machines (ELM)—were utilized for a comparative analysis. Their respective accuracy rates were 96.17% (ELM), 96.95% (SVM), 96.60% (KNN), and 96.38% (NB). Neural networks and Naive Bayesian classification methods based on data mining were applied in [148] for gesture learning and recognition, with the neural network attaining an accuracy of 98.11% and the plain Bayesian classification method having an accuracy of 88.84%.

Deep learning

A recent area of research in machine learning is deep learning. The latter is the process of discovering the innate patterns and depths of sample data representation; the knowledge gained from such learning can be very useful in understanding the meaning of data such as text, photos, and sounds. Its ultimate objective is to provide machines with analytical learning abilities similar to those of humans, enabling machines to recognize data types including text, images, and sounds. Deep learning does not require manual engineering, in contrast to conventional learning algorithms, making it possible to utilize the rapidly expanding amounts of data and computational power that are currently available [149]. Convolutional neural networks and recurrent neural networks are two popular deep learning algorithms.

(1) Recurrent neural network
Saratha Sathasivam proposed the recurrent neural network (RNN), a Hopfield network, in 1982 [150]. Each moment t in a recurrent neural network is processed sequentially and is closely related to the moment t before it. The RNN's potent temporal modeling ability introduces a novel strategy for gesture recognition. However, if there are more than 10-time steps between the relevant input and the target event, it is challenging to train a simple RNN structure [151].
Neverova et al. [152] were the first to use recurrent neural networks for gesture recognition. The researchers' proposal included a multimodal gesture recognition system for speech, skeleton pose, and depth. Prior to a long time-dependent model being built by an RNN for data fusion and final classification, each modality was first processed independently on a short time series, and its features were manually extracted or obtained by learning. To examine the benefits of RNN using various training methods and to suggest an efficient learning process based on suitable

adjustments to the real-time recursive learning algorithm, RNN was also utilized to recognize gestures in [48] and compared with HMM. For skeleton action recognition, Geng et al. [153] proposed a sequence-to-sequence hierarchical RNN structure. Shin and Kim [154] separated the features into various components and fed each hand's input into a GRU-RNN. This enhanced performance and lowered the number of parameters needed for the neural network. Zhang et al. [155] proposed a variant of a long short-term memory(LSTM) model for dynamic gesture recognition by combining ResC3D and ConvLSTM.

(2) Convolutional neural network

A convolutional neural network (CNN) [156] is a feed-forward neural network that has emerged quickly in the field of image analysis and processing. CNN effectively avoids the preprocessing stage and a substantial amount of manual involvement in the project compared to traditional image processing algorithms. However, a large amount of data consists of more than just one image. To process video data, a 3D CNN [157] should be created and used as soon as possible for the task of behavior recognition in surveillance videos.

Two-dimensional convolutional neural networks are mostly used to process static gestures or dynamic gesture sequences on a frame-by-frame basis. John et al. [158] used a long-term recursive neural network to classify gesture video sequences. All 24 motions from Thomas Moeslund's gesture recognition database were used to apply deep learning to the gesture identification problem in [159], demonstrating that deep neural networks were capable of learning complicated gesture categorization tasks with a low error rate. The approach in [160] combined a skeletonization algorithm with CNN, which lessened the impact of capture angle and surroundings on the recognition effectiveness and increased the precision of gesture recognition in complicated contexts. In [161], a transformation of Arabic sign language letters into Arabic voice using a vision- and CNN-based system was suggested. In [162], a comparison of various gesture recognition techniques showed that CNN outperformed other classification systems. Noreen et al. [163] proposed a multiparallel streaming two-dimensional CNN model to recognize hand gestures.

Several three-dimensional convolutional neural network (3D-CNN) models have been proposed for gesture recognition. To address the lack of a large number of labeled gesture datasets, an efficient deep convolutional neural network method called 3D-CNN was proposed in [164]. A 3D-CNN model is suggested by Molchanov et al. [165] to identify driving gestures based on depth and intensity data and to combine data from various spatial scales for the final prediction. Using a recurrent mechanism for dynamic gesture detection and classification, Molchanov et al. [166] enhanced the 3D-CNN model. A 3D-CNN structure for extracting spatiotemporal features and a recurrent layer for global temporal modeling were both included in the network model. Li et al. [167] enhanced the 3D-CNN model of Tran et al. [168] for large-scale gesture recognition using depth and RGB videos. Similarly to the above study of Tran et al., Camgoz et al. [169] developed an end-to-end 3D-CNN model for extensive gesture recognition.

Lightweight convolutional neural networks were created in recent years as a result of the development of convolutional neural networks by numerous academics. A hardware-friendly neural network was made possible by a lightweight neural network, which was a lighter model that performed on par with a heavier model. The accuracy of the MobileNetv2- and CNN-based gesture recognition in [170] was 99.96%. Baumgartl et al. [170] proposed a lightweight, robust and fast CNN for manual gesture recognition by image classification. In [171], a hybrid network structure of a lightweight VGG16 model and a random forest was presented for recognition of gestures based on visual input.

Experimental evaluation

In "Hand gesture recognition process", the process of gesture recognition has been described. In this section, several evaluation metrics for gesture recognition and segmentation are presented.

Accuracy

Accuracy is the ratio of the number of samples correctly classified by a classifier to the total number of samples. In gesture recognition and segmentation, the accuracy rate can be used to measure the overall performance of the classifier. The calculation formula is

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}},$$

where TP denotes true-positive cases, TN denotes true-negative cases, FP denotes false-positive cases, and FN denotes false-negative cases.

Precision

The precision rate is the percentage of samples identified by the classifier as belonging to positive classes that indeed belong to positive classes. In gesture recognition and segmentation, the precision rate can be used to measure the accuracy of the classifier. The calculation formula is

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

Recall

The recall rate is the proportion of samples that indeed belong to positive classes and that the classifier correctly identifies as belonging to positive classes. In gesture recognition and segmentation, recall can be used to measure the completeness of the classifier. The formula is as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

F1 score

The F1 score is the summed average of accuracy and recall, and hence assesses both the accuracy and the completeness of the classifier. In gesture recognition and segmentation, the F1 score can be used as an evaluation metric to help select the optimal classifier. The calculation formula is

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

Intersection over union (IoU)

IoU is the ratio of the overlapping part between the predicted region and the real one to the overall size. In gesture segmentation, IoU can be used to measure the segmentation effectiveness of the model. The calculation formula is

$$\text{IoU} = \frac{\text{Intersection}}{\text{Union}},$$

where intersection denotes the intersection of the predicted frame and the true frame, and union indicates the merging of the predicted frame and the true frame. The larger the intersection is, the closer the predicted result is the truth.

Hand gesture recognition based on RGB-D cameras

The paper has introduced the use of three widely used pairs of depth cameras, namely, Kinect, Leap Motion and RealSense, in "Introduction". The data structure of RGB-D images produced by depth cameras is more complicated than that of earlier 2D images, opening new possibilities for gesture recognition studies. A simple thresholding algorithm can accurately split the hand zone in the depth map using depth data, reducing the gesture detection problem to a problem of recognition of the 3D shape of the hand.

Given the popularity of depth sensors, scholars have conducted extensive research on gesture segmentation based on depth information. In [172], Jiang used the Kinect sensor to gather depth information, established a threshold for each frame in accordance with each pixel's depth value, extracted the largest region as the foreground, and then removed the other patches with smaller areas. Kane and Khanna [173] described the creation of an acquisition module that used depth thresholding and velocity tracking to execute pen lifting and falling movements. The hand needed to be in the foreground of the camera for the depth thresholding-based hand segmentation method to work well. Zhao and Jia [59] presented an enhanced hand segmentation approach based on the random choice forest framework for depth sensor-acquired images by manually integrating the essentials of depth thresholding-based segmentation methods. To ensure the accuracy of hand segmentation, the method generated new depth features from the centroid of the hand structure, improved the generalizability of earlier depth features, and specified the depth invariance of hand pixels as much as possible.

Using a depth camera, more information about the appearance features can be obtained. In [174], a hand contour model was suggested to make gesture matching easier by incorporating the Kinect sensor, which could make gesture matching less computationally complex. Brazilian sign language's phonetic structure was investigated in [175], relying on RGB-D sensors to collect intensity, location, and depth data. Seven vision-based traits were extracted by Almeida et al. [175] from RGB-D photos. Almeida et al. studied the relationship between the extracted features and the structural elements based on hand shape, motion, and location in Brazilian sign language. A 3D hand shape was separated from a cluttered background using a depth map of hand gestures recorded by the Kinect sensor to extract patterns of 3D shape features, and a 3D shape context description approach was proposed in [176] for 3D gesture representation. A TOF depth camera was used in [177] to gather depth data, determine the wrist cut edge, and and capture palm images. In [178], the coordinates of the 21 bonding points of the human hand were recorded using Leap Motion, and the motion images were captured using an RGB camera.

One of the more popular state space-based techniques for matching time-varying data is a hidden Markov model with two main stages, namely, training and classification. The Baum–Welch algorithm is a fundamental algorithm used to solve the training problem, whereas the Viterbi algorithm is a fundamental algorithm used to solve the classification problem [179]. Hoque et al. [180] proposed a real-time gesture recognition system based on Kinect that could manipulate desktop objects by identifying the hand's 3D position using Kinect's depth sensor. To identify predefined gestures, such location points were subsequently examined. The HMM was

Table 4 Results of the studies

References	Experimental results
[172]	Average task completion time for target capture: "semiautomatic" (176.9s) and "manual" (287.4s)
[173]	On the self-acquisition dataset, the EPS solution reached an accuracy of over 96.5% with an average runtime of 30 ms. On the AIR Handwriting dataset, the recognition time per gesture was 24.3 ms with an average accuracy of 95.5%
[59]	The percentages of false negatives and false positives were 2.00% and 4.38%, respectively. Training time was approximately 16min. Real-world image classification time was approximately 2.1 sec/frame
[174]	The accuracy rate of hand inspection procedures improved to 95.43%. The average accuracy of hand part classification improved to 74.65%
[175]	The average recognition rate was above 80%
[177]	Average recognition success rate of 84.5%
[178]	Highest recognition accuracy of up to 99.66%
[180]	The accuracy rate reached 89%
[181]	The best accuracy for static gesture recognition was 95.6%. The best accuracy for dynamic gesture recognition was 97.2%
[176]	On the NTU Hand Digit Dataset, the best obtained performance was 98.7%. On the Kinect Leap DataSet, an accuracy of 96.8% was reached. On the Senz3d Dataset, an accuracy of 100% was attained. On the ASL-FS Dataset, an accuracy of 87.1% was obtained. On the ChaLearn LAP IsoGD Dataset, the accuracy of 60.12% was reached. The average runtime per query on an average PC (without a GPU) was only 6.3 ms

trained using the Baum–Welch algorithm, which resulted in the accuracy rate of 89%. A dynamic hand gesture detection system based on an RGB-D camera was proposed by Simao et al. [181]. Hand segmentation in color photos was performed using a large illumination-invariant skin tone model, and hand detection in depth images was performed using a chamfer distance matching-based technique. Hand movements were modeled and classified using the HMM with the left-right band state graph topology.

Table 4 shows the results of the studies mentioned above.

Hand gesture recognition applications

Gesture recognition has a wide range of applications, such as healthcare, safe driving, sign language awareness, virtual reality, and device control. This section mainly focuses on human-robot interaction using vision-based hand gestures captured by monocular cameras and RGB-D cameras. The main application areas for gesture recognition technologies are listed below.

- Healthcare: Emergency rooms and operating rooms can be chaotic, with a significant amount of noise from individuals and equipment. In such an environment, voice commands are not as effective as hand gestures. Touchscreens are also not an option because of the strict boundaries between sterile and nonsterile domains. However, accessing information and images during surgery or other procedures is possible with gesture recognition technology, as demonstrated by Microsoft. GestSure, a gesture control technology that can be used to control medical devices, allows physicians to examine MRI, CT and other images with simple gestures without scrubbing. This touch-free interaction reduces the number of times doctors and nurses touch patients, reducing the risk of cross-contamination.
- Safe driving: Advanced driver assistance systems that incorporate gesture recognition can somewhat increase driving safety. Through an advanced driver assistance system, drivers can modify many parameters inside the automobile using gestures, allowing them to focus more on the road and perhaps reducing traffic accidents. The BMW 7 Series has an integrated hand gesture recognition system that recognizes five gestures to control music, incoming calls, etc. Reducing interaction with the touchscreen makes the driving experience safer and more convenient.
- Sign language awareness: The primary means of communication for hearing-impaired individuals is sign language; however, understanding sign language is difficult for those who have not received formal instruction. The ability of hearing-impaired and other individuals to communicate will be enhanced substantially using sign recognition technology for sign language cognition. The Italian startup Limix combines IoT and dynamic gesture recognition technology to record sign language, translate it to text, and then play it back on a smartphone via a voice synthesizer.
- Virtual reality: Gesture recognition allows users to interact with and control virtual reality scenes more naturally, enhancing users' immersion and experience. In 2016, Leap Motion demonstrated updated gesture recognition software that allowed users to track gestures in virtual reality in addition to controlling computers. ManoMotion's hand-tracking application recognizes 3D gestures through a smartphone camera (on Android and iOS) and can be applied to AR and VR environments. Use cases for this technology include gaming, IoT devices, consumer electronics, and robotics.
- Device control: Intelligent robots can also be controlled by gestures. With the advancement of artificial intelligence, home robots or smart home equipment will progressively appear in millions of households, and consumers will feel more at ease using gesture control as

opposed to traditional button or touch screen input. A company called uSens develops hardware and software that enables Smart TVs to recognize finger movements and gestures. Gestoo's artificial intelligence platform uses gesture recognition technology to enable touchless control of lighting and audio systems. With Gestoo, gestures can be created and assigned from a smartphone or another device, and a single gesture can be used to activate multiple commands.

Robots are becoming more prevalent in our daily lives as robotics develops quickly. Robots must learn how to communicate with people to be fully integrated into the human civilization. Researchers choose gesture recognition over other emerging human-robot interface technologies because of its straightforward and natural interaction qualities, rich expressive capabilities, and potential for a wide range of applications. Tables 5 and 6 present the applications of gesture recognition in nonmobile and moving robots, respectively.

Problems, outlook, and conclusion

Problems

Gesture recognition technology has developed rapidly in recent years. However, due to the interference of external environmental factors and the limitations of gestures themselves, it is very easy to introduce various effects into the system; thus, gesture recognition still faces insurmountable difficulties. To facilitate the improvement of gesture recognition-driven human-computer interaction, this paper summarizes the following problems.

Data gathering

Most existing studies of hand detection assume a simple background for gestures during data acquisition; indeed, it is challenging to capture a hand because it is a relatively small object with many complex articulations [29]. However, in the practical application of robotics, workers are typically in complex environments, and thus we need to improve gesture recognition methods and apply them to real-world scenarios. For instance, the skeleton segmentation method can be utilized to distinguish locations other than the hand due to the specificity of skin color segmentation. One of the main challenges in identifying hand motion is frequently the complex posture of the hand, which is affected by the occlusion of the fingers [203, 204].

Training data environment

Scholars have compared gesture recognition methods, training on various datasets and obtaining noticeably different results, hence proving the importance of datasets in gesture training. The accuracy of various methods on different datasets was reviewed by Rawat et al. [205]. Future research can concentrate on honing the training dataset, making it as rich and varied as possible, and enhancing the capability of gesture recognition techniques to recognize gestures in any situation because of the nature of deep learning, e.g., by using convolutional neural networks.

The lack of high-quality datasets captured from various angles makes it difficult to create a highly realistic model that takes into account the actual contours of the hand [206, 207]. Due to the hand's complex structure and varied dimensions, it has many degrees of freedom, which increases occlusion in noncomplex environments [208, 209].

It has also been noted that most databases used in gesture recognition research originate from various nations; in the case of the sign language, for instance, different nations have different sign languages and gestures. Therefore, more focus on uncontrolled situations is required to improve vision-based gesture recognition systems for real-world applications. To assess the transfer learning ability, future experiments may try to transfer knowledge from the gestural symbol system proposed by Zengeler et al. [210] to other gestural languages.

Identification speed

We also need to consider real-time networks: some have strong identification rates but poor real-time performance, while others have the opposite problem. When used in the real world, for instance, in surgical robots, this poses serious and troubling issues. Since gesture recognition systems require both high recognition accuracy and good real-time performance, we must increase the latter and reduce time consumption without sacrificing recognition accuracy. To further reduce computing costs and boost recognition effectiveness for application-level and real-time gesture identification, Liu et al. [211] chose to incorporate image segmentation methods in the future.

Segmentation in complex background

Currently, most existing studies of gesture recognition processes assume that the background of gestures is simple; however, the background in applications is complex. For example, during the human–machine interaction between a machine and a worker, the gestures captured by a sensor device will contain many complex influences of the background environment, including changes in lighting and the

Table 5 Gesture recognition applied to nonmobile robots

References	Cameras	Focus
[182]	Monocular	Built a completely automated hand gesture detection system and used it for a robot that assisted individuals in libraries
[183]	Monocular	Used a hand gesture detection system to improve conventional teaching techniques by replacing them with a simple gesture-based control scheme
[184]	Monocular	Suggested Eureka, a deep learning-based gesture recognition method that combined a feature extractor and a neural network
[185]	Monocular	Fed many preprocessed sample images into a CNN, which subsequently performed feature extraction, to apply the YOLOv4 method to distinguish gesture features
[186]	RGB-D	Applied Kalman filtering to the raw data to lessen the jitter or jumps produced during data capture by Leap Motion
[187]	RGB-D	Built a multisensor data fusion model and proposed a multilayer RNN consisting of an LSTM module
[188]	RGB-D	Enhanced communication between the Arduino Mega chip and Microsoft Kinect V2 to construct an industrial robot with 7 degrees of freedom operated by hand gestures
[189]	RGB-D	Presented a deep learning-based hand gesture categorization network and hand detection model, and performed pixel-level fusion of RGB and depth images with hand information
[190]	RGB-D	Used Leap Motion to collect hand data, and developed a neural network method to categorize nine hand movements, and a finite state machine to control the robot

background environment, which will increase the difficulty of gesture detection and lead to a decrease in the accuracy of gesture recognition. Many researchers have sought to enhance the robustness of gesture recognition in complex backgrounds and to improve the interaction capability of gesture recognition in complex scenes.

Sheenu et al. [212] proposed a new method for gesture recognition in images with complex backgrounds based on histograms of the orientation gradient and sequential minimal optimization, which had an overall recognition rate of 93.12% for complex backgrounds. Chen et al. [213] suggested a gesture recognition method based on an improved YOLOv5 approach that reduced various types of interference in gesture images with complex backgrounds and improved the robustness of the network to complex backgrounds. Zhang et al. [214] proposed a two-stage gesture recognition method. In the first stage, the convolutional pose machine was used to localize a hand's key points, which could effectively localize the hand's key points even in cases of complex backgrounds. Vishwakarma [215] researched and developed a method for effective detection and classification of hand gestures in cases of complex backgrounds. Pabendon et al. [216] suggested a gesture recognition method based on spatiotemporal domain pattern analysis, which could significantly reduce the irregular noise affecting gesture recognition in cases of complex backgrounds. Elsayed et al. [217] described a robust gesture segmentation method based on adaptive background subtraction with skin color thresholding, which aimed to automatically segment gestures from a given video in cases of different lighting conditions and complex backgrounds. Qi et al. [218] suggested an improved atrous spatial pyramid pool to improve the accuracy of gesture feature representation in images. Zhou et al. [219] proposed a two-stage gesture recognition system to solve the problem of recognizing gestures in cases of complex backgrounds.

Distance and hand anatomy

The distance between the cameras and the person is an important factor in hand gesture segmentation. If the cameras are too far away, hand gestures may not be captured accurately, resulting in incorrect segmentation. However, if the cameras are too close, there may be occlusion as hands move in and out of the frame, leading to incomplete segmentation.

Additionally, the hand anatomy should be considered. Different types of hand gestures involve different parts of the hand, and some gestures may be more difficult to segment accurately than others. For example, gestures that involve the fingers being close together may be more challenging to distinguish from each other.

To improve hand gesture segmentation accuracy, researchers may use various techniques, such as depth sensing (e.g., RealSense and Kinect), machine learning algorithms, and hand-tracking algorithms. These methods can help identify the different parts of the hand and track their movement accurately even in complex gesture sequences.

Table 6 Gesture recognition applied to moving robots

References	Cameras	Focus
[191]	Monocular	Proposed an HRI system based on the HMM that could recognize meaningful gestures composed of continuous hand movements in real time
[192]	Monocular	Tracked a robot's subsequent gestures and used them to transmit data for movement control
[193]	Monocular	Proposed a template matching algorithm to recognize gestures and control the motion of mobile carts by combining invariant moment matching techniques
[194]	Monocular	Integrated the improved YOLOv5 algorithm for hand pinpointing and the Resnet-152 method for hand classification
[195]	Monocular	Proposed E-MobileNetv2, an enhanced lightweight CNN, for classification
[196]	Monocular	Built a gesture change detection technique by using an upgraded residual neural network, as well as a hand segmentation algorithm enhanced by skin color detection and skeletal joint tracking
[197]	Monocular	Estimated gesture motion from time-series data of hand coordinates by using a one-dimensional fast Fourier transform and estimated two-dimensional coordinates of hand areas in images from color information
[198]	Monocular	Applied the 3DCNN architecture to gesture recognition and implemented a system for directing robots with video-recorded gestures in real-world scenarios
[199]	RGB-D	Improved a robot's capacity to recognize gestures by training on the data captured by a Leap Motion controller using classifiers (SVM, KNN, and HMM)
[200]	RGB-D	Suggested a two-handed gesture recognition method based on depth cameras for real-time control of a mechanical wheeled mobile robot
[58]	RGB-D	Described a dynamic gesture recognition system based on depth sensors for a continuous operation of material-handling robots
[201]	RGB-D	Created a real-time skeleton-based five-gesture detection system using depth cameras and machine learning
[202]	RGB-D	Proposed a dynamic gesture detection technique based on 3D hand posture estimation

Future outlook

With the rise of artificial intelligence, deep learning is undoubtedly a benign accelerator for gesture recognition, and gesture recognition systems will become more accurate and stable. Future gesture recognition systems will also be more diversified and applicable to more fields such as medical care, education, entertainment, etc., bringing more convenience and innovation to people. In the future, gesture recognition technology will continue to develop in the following directions:

- More intelligent: Gesture recognition will become more intelligent with the continued development of deep learning and artificial intelligence technology. Training a model will allow it to understand more complex gestures while reducing the user requirements, making gesture recognition more natural and intelligent.
- More accurate: As computer vision and sensor technology continue to improve, gesture recognition will become more accurate. For example, higher-resolution cameras and more sensitive sensors can capture more subtle hand movements, improving the accuracy of gesture recognition.
- More capable of real-time performance: Future gesture recognition technology will operate closer to real time, and be capable of processing large numbers of gestures and translating them into commands or actions. This will enable gesture recognition's wider use in virtual reality, gaming, medical, and other fields.
- More reliable: As the applications of gesture recognition technology expand, its reliability becomes increasingly important. Future gesture recognition technologies will require more rigorous testing and validation to ensure their reliable operation in a variety of environments.
- More personalized: Future gesture recognition technologies will be more personalized and able to adapt to different users' gesture habits and preferences. For example, users may be able to customize specific gestures to accomplish a particular operation or function.

Conclusions

This paper focuses on the processing steps and techniques for gesture recognition. Gesture recognition methods are divided according to four steps: data acquisition, gesture detection and segmentation, feature extraction and gesture classification. The focus of this paper is on RGB-D camera-based gesture recognition techniques, but it also covers several related studies in the field that use monocular and depth cameras. Contrasting these two methods, we observe that the depth camera-based gesture recognition method is more practical and effective. It can be used to perform both dynamic and static gesture recognition. The research on gesture recognition's implementation in robotic scenarios is then reviewed and analyzed, and algorithms for gesture recognition in human-robot interaction are discussed. The problems faced by vision-based gesture recognition methods over the years, the progress that can be made, and the possible future directions are reviewed.

Acknowledgements The work is supported by the National Natural Science Foundation of China (Grant No. 62173037), Scientific Research Foundation of Hebei University for Distinguished Young Scholars (Grant No. 521100221081), Scientific Research Foundation of Colleges and Universities in Hebei Province (Grant No. QN2022107), Science and Technology Program of Hebei Province (Grant No. 22370301D), Key Program of Hebei Province (Grant No. 22370301D), and Innovation and Entrepreneurship Training Program for College students of Hebei University (Grant No. 2022156).

Author contributions LM wrote the manuscript, ZC and YY modified the manuscript, JQ gave some guidance.

Funding National Natural Science Foundation of China (Grant No.62173037), Scientific Research Foundation of Hebei University for Distinguished Young Scholars (Grant No. 521100221081), Scientific Research Foundation of Colleges and Universities in Hebei Province (Grant No. QN2022107), Science and Technology Program of Hebei Province (Grant No. 22370301D), Key Program of Hebei Province (Grant No. 22370301D), and Innovation and Entrepreneurship Training Program for College students of Hebei University (Grant No. 2022156)

Availability of data and materials Data sharing is not applicable to this article as no new data were created oranalyzed in this study.

Code Availability Not applicable.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

1. Komura T, Lam W-C (2006) Real-time locomotion control by sensing gloves. Comput Anim Virtual Worlds 17(5):513–525
2. Kim M, Cho J, Lee S, Jung Y (2019) Imu sensor-based hand gesture recognition for human-machine interfaces. Sensors 19(18):3827
3. Rodriguez G, Jofre N, Alvarado Y, Fernández J, Guerrero R (2017) Gestural interaction for virtual reality environments through data gloves. Adv Sci Technol Eng Syst J 2(3):284–290
4. Helen Jenefa R, Gokulakrishnan K (2018) Bluetooth enabled electronic gloves for hand gesture recognition. In: International conference on computer networks, big data and IoT. Springer, pp 771–777
5. Huang H, Liang Z, Sun F, Dong M et al (2022) Virtual interaction and manipulation control of a hexacopter through hand gesture recognition from a data glove. Robotica 40(12):4375–4387
6. Antillon DWO, Walker CR, Rosset S, Anderson IA (2022) Glove-based hand gesture recognition for diver communication. IEEE Trans Neural Netw Learn Syst
7. Mummadi CK, Philips Peter Leo F, Deep Verma K, Kasireddy S, Scholl PM, Kempfle J, Van Laerhoven K (2018) Real-time and embedded detection of hand gestures with an imu-based glove. In: Informatics, vol 5. MDPI, p 28
8. Vuskovic M, Du S (2002) Classification of prehensile emg patterns with simplified fuzzy artmap networks. In: International joint conference on neural networks, pp 2539–2544
9. Nazarpour K, Sharafat AR, Firoozabadi S (2005) Surface emg signal classification using a selective mix of higher order statistics. In: Conference Proceedings: ... annual international conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society, pp 4208–4211
10. Wu Y, Liang S, Zhang L, Chai Z, Cao C, Wang S (2018) Gesture recognition method based on a single-channel semg envelope signal. EURASIP J Wirel Commun Netw 2018:1–8
11. Guo S, Pang M, Gao B, Hirata H, Ishihara H (2015) Comparison of semg-based feature extraction and motion classification methods for upper-limb movement. Sensors 15(4):9022–9038
12. Kim J, Cho D, Lee KJ, Lee B (2014) A real-time pinch-to-zoom motion detection by means of a surface emg-based human-computer interface. Sensors 15(1):394–407
13. Bahl P, Padmanabhan VN (2000) Radar: an in-building rf-based user location and tracking system. In: Proceedings IEEE INFOCOM 2000. Conference on computer communications. Nineteenth annual joint conference of the IEEE computer and communications societies (Cat. No. 00CH37064), vol 2. IEEE, pp 775–784
14. Pu Q, Gupta S, Gollakota S, Patel S (2013) Whole-home gesture recognition using wireless signals

15. Zhang O, Srinivasan K (2016) Mudra: user-friendly fine-grained gesture recognition using wifi signals. In: Proceedings of the 12th international on conference on emerging networking experiments and technologies, pp 83–96
16. Wang Y, Wu K, Ni LM (2016) Wifall: device-free fall detection by wireless networks. IEEE Trans Mob Comput 16(2):581–594
17. Wu X, Chu Z, Yang P, Xiang C, Zheng X, Huang W (2018) Tw-see: human activity recognition through the wall with commodity wi-fi devices. IEEE Trans Veh Technol 68(1):306–319
18. Hisham B, Hamouda A (2019) Supervised learning classifiers for Arabic gestures recognition using kinect v2. SN Appl Sci 1(7):1–21
19. De Smedt Q, Wannous H, Vandeborre J-P (2016) Skeleton-based dynamic hand gesture recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 1–9
20. Rautaray SS, Agrawal A (2015) Vision based hand gesture recognition for human computer interaction: a survey. Artif Intell Rev 43:1–54
21. Ji Y, Kim S, Lee K-B (2017) Sign language learning system with image sampling and convolutional neural network. In: 2017 first IEEE international conference on robotic computing (IRC). IEEE, pp 371–375
22. ElBadawy M, Elons A, Shedeed HA, Tolba M (2017) Arabic sign language recognition with 3d convolutional neural networks. In: 2017 eighth international conference on intelligent computing and information systems (ICICIS). IEEE, pp 66–71
23. Ĉadík M (2008) Perceptual evaluation of color-to-grayscale image conversions. In: Computer graphics forum, vol 27. Wiley Online Library, pp 1745–1754
24. Benedetti L, Corsini M, Cignoni P, Callieri M, Scopigno R (2012) Color to gray conversions in the context of stereo matching algorithms: an analysis and comparison of current methods and an ad-hoc theoretically-motivated technique for image matching. Mach Vis Appl 23:327–348
25. Fairchild MD (2013) Color appearance models. Wiley, New York
26. Rosenfeld A (1976) Digital picture processing. Academic Press, Cambridge
27. Xu Y, Gu J, Tao Z, Wu D (2009) Bare hand gesture recognition with a single color camera. In: 2009 2nd international congress on image and signal processing. IEEE, pp 1–4
28. Zhang H, Wang Y, Deng C (2011) Application of gesture recognition based on simulated annealing bp neural network. In: Proceedings of 2011 international conference on electronic & mechanical engineering and information technology, vol 1. IEEE, pp 178–181
29. Lahiani H, Elleuch M, Kherallah M (2015) Real time hand gesture recognition system for android devices. In: 2015 15th international conference on intelligent systems design and applications (ISDA). IEEE, pp 591–596
30. Canny J (1986) A computational approach to edge detection. IEEE Trans Pattern Anal Mach Intell 6:679–698
31. Panda CS, Patnaik S (2010) Better edgegap in grayscale image using gaussian method. Int J Comput Appl Math 5(1):53–66
32. Deng G, Pinoli J-C (1998) Differentiation-based edge detection using the logarithmic image processing model. J Math Imaging Vis 8:161–180
33. Sonka M, Hlavac V, Boyle R (2014) Image processing, analysis, and machine vision. Cengage Learning, Boston
34. Otsu N (1979) A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern 9(1):62–66
35. Niblack W (1985) An introduction to digital image processing. Strandberg Publishing Company, Copenhagen
36. Malima AK, Özgür E, Çetin M (2006) A fast algorithm for vision-based hand gesture recognition for robot control
37. Zaki MM, Shaheen SI (2011) Sign language recognition using a combination of new vision based features. Pattern Recognit Lett 32(4):572–577
38. Shangeetha R, Valliammai V, Padmavathi S (2012) Computer vision based approach for Indian sign language character recognition. In: 2012 international conference on machine vision and image processing (MVIP). IEEE, pp 181–184
39. Bhame V, Sreemathy R, Dhumal H (2014) Vision based hand gesture recognition using eccentric approach for human computer interaction. In: 2014 international conference on advances in computing, communications and informatics (ICACCI). IEEE, pp 949–953
40. Dhule C, Nagrare T (2014) Computer vision based human-computer interaction using color detection techniques. In: 2014 fourth international conference on communication systems and network technologies. IEEE, pp 934–938
41. Ahuja MK, Singh A (2015) Static vision based hand gesture recognition using principal component analysis. In: 2015 IEEE 3rd international conference on moocs, innovation and technology in education (MITE). IEEE, pp 402–406
42. Veluchamy S, Karlmarx L, Sudha JJ (2015) Vision based gesturally controllable human computer interaction system. In: 2015 international conference on smart technologies and management for computing, communication, controls, energy and materials (ICSTM). IEEE, pp 8–15
43. Sreekanth N, Narayanan N (2017) Dynamic gesture recognition—a machine vision based approach. In: Proceedings of the international conference on signal, networks, computing, and systems. Springer, pp 105–115
44. Wang K, Xiao B, Xia J, Li D, Luo W (2016) A real-time vision-based hand gesture interaction system for virtual east. Fusion Eng Des 112:829–834
45. Patel P, Patel N (2019) Vision based real-time recognition of hand gestures for Indian sign language using histogram of oriented gradients features. Int J Next-Gener Comput 10:92–102
46. Zhou W, Lyu C, Jiang X, Li P, Chen H, Liu Y-H (2017) Real-time implementation of vision-based unmarked static hand gesture recognition with neural networks based on fpgas. In: 2017 IEEE international conference on robotics and biomimetics (ROBIO). IEEE, pp 1026–1031
47. Gupta L, Ma S (2001) Gesture-based interaction and communication: automated classification of hand gesture contours. IEEE Trans Syst Man Cybern Part C (Applications and Reviews) 31(1):114–120
48. Ng CW, Ranganath S (2002) Real-time gesture recognition system and application. Image Vis Comput 20(13–14):993–1007
49. Sharma N, Maringanti HB, Asawa K (2012) Upper body pose recognition and classifier. In: Acm compute conference: intelligent & scalable system technologies
50. Sun J, Zhang Z, Yang L, Zheng J (2020) Multi-view hand gesture recognition via pareto optimal front. IET Image Proc 14(14):3579–3587
51. Li Y (2012) Hand gesture recognition using kinect. In: 2012 IEEE international conference on computer science and automation engineering. IEEE, pp 196–199
52. Anant S, Veni S (2018) Safe driving using vision-based hand gesture recognition system in non-uniform illumination conditions. J ICT Res Appl 12(2)
53. Singha J, Roy A, Laskar RH (2018) Dynamic hand gesture recognition using vision-based approach for human-computer interaction. Neural Comput Appl 29(4):1129–1141
54. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. Comput Sci
55. Paul S, Bhattacharyya A, Mollah AF, Basu S, Nasipuri M (2020) Hand segmentation from complex background for gesture recog-

nition. In: Emerging technology in modelling and graphics: proceedings of IEM graph 2018. Springer, pp 775–782

56. Shelhamer E, Long J, Darrell T (2016) Fully convolutional networks for semantic segmentation. IEEE Trans Pattern Anal Mach Intell 1
57. Badrinarayanan V, Kendall A, Cipolla R (2017) Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell 1
58. Ding I-J, Su J-L (2022) Designs of human-robot interaction using depth sensor-based hand gesture communication for smart material-handling robot operations. Proc Inst Mech Eng Part B J Eng Manuf 237(3):392–413
59. Zhao M, Jia Q (2016) Hand segmentation using randomized decision forest based on depth images. In: 2016 international conference on virtual reality and visualization (ICVRV). IEEE, pp 110–113
60. Fukunaga K, Hostetler L (1975) The estimation of the gradient of a density function, with applications in pattern recognition. IEEE Trans Inf Theory
61. Guo Y, Şengür A, Akbulut Y, Shipley A (2018) An effective color image segmentation approach using neutrosophic adaptive mean shift clustering. Measurement 119:28–40
62. Khan B, Khan AK, Raja G, Yousaf MH (2013) Implementation of modified mean-shift tracking algorithm for occlusion handling. Life Science Journal 10(11s):337–342
63. Bradski GR (1998) Computer vision face tracking for use in a perceptual user interface. Intel Technol J
64. Allen JG, Xu R, Jin JS (2004) Object tracking using camshift algorithm and multiple quantized feature spaces
65. Ghotkar A, Kharate G (2012) Hand segmentation techniques to hand gesture recognition for natural human computer interaction. Int J Hum Comput Interact 3(1):15–25
66. Akmeliawati R, Dadgostar F, Demidenko S, Gamage N, Sengupta G (2009) Towards real-time sign language analysis via markerless gesture tracking. In: IEEE instrumentation & measurement technology conference
67. Collins RT, Lipton AJ, Kanade T, Fujiyoshi H, Burt P (2000) A system for video surveillance and monitoring. VSAM final report, Carnegie Mellon University Technical Report
68. Shen Y, Wen H, Yang M, Liu J, Chou CT (2012) Efficient background subtraction for tracking in embedded camera networks. ACM
69. Apolinário L, Armesto N, Cunqueiro L (2012) An analysis of the influence of background subtraction and quenching on jet observables in heavy-ion collisions
70. Denman S, Chandran V, Sridharan S (2007) An adaptive optical flow technique for person tracking systems. Pattern Recognit Lett 28(10):1232–1239
71. Jayabalan E, Krishnan A, Pugazendi R (2007) Non rigid object tracking in aerial videos by combined snake and optical flow technique. In: Computer graphics, imaging & visualisation
72. Chanda K, Ahmed W, Mitra S (2015) A new hand gesture recognition scheme for similarity measurement in a vision based barehanded approach. In: International conference on image information processing, pp 17–22
73. Liao C-J, Su S-F, Chen M-C (2015) Vision-based hand gesture recognition system for a dynamic and complicated environment. In: 2015 IEEE international conference on systems, man, and cybernetics. IEEE, pp 2891–2895
74. De O, Deb P, Mukherjee S, Nandy S, Chakraborty T, Saha S (2016) Computer vision based framework for digit recognition by hand gesture analysis. In: 2016 IEEE 7th annual information technology, electronics and mobile communication conference (IEMCON). IEEE, pp 1–5
75. Panigrahi A, Mohanty JP, Swain AK, Mahapatra K (2018) Real-time efficient detection in vision based static hand gesture recognition. In: 2018 IEEE international symposium on smart electronic systems (iSES)(Formerly iNiS). IEEE, pp 265–268
76. Wachs JP, Stern HI, Edan Y, Gillam M, Handler J, Feied C, Smith M (2008) A gesture-based tool for sterile browsing of radiology images. J Am Med Inform Assoc 15(3):321–323
77. Ganokratanaa T, Pumrin S (2017) The vision-based hand gesture recognition using blob analysis. In: 2017 international conference on digital arts, media and technology (ICDAMT). IEEE, pp 336–341
78. Shan C, Wei Y, Tan T, Ojardias F (2004) Real time hand tracking by combining particle filtering and mean shift. In: IEEE international conference on automatic face & gesture recognition, pp 669–674
79. Shan C, Tan T, Wei Y (2007) Real-time hand tracking using a mean shift embedded particle filter. Pattern Recognit 40(7):1958–1970
80. Li P, Zhang T, Pece A (2003) Visual contour tracking based on particle filters. Image Vis Comput 21(1):111–123
81. Ma C, Wang A, Ge C, Chi X (2018) Hand joints-based gesture recognition for noisy dataset using nested interval unscented kalman filter with lstm network. Vis Comput 34(6–8):1053–1063
82. Lech M, Kostek B (2012) Hand gesture recognition supported by fuzzy rules and kalman filters. Int J Intell Inf Database Syst 6(5):407–420
83. Kumar G, Bhatia PK (2014) A detailed review of feature extraction in image processing systems. In: 2014 fourth international conference on advanced computing & communication technologies. IEEE, pp 5–12
84. Luan S, Chen C, Zhang B, Han J, Liu J (2018) Gabor convolutional networks. IEEE Trans Image Process 27(9):4357–4366
85. Dalal N, Triggs B, Schmid C (2006) Human detection using oriented histograms of flow and appearance. In: Computer vision–ECCV 2006: 9th European conference on computer vision, Graz, Austria, May 7–13, 2006. Proceedings, Part II 9. Springer, pp 428–441
86. Surasak T, Takahiro I, Cheng C-h, Wang C-e, Sheng P-y (2018) Histogram of oriented gradients for human detection in video. In: 2018 5th international conference on business and industrial research (ICBIR). IEEE, pp 172–176
87. Ojala T, Pietikainen M, Maenpaa T (2002) Multiresolution grayscale and rotation invariant texture classification with local binary patterns. IEEE Trans Pattern Anal Mach Intell 24(7):971–987
88. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60:91–110
89. Bay H, Tuytelaars T, Van Gool L (2006) Surf: speeded up robust features. Lect Notes Comput Sci 3951:404–417
90. Sykora P, Kamencay P, Hudec R (2014) Comparison of sift and surf methods for use on hand gesture recognition based on depth map. Aasri Proc 9:19–24
91. Suriya M, Sathyapriya N, Srinithi M, Yesodha V (2016) Survey on real time sign language recognition system: an lda approach. In: International conference on exploration and innovations in engineering and technology, ICEIET, pp 219–225
92. Ahmed AA, Aly S (2014) Appearance-based Arabic sign language recognition using hidden markov models. In: 2014 international conference on engineering and technology (ICET). IEEE, pp 1–6
93. Hsieh C-C, Liou D-H (2015) Novel haar features for real-time hand gesture recognition using svm. J Real-Time Image Proc 10:357–370
94. Tharwat A, Gaber T, Hassanien AE, Shahin MK, Refaat B (2015) Sift-based Arabic sign language recognition system. In: Afro-European conference for industrial advancement: proceedings of the first international afro-european conference for industrial advancement AECIA 2014. Springer, pp 359–370
95. Hartanto R, Susanto A, Santosa PI (2014) Real time static hand gesture recognition system prototype for Indonesian sign lan-

guage. In: 2014 6th international conference on information technology and electrical engineering (ICITEE). IEEE, pp 1–6

96. Yun L, Lifeng Z, Shujun Z (2012) A hand gesture recognition method based on multi-feature fusion and template matching. Proc Eng 29:1678–1684
97. Pan T-Y, Lo L-Y, Yeh C-W, Li J-W, Liu H-T, Hu M-C (2016) Real-time sign language recognition in complex background scene based on a hierarchical clustering classification method. In: 2016 IEEE second international conference on multimedia big data (BigMM). IEEE, pp 64–67
98. Rokade US, Doye D, Kokare M (2009) Hand gesture recognition using object based key frame selection. In: 2009 international conference on digital image processing. IEEE, pp 288–291
99. Bao J, Song A, Guo Y, Tang H (2011) Dynamic hand gesture recognition based on surf tracking. In: 2011 international conference on electric information and control engineering. IEEE, pp 338–341
100. Baranwal N, Nandi GC (2017) An efficient gesture based humanoid learning using wavelet descriptor and mfcc techniques. Int J Mach Learn Cybern 8:1369–1388
101. Ibrahim NB, Selim MM, Zayed HH (2018) An automatic Arabic sign language recognition system (arslrs). J King Saud Univ Comput Inf Sci 30(4):470–477
102. Chen J, Han M, Yang S, Chang Y (2016) A fingertips detection method based on the combination of centroid and Harris corner algorithm. In: 2016 17th IEEE/ACIS international conference on software engineering, artificial intelligence, networking and parallel/distributed computing (SNPD). IEEE, pp 225–230
103. Dardas N, Chen Q, Georganas ND, Petriu EM (2010) Hand gesture recognition using bag-of-features and multi-class support vector machine. In: 2010 IEEE international symposium on haptic audio visual environments and games. IEEE, pp 1–5
104. Gupta B, Shukla P, Mittal A (2016) K-nearest correlated neighbor classification for Indian sign language gesture recognition using feature fusion. In: 2016 international conference on computer communication and informatics (ICCCI). IEEE, pp 1–5
105. Huong TNT, Huu TV, Le Xuan T et al (2015) Static hand gesture recognition for Vietnamese sign language (vsl) using principle components analysis. In: 2015 international conference on communications, management and telecommunications (ComManTel). IEEE, pp 138–141
106. Ohn-Bar E, Trivedi MM (2014) Hand gesture recognition in real time for automotive interfaces: a multimodal vision-based approach and evaluations. IEEE Trans Intell Transp Syst 15(6):2368–2377
107. Chaudhary A, Raheja J (2018) Light invariant real-time robust hand gesture recognition. Optik 159:283–294
108. Wen X, Niu Y (2010) A method for hand gesture recognition based on morphology and fingertip-angle. In: 2010 the 2nd international conference on computer and automation engineering (ICCAE), vol 1. IEEE, pp 688–691
109. Shin J, Kim CM (2016) Character input system using fingertip detection with kinect sensor. In: Proceedings of the international conference on research in adaptive and convergent systems, pp 74–79
110. Meng G, Wang M (2013) Hand gesture recognition based on fingertip detection. In: 2013 fourth global congress on intelligent systems (GCIS). IEEE, pp 107–111
111. Wang M, Lin J-S, Meng GQ (2015) Fingertip detection and gesture recognition based on contour approximation. Int J Pattern Recognit Artif Intell 29(07):1555016
112. Rakthanmanon T, Campana B, Mueen A, Batista G, Keogh E (2012) Searching and mining trillions of time series subsequences under dynamic time warping. ACM
113. Ahmed W, Chanda K, Mitra S (2016) Vision based hand gesture recognition using dynamic time warping for Indian sign language. In: 2016 international conference on information science (ICIS). IEEE, pp 120–125
114. Zhi D, de Oliveira TEA, da Fonseca VP, Petriu EM (2018) Teaching a robot sign language using vision-based hand gesture recognition. In: 2018 IEEE international conference on computational intelligence and virtual environments for measurement systems and applications (CIVEMSA). IEEE, pp 1–6
115. Rabiner LR (1989) A tutorial on hidden markov models and selected applications in speech recognition. In: Proc IEEE, p 77
116. Oka K, Sato Y, Koike H (2002) Real-time fingertip tracking and gesture recognition. IEEE Comput Graph Appl 22(6):64–71
117. Chen FS, Fu CM, Huang CL (2003) Hand gesture recognition using a real-time tracking method and hidden Markov models. Image Vis Comput 21(8):745–758
118. Malgireddy MR, Nwogu I, Govindaraju V (2013) Language-motivated approaches to action recognition. Springer, Cham
119. Jebali M, Dakhli A, Jemni M (2021) Vision-based continuous sign language recognition using multimodal sensor fusion. Evolut Syst 12(4):1031–1044
120. Tutsoy O (2022) Pharmacological, non-pharmacological policies and mutation: an artificial intelligence based multi-dimensional policy making algorithm for controlling the casualties of the pandemic diseases. IEEE Trans Pattern Anal Mach Intell 44(12):9477–9488
121. Tutsoy O, Çolak Ş, Polat A, Balikci K (2020) A novel parametric model for the prediction and analysis of the covid-19 casualties. IEEE Access 8:193898–193906
122. Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297
123. Vapnik V, Vapnik V et al (1998) Statistical learning theory. Wiley, New York
124. Keerthi SS, Gilbert EG (2002) Convergence of a generalized smo algorithm for svm classifier design. Mach Learn 46(1):351–360
125. Song Y, Demirdjian D, Davis R (2012) Continuous body and hand gesture recognition for natural human-computer interaction. ACM Trans Interact Intell Syst (TiiS) 2(1):1–28
126. Trigueiros P, Ribeiro F, Reis LP (2014) Vision-based Portuguese sign language recognition system. In: New perspectives in information systems and technologies, vol 1. Springer, pp 605–617
127. Al Farid F, Hashim N, Abdullah J (2019) Vision-based hand gesture recognition from rgb video data using svm. In: International workshop on advanced image technology (IWAIT) 2019, vol 11049. SPIE, pp 265–268
128. Athira P, Sruthi C, Lijiya A (2019) A signer independent sign language recognition with co-articulation elimination from live videos: an Indian scenario. J King Saud Univ Comput Inf Sci
129. Trigueiros P, Ribeiro F, Reis LP (2013) Vision-based gesture recognition system for human-computer interaction. In: Computational vision and medical image processing IV: VIPIMAGE 2013, pp 137–142
130. Sahoo JP, Ari S, Ghosh DK (2018) Hand gesture recognition using dwt and f-ratio based feature descriptor. IET Image Proc 12(10):1780–1787
131. Maqueda AI, del-Blanco CR, Jaureguizar F, García N (2015) Human-computer interaction based on visual hand-gesture recognition using volumetric spatiograms of local binary patterns. Comput Vis Image Underst 141:126–137
132. Kubat M (1999) Neural networks: a comprehensive foundation by Simon Haykin, Macmillan, 1994, isbn 0-02-352781-7. Knowl Eng Rev 13(4):409–412
133. Haykin SS, Haykin SS (2001) Kalman filtering and neural networks, vol 284. Wiley Online Library
134. Balasundaram A, Chellappan C (2017) Vision based gesture recognition: a comprehensive study. IIOAB J 8:20–28

135. Schwenker F, Kestler HA, Palm G (2001) Three learning phases for radial-basis-function networks. Neural Netw 14(4–5):439–458
136. Ghosh DK, Ari S (2016) On an algorithm for vision-based hand gesture recognition. SIViP 10(4):655–662
137. Lafferty J, McCallum A, Pereira FC (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data
138. Laskar MA, Das AJ, Talukdar AK, Sarma KK (2015) Stereo vision-based hand gesture recognition under 3d environment. Proc Comput Sci 58:194–201
139. Jiang S, Pang G, Wu M, Kuang L (2012) An improved k-nearest-neighbor algorithm for text categorization. Expert Syst Appl 39(1):1503–1509
140. Su M-Y (2011) Using clustering to improve the knn-based classifiers for online anomaly network traffic identification. J Netw Comput Appl 34(2):722–730
141. Mejdoub M, Ben Amar C (2013) Classification improvement of local feature vectors over the knn algorithm. Multimed Tools Appl 64(1):197–218
142. Sankaranarayanan J, Samet H, Varshney A (2007) A fast all nearest neighbor algorithm for applications involving large point-clouds. Comput Graph 31(2):157–174
143. Jasim M, Zhang T, Hasanuzzaman M (2014) A real-time computer vision-based static and dynamic hand gesture recognition system. Int J Image Graph 14(01n02):1450006
144. Venkatesh, Ranjitha KV (2019) Classification and optimization scheme for text data using machine learning nave Bayes classifier. In: 2018 IEEE world symposium on communication engineering (WSCE)
145. Argyros AA, Lourakis MI (2006) Vision-based interpretation of hand gestures for remote control of a computer mouse. In: European conference on computer vision. Springer, pp 40–51
146. Kharate GK, Ghotkar AS (2016) Vision based multi-feature hand gesture recognition for Indian sign language manual signs. Int J Smart Sens Intell Syst 9(1):124
147. Misra S, Singha J, Laskar RH (2018) Vision-based hand gesture recognition of alphabets, numbers, arithmetic operators and ascii characters in order to develop a virtual text-entry interface system. Neural Comput Appl 29(8):117–135
148. Heickal H, Zhang T, Hasanuzzaman M (2015) Computer vision-based real-time 3d gesture recognition using depth image. Int J Image Graph 15(01):1550004
149. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–44
150. Yamada T, Murata S, Arie H, Ogata T (2017) Representation learning of logic words by an rnn: from word sequences to robot actions. Front Neurorobot 11:70
151. Auephanwiriyakul S, Phitakwinai S, Suttapak W, Chanda P, Theera-Umpon N (2013) Thai sign language translation using scale invariant feature transform and hidden Markov models. Pattern Recognit Lett 34(11):1291–1298
152. Neverova N, Wolf C, Paci G, Sommavilla G, Taylor G, Nebout F (2013) A multi-scale approach to gesture detection and recognition. In: Proceedings of the IEEE international conference on computer vision workshops, pp 484–491
153. Geng L, Ma X, Wang H, Gu J, Li Y (2014) Chinese sign language recognition with 3d hand motion trajectories and depth images. In: Proceeding of the 11th world congress on intelligent control and automation. IEEE, pp 1457–1461
154. Shin S, Kim W-Y (2020) Skeleton-based dynamic hand gesture recognition using a part-based gru-rnn for gesture-based interface. IEEE Access 8:50236–50243
155. Zhang L, Zhu G, Mei L, Shen P, Shah SAA, Bennamoun M (2018) Attention in convolutional lstm for gesture recognition. Advances in neural information processing systems, p 31
156. Anastassiou D, Kollias S (1988) Digital image halftoning using neural networks. In: Visual communications and image processing'88: third in a series, vol 1001. SPIE, pp 1062–1069
157. Ji S, Xu W, Yang M, Yu K (2013) 3d convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell 35(1):221–231
158. John V, Boyali A, Mita S, Imanishi M, Sanma N (2016) Deep learning-based fast hand gesture recognition using representative frames. In: 2016 international conference on digital image computing: techniques and applications (DICTA). IEEE, pp 1–8
159. Oyedotun OK, Khashman A (2017) Deep learning in vision-based static hand gesture recognition. Neural Comput Appl 28(12):3941–3951
160. Jiang D, Li G, Sun Y, Kong J, Tao B (2019) Gesture recognition based on skeletonization algorithm and cnn with asl database. Multimed Tools Appl 78(21):29953–29970
161. Kamruzzaman M (2020) Arabic sign language recognition and generating Arabic speech using convolutional neural network. Wirel Commun Mob Comput 2020
162. Chhajed RR, Parmar KP, Pandya MD, Jaju NG (2021) Messaging and video calling application for specially abled people using hand gesture recognition. In: 2021 6th international conference for convergence in technology (I2CT). IEEE, pp 1–4
163. Noreen I, Hamid M, Akram U, Malik S, Saleem M (2021) Hand pose recognition using parallel multi stream cnn. Sensors 21(24):8469
164. Al-Hammadi M, Muhammad G, Abdul W, Alsulaiman M, Bencherif MA, Mekhtiche MA (2020) Hand gesture recognition for sign language using 3dcnn. IEEE Access 8:79491–79509
165. Molchanov P, Gupta S, Kim K, Kautz J (2015) Hand gesture recognition with 3d convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 1–7
166. Molchanov P, Yang X, Gupta S, Kim K, Tyree S, Kautz J (2016) Online detection and classification of dynamic hand gestures with recurrent 3d convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4207–4215
167. Li Y, Li W, Mahadevan V, Vasconcelos N (2016) Vlad3: encoding dynamics of deep features for action recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1951–1960
168. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision, pp 4489–4497
169. Camgoz NC, Hadfield S, Koller O, Bowden R (2016) Using convolutional 3d neural networks for user-independent continuous gesture recognition. In: 2016 23rd international conference on pattern recognition (ICPR). IEEE, pp 49–54
170. Baumgartl H, Sauter D, Schenk C, Atik C, Buettner R (2021) Vision-based hand gesture recognition for human-computer interaction using mobilenetv2. In: 2021 IEEE 45th annual computers, software, and applications conference (COMPSAC). IEEE, pp 1667–1674
171. Ewe ELR, Lee CP, Kwek LC, Lim KM (2022) Hand gesture recognition via lightweight vgg16 and ensemble classifier. Appl Sci 12(15):7643
172. Jiang H, Wachs JP, Duerstock BS (2013) Integrated vision-based robotic arm interface for operators with upper limb mobility impairments. In: 2013 IEEE 13th international conference on rehabilitation robotics (ICORR). IEEE, pp 1–6
173. Kane L, Khanna P (2017) Vision-based mid-air unistroke character input using polar signatures. IEEE Trans Hum Mach Syst 47(6):1077–1088

174. Yao Y, Fu Y (2014) Contour model-based hand-gesture recognition using the kinect sensor. IEEE Trans Circuits Syst Video Technol 24(11):1935–1944
175. Almeida SGM, Guimarães FG, Ramírez JA (2014) Feature extraction in Brazilian sign language recognition based on phonological structure and using rgb-d sensors. Expert Syst Appl 41(16):7259–7271
176. Zhu C, Yang J, Shao Z, Liu C (2019) Vision based hand gesture recognition using 3d shape context. IEEE/CAA J Autom Sin 8(9):1600–1613
177. Yu C-W, Liu C-H, Chen Y-L, Lee P, Tian M-S (2018) Vision-based hand recognition based on tof depth camera. Smart Sci 6(1):21–28
178. Wu B-X, Yang C-G, Zhong J-P (2021) Research on transfer learning of vision-based gesture recognition. Int J Autom Comput 18(3):422–431
179. Starner TE (1995) Visual recognition of American sign language using hidden Markov models. Technical report, Massachusetts Inst of tech Cambridge Dept of brain and cognitive sciences
180. Hoque SA, Haq MS, Hasanuzzaman M (2018) Computer vision based gesture recognition for desktop object manipulation. In: 2018 International conference on innovation in engineering and technology (ICIET). IEEE, pp 1–6
181. Simao MA, Gibaru O, Neto P (2019) Online recognition of incomplete gesture data to interface collaborative robots. IEEE Trans Ind Electron 66(12):9372–9382
182. Nguyen V-T, Tran T-H, Le T-L, Mullot R, Courboulay V (2015) Using hand postures for interacting with assistant robot in library. In: 2015 seventh international conference on knowledge and systems engineering (KSE). IEEE, pp 354–359
183. Grzejszczak T, Legowski A, Niezabitowski M (2015) Robot manipulator teaching techniques with use of hand gestures. In: 2015 20th international conference on control systems and computer science. IEEE, pp 71–77
184. Peral M, Sanfeliu A, Garrell A (2022) Efficient hand gesture recognition for human-robot interaction. IEEE Robot Autom Lett 7(4):10272–10279
185. Shang-Liang C, Li-Wu H (2021) Using deep learning technology to realize the automatic control program of robot arm based on hand gesture recognition. Int J Eng Technol Innov 11(4):241
186. Wu B, Zhong J, Yang C (2021) A visual-based gesture prediction framework applied in social robots. IEEE/CAA J Autom Sin 9(3):510–519
187. Qi W, Ovur SE, Li Z, Marzullo A, Song R (2021) Multi-sensor guided hand gesture recognition for a teleoperated robot using a recurrent neural network. IEEE Robot Autom Lett 6(3):6039–6045
188. Torres SHM, Kern MJ, et al (2017) 7 dof industrial robot controlled by hand gestures using microsoft kinect v2. In: 2017 IEEE 3rd Colombian conference on automatic control (CCAC). IEEE, pp 1–6
189. Gao Q, Ju Z, Chen Y, Wang Q, Chi C (2022) An efficient rgb-d hand gesture detection framework for dexterous robot hand-arm teleoperation system. IEEE Trans Hum Mach Syst
190. Xue Z, Chen X, He Y, Cao H, Tian S (2022) Gesture-and vision-based automatic grasping and flexible placement in teleoperation. Int J Adv Manuf Technol 1–16
191. Fahn C-S, Chu K-Y (2011) Hidden-markov-model-based hand gesture recognition techniques used for a human-robot interaction system. In: International conference on human-computer interaction. Springer, pp 248–258
192. Wang M, Chen W-Y, Li XD (2016) Hand gesture recognition using valley circle feature and hu's moments technique for robot movement control. Measurement 94:734–744
193. Zhao H, Hu J, Zhang Y, Cheng H (2017) Hand gesture based control strategy for mobile robots. In: 2017 29th Chinese control and decision conference (CCDC). IEEE, pp 5868–5872
194. Zhang T, Su Z, Cheng J, Xue F, Liu S (2022) Machine vision-based testing action recognition method for robotic testing of mobile application. Int J Distrib Sens Netw 18(8):15501329221115376
195. Wang W, He M, Wang X, Song H, Ma J (2022) Medical gesture recognition method based on improved lightweight network. Available at SSRN 4102589
196. Xu J, Li J, Zhang S, Xie C, Dong J (2020) Skeleton guided conflict-free hand gesture recognition for robot control. In: 2020 11th international conference on awareness science and technology (iCAST). IEEE, pp 1–6
197. Togo S, Ukida H (2021) Gesture recognition using hand region estimation in robot manipulation. In: 2021 60th annual conference of the society of instrument and control engineers of Japan (SICE). IEEE, pp 1122–1127
198. Castro-Vargas J, Zapata-Impata B, Gil P, Garcia-Rodriguez J, Torres F (2019) 3dcnn performance in hand gesture recognition applied to robot
199. Almarzuqi AA, Buhari SM (2016) Enhance robotics ability in hand gesture recognition by using leap motion controller. In: International conference on broadband and wireless computing, communication and applications. Springer, pp 513–523
200. Luo X, Amighetti A, Zhang D (2019) A human-robot interaction for a mecanum wheeled mobile robot with real-time 3d two-hand gesture recognition. J Phys Conf Ser 1267:012056
201. Moysiadis V, Katikaridis D, Benos L, Busato P, Anagnostis A, Kateris D, Pearson S, Bochtis D (2022) An integrated real-time hand gesture recognition framework for human-robot interaction in agriculture. Appl Sci 12(16):8160
202. Gao Q, Chen Y, Ju Z, Liang Y (2021) Dynamic hand gesture recognition based on 3d hand pose estimation for human-robot interaction. IEEE Sens J
203. Vishwakarma DK, Maheshwari R, Kapoor R (2015) An efficient approach for the recognition of hand gestures from very low resolution images. In: 2015 fifth international conference on communication systems and network technologies. IEEE, pp 467–471
204. Tsai T-H, Huang C-C, Zhang K-L (2020) Design of hand gesture recognition system for human-computer interaction. Multimed Tools Appl 79(9):5989–6007
205. Rawat P, Kane L, Goswami M, Jindal A, Sehgal S (2022) A review on vison-based hand gesture recognition targeting rgb-d sensors. Int J Inf Technol Decis Mak
206. Chanu OR, Pillai A, Sinha S, Das P (2017) Comparative study for vision based and data based hand gesture recognition technique. In: 2017 international conference on intelligent communication and computational techniques (ICCT). IEEE, pp 26–31
207. Hasan MM, Mishra PK (2010) Hsv brightness factor matching for gesture recognition system. Int J Image Process (IJIP) 4(5):456–467
208. Xu C, Govindarajan LN, Zhang Y, Cheng L (2017) Lie-x: depth image based articulated object pose estimation, tracking, and action recognition on lie groups. Int J Comput Vis 123(3):454–478
209. Islam M et al (2020) An efficient human computer interaction through hand gesture using deep convolutional neural network. SN Comput Sci 1(4):1–9
210. Zengeler N, Kopinski T, Handmann U (2018) Hand gesture recognition in automotive human-machine interaction using depth cameras. Sensors 19(1):59
211. Liu Y, Song S, Yang L, Bian G, Yu H (2022) A novel dynamic gesture understanding algorithm fusing convolutional neural networks with hand-crafted features. J Vis Commun Image Represent 83:103454
212. Joshi G, Vig R et al (2015) A multi-class hand gesture recognition in complex background using sequential minimal optimization. In: 2015 international conference on signal processing, computing and control (ISPCC). IEEE, pp 92–96

213. Chen R, Tian X (2023) Gesture detection and recognition based on object detection in complex background. Appl Sci 13(7):4480
214. Zhang T, Lin H, Ju Z, Yang C (2020) Hand gesture recognition in complex background based on convolutional pose machine and fuzzy gaussian mixture models. Int J Fuzzy Syst 22:1330–1341
215. Vishwakarma DK (2017) Hand gesture recognition using shape and texture evidences in complex background. In: 2017 international conference on inventive computing and informatics (ICICI). IEEE, pp 278–283
216. Pabendon E, Nugroho H, Suheryadi A, Yunanto PE (2017) Hand gesture recognition system under complex background using spatio temporal analysis. In: 2017 5th international conference on instrumentation, communications, information technology, and biomedical engineering (ICICI-BME). IEEE, pp 261–265
217. Elsayed RA, Sayed MS, Abdalla MI (2015) Skin-based adaptive background subtraction for hand gesture segmentation. In: 2015 IEEE international conference on electronics, circuits, and systems (ICECS). IEEE, pp 33–36
218. Cui Z, Lei Y, Wang Y, Yang W, Qi J (2022) Hand gesture segmentation against complex background based on improved atrous spatial pyramid pooling. J Ambient Intell Humaniz Comput 1–13
219. Zhou W, Chen K (2022) A lightweight hand gesture recognition in complex backgrounds. Displays 74:102226

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.