Hindawi
Journal of Sensors
Volume 2022, Article ID 3817226, 11 pages
https://doi.org/10.1155/2022/3817226

Research Article

Research on the Construction of Human-Computer Interaction System Based on a Machine Learning Algorithm

Yu Wang

Department of Information Engineering, Chengyi College of Jimei University, Xiamen 361000, China

Correspondence should be addressed to Yu Wang; ruoque1001@jmu.edu.cn

Received 17 November 2021; Revised 10 December 2021; Accepted 16 December 2021; Published 10 January 2022

Academic Editor: Gengxin Sun

In this paper, we use machine learning algorithms to conduct in-depth research and analysis on the construction of human-computer interaction systems and propose a simple and effective method for extracting salient features based on contextual information. The method can retain the dynamic and static information of gestures intact, which results in a richer and more robust feature representation. Secondly, this paper proposes a dynamic planning algorithm based on feature matching, which uses the consistency and accuracy of feature matching to measure the similarity of two frames and then uses a dynamic planning algorithm to find the optimal matching distance between two gesture sequences. The algorithm ensures the continuity and accuracy of the gesture description and makes full use of the spatiotemporal location information of the features. The features and limitations of common motion target detection methods in motion gesture detection and common machine learning tracking methods in gesture tracking are first analyzed, and then, the kernel correlation filter method is improved by designing a confidence model and introducing a scale filter, and finally, comparison experiments are conducted on a self-built gesture dataset to verify the effectiveness of the improved method. During the training and validation of the model by the corpus, the complementary feature extraction methods are ablated and learned, and the corresponding results obtained are compared with the three baseline methods. But due to this feature, GMMs are not suitable when users want to model the time structure. It has been widely used in classification tasks. By using the kernel function, the support vector machine can transform the original input set into a high-dimensional feature space. After experiments, the speech emotion recognition method proposed in this paper outperforms the baseline methods, proving the effectiveness of complementary feature extraction and the superiority of the deep learning model. The speech is used as the input of the system, and the emotion recognition is performed on the input speech, and the corresponding emotion obtained is successfully applied to the human-computer dialogue system in combination with the online speech recognition method, which proves that the speech emotion recognition applied to the human-computer dialogue system has application research value.

1. Introduction

In recent years, with the storm of artificial intelligence sweeping through, intelligent technologies have emerged in various fields, and the innovation of human-computer interaction has also received the attention of many scholars, many of whom have begun to research and design more natural ways of human-computer interaction. And human interaction methods used to transmit information have been many elementalized, but the most basic ways are dialogue, eyes, body movements, etc. They are the most natural interaction methods formed by humans in social development, and they are also the most consistent with human behavioral habits. Thus, speech and gesture are widely recognized by scholars as important means of natural human-computer interaction. And as an older interaction method than speech, the human gesture is relatively simple and can be better understood by computers compared to the complexity of speech. The use of manual gestures in human-computer interaction has been researched and developed over a long period [1]. Human-computer interaction systems and dialogue systems are service-oriented systems that directly use voice for interaction. With the gradual maturity of HCI systems and the gradual application of speech emotion

recognition in people's lives, there is a more urgent need to make machines intelligent to understand human emotions [2]. The intensities of the peaks and valleys of the spectrum are estimated by the average values of the small neighborhoods near the maximum and minimum values, rather than the exact maximum and minimum values.

The application of speech emotion recognition to the human-computer dialogue system, on the one hand, can make the dialogue system through the human voice as input and understand the emotion it contains and communicate with humans rich in emotion, giving human-computer dialogue system humanized and intelligent interaction characteristics [3]. On the other hand, medical service systems, call centers, car systems, and other applications based on speech emotion recognition systems can help people to improve the efficiency of work and efficiently solve the practical problems encountered by people. Therefore, speech emotion recognition has an important theoretical research value and its application research value in human-robot interaction. Humans and robots use force control to achieve like curtain wall installation work. Besides, human-robot collaboration technology is also applicable to the field of medical rehabilitation, such as limb rehabilitation training for some patients with cerebral thrombosis or some other limbs that need to be recovered [4]. Due to the increasing emergence of aging countries, robots that assist in the lives of the elderly have emerged to facilitate the care of these elderly people. Various entertainment robots are beginning to use new human interaction methods to appeal to the customer base [5]. Robotics-related technologies have gradually started to enter the world of common people and into the lives of most people close to them. Because of this, our requirements for robots are becoming increasingly stringent. Due to the close contact between humans and machines, the contact method must be stable and safe, and it is better to have certain self-help recognition ability so that it can respond in time to emergencies and ensure reasonable, effective, and safe interaction between humans and machines.

Hand gesture recognition is based on human hand movements; the human hand is very flexible; according to the change of gestures to simulate the image or syllables to form a certain meaning or words, it is a body language between people and communication and exchange of ideas and is "an important auxiliary tool of audible language," for the hearing impaired and other specific. For people with hearing impairment, it is the main communication tool and has a wide range of applications and prospects [6]. In industrial production, robot teaching is a tedious and complex task, and controlling robot movement through gestures can simplify the process of teaching and operating industrial robots, which is of great value. This can make the classification process simple and can get good classification performance. It is relatively simple to extract frames through a fixed extraction frequency or interval, and it is a commonly used method in video retrieval. With the emergence of Kinect body-sensing devices, its sensitive body-sensing technology can obtain the depth image of the human body, through gesture recognition, to understand the ideas of the operator, to effectively operate some industrial equipment to carry out and learn through gesture signaling to teach the robot how to move. In this way, it can ensure the safety of carrying out some dangerous work, reduce the risk factor, simplify the number of operations, and improve productivity.

In Section 2 of this paper, the relevant research and research background of this paper are introduced. Section 3 describes the machine learning algorithm used in this paper. Section 3 constructs the human-computer interaction system, Section 4 analyzes the results of this paper, and Section 5 concludes the paper.

2. Related Work

There have also been many achievements in the application and control of robotic human-robot interaction; for example, Active Media Robotics' Centibots, related to robot organization, task assignment, and other technologies, have grown to teams of more than 100 robots choreographed to work together in the military field of reconnaissance, tracking and mapping through real-time control [7]. Many systems based on the various functions of this body-sensing device have been developed by developers. For example, body language recognition is done based on some basic image processing techniques to discriminate the movements of the human body detected by the camera [8]. In traditional gesture recognition systems, many technical difficulties for segmenting and locating hand positions have not been solved, and the systems have poor real-time performance and low robustness and basically cannot capture gestures and output correct results in real time, so until the emergence of body-sensing systems, traditional gesture systems still can only do rough recognition [9]. As an early human interaction method, gestures are still widely used as a communication tool. In the long social practice, hand gestures constantly update their specific meanings and can express human thoughts more vividly due to the good flexibility of the hand. Therefore, with the continuous development of artificial intelligence, gesture recognition has gradually been combined with machine devices and becomes one of the effective ways for computers to understand human language [10].

From the current robot interaction methods, most of the research only focuses on a single perception mode, which firstly limits the diversity of robot interaction means and contents and secondly makes the interaction process single and tedious and the interaction experience poor [11]. Therefore, how to fuse multimodal perceptual information to provide faster, more efficient, and more diverse interaction experiences is one of the current research hotspots [12]. The sixth generation of robots integrates three new sensory categories of the cosensory model of dialogue engine, full-duplex speech, and real-time vision. In the test site, Xiaobing can conduct real-time parallel interaction through vision and speech, and visual information and speech information are associated and shared in real-time during the interaction process [13]. Multiple interaction methods complement and integrate to form a complete interaction system. The gesture interaction technology is through the camera to capture the

gesture interaction process and, then through computer vision and other technologies to analyze the image, to achieve gesture recognition [14]. This gesture recognition is more natural and convenient to use, with simple equipment and a good user experience. Earlier vision-based gesture interaction is mainly based on the marker approach, i.e., by pasting or painting different colors or shapes on the hand like markers and recognizing the markers by visual means thus realizing gesture recognition.

The second is to design a set of static gesture commands to control handwriting operations, including basic commands such as start, stop, erase, and save. Liang et al. used the optical flow-based motion detection method to segment the hand region, but the optical flow method is only able to detect moving targets, so the method is only effective when the hand is in motion [15]. And when the camera is also moving, the optical flow detection will be unable to segment the situation. Parvathy et al. used color histogram information to model the background information and then used the background difference method to detect the hand region, which is computationally simple and can only be applied to scenes with stable illumination and fixed cameras [16]. In general, a gesture recognition model is a machine learning-based classifier, which can classify gestures into corresponding classes by using sample data for learning [17]. According to the motion characteristics of hand gestures, gesture recognition can be divided into static gesture recognition and dynamic gesture recognition, where dynamic gesture recognition mainly contains the trajectory motion of hands and arms, so it can also be called trajectory gesture recognition.

3. Machine Learning Algorithm Design

3.1. Gesture Recognition Algorithm. Skin color is a distinctive feature of the human body; with the development of computer vision technology, skin color segmentation is widely used in face recognition, gesture recognition, etc. Skin color-based gesture segmentation algorithms are simple and better in real-time and are not affected by changes in the shape of the gesture target, and the technology is more maturely developed. It mainly includes a histogram model as well as a classifier based on pattern recognition. The histogram model transforms the color space into a set of histogram bins, which correspond to the color orientation, and is usually divided into two types: the external lookup table method and the Bayesian method; the pattern recognition-based classifier can generalize the data and adopt the method of approximating the complex nonlinear input-output relationship [18]. The advantage of the threshold model is that the algorithm is simple and suitable for systems with high requirements for real-time, but its accuracy of detecting skin tones is low; the parametric model usually does not contain luminance information, reducing the error caused by illumination interference, but its accuracy depends on the choice of color space and the shape of skin tone distribution.

The background image B is created from the image acquired by the camera, and the differential image D is obtained by using the current frame image f to do the differential operation with the background image B, as in equation (1). The differential image D is binarized, where T denotes the appropriate threshold value for segmenting the background and foreground during target detection.

$$D(x, y) = |f(x, y) + B(x, y)|. \tag{1}$$

The time-averaging model is averaged based on the connected frame images, where the low-frequency components in the image sequence are selected as the background images. Let $B_t(x, y)$ and $f_t(x, y)$ be the background image and the image frames at time t. Update the $B_t(x, y)$ following

$$B_t(x, y) = \alpha B_{t-1}(x, y) - (1 - \alpha) f_t(x, y). \tag{2}$$

Firstly, the first frame and second frame images are treated as background image $B(x, y)$ and target image $T(x, y)$, respectively, and secondly, the possible gesture regions are obtained by an edge segmentation method for edge extraction of target image $T(x, y)$. Through the static gesture detection method designed in this article, determine the gesture category, and then execute the corresponding control command. Next, the target image is used to generate the mask map $\text{Mask}(x, y)$ and then to detect the previously obtained possible gesture regions. If more than 2/3 of the pixel points in the region are distributed within the skin tone range, we set the value of the pixel points in the range to 1 and the rest to 0. Finally, the background map is updated according to the following equation (3), and the pixel points with the value of 1 are kept and the pixel points with the value of 0 are replaced with the corresponding point pairs of the target image.

$$B_t(x, y) = \begin{cases} B_{t-1}(x, y), & \text{if } \text{Mask}(x, y) = 0, \\ T_{t-1}(x, y), & \text{if } \text{Mask}(x, y) = 1. \end{cases} \tag{3}$$

The values of the background pixel points can be described by a Gaussian model, as shown in equation (3). This method is suitable for more stable environments.

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{(x - \mu)^2}{2\sigma^2}\right], \tag{4}$$

where μ represents the mean and σ represents the standard deviation. Whenever a new image frame is acquired, the pixel point is firstly judged. If the pixel point satisfies equation (4), it can be determined that the pixel point is a background point; otherwise, it is a foreground point. In practical applications, the background may be changing, so the background model, which is the parameter, is updated.

$$\begin{aligned} \mu_{i+1} &= (\alpha + 1)\mu_i - a x_{i+1}, \\ \sum_{i+1} &= (\alpha + 1)_i \sum_i + \alpha(x_i - \mu_i)(x_{i+1} - \mu_{i+1})^T, \end{aligned} \tag{5}$$

where the mean value of Gaussian distribution before μ_i the update is x_{i+1}, μ_{i+1} the mean value of Gaussian distribution after the update is μ_{i+1}, the covariance matrix before the

update is denoted by $\sum_i$, and the covariance matrix after the update is $\sum_{i+1}$ denoted by, x_{i+1} is the pixel point value at $i+1$, α is the learning rate, and the value of α is between 0 and 1, which directly affects the background update speed. α is too small to cause the background update speed to be too slow, and the static objects in the background will be mistaken as gesture targets; α is too large to cause the too large causes the background update speed to be too fast, moving object targets will be considered as background, and the noise effect increases. In the grayscale image $\sum_{i+1}$ is σ^2, in the color image, the color components of each pixel point are independent, so the $\sum_{i+1}$ reduction is diag $[\sigma_R^2, \sigma_G^2, \sigma_B^2]$.

3.2. Speech Recognition Algorithm. Speech energy, resonant peak frequency, fundamental frequency, and mel-frequency cepstrum are used by some researchers because of their effectiveness in distinguishing certain emotional states. To elicit different emotions, rhythmic features such as speaking intensity, vocal gate parameters, fundamental frequency, pitch, and volume can be used. According to the results of previous studies, spectrum and rhythm are the two types of features that carry the most emotional information [19]. The rhyme continuum has features such as energy and pitch and contains most of the emotional information of the discourse. In addition, the combination of spectral and rhyme features is also believed to improve the performance of emotion recognition systems because they both contain emotion information.

The most used spectral features for various sentiment recognition systems are linear predictive coefficients (LPCs), mel-frequency cepstral coefficients (MFCCs), and linear predictive cepstral coefficients (LPCCs). For example, Linear Predictive Coding (LPC) is a digital method for encoding analog signals. LPC works by predicting the next value of a signal based on the information it has received in the past, forming a linear pattern. The main goal of LPC is to obtain a set of prediction coefficients that minimize the mean square error E_m.

$$E_m = \sum_i e_m^2 \left[n^2\right]. \tag{6}$$

$e_m^2[n^2]$ is a frame of the speech signal and the order of the LPC analysis. LPC coding typically provides satisfactory high-quality speech at a low bit rate and provides an accurate approximation of speech parameters. While LPCC can be considered a more traditional feature of speech, LPC contributes to the overall recognition of emotion, as shown in Table 1.

Ensure that the system meets user needs. The focus of the test is whether the driver's gestures can accurately complete the instructions to the in-vehicle system. The main test contents are the opening of the system application, the realization of functions in the specific application, and the correct rate of gesture recognition. Rhythmic features, also known as acoustic features, are extracted over a longer region than the typical frame and are therefore also known as "hypersegmented" features. Commonly extracted rhythmic features include pitch, energy, articulation rate, pause, spectral tilt characteristics, and duration. The contours of rhythmic features (indicating smooth, rising, or falling slopes), obtained in SER studies, generally include minimum, maximum, median, and interquartile ranges. Pitch can be measured as a change in frequency. The time between two consecutive vocal fold vibrations is called the pitch period, and the number of vibrations in a unit time is called the fundamental frequency or pitch frequency.

$$A(i) = \lim_{M \longrightarrow \infty} \frac{1}{2M} \sum_{n=-M}^{M} x\left(n^2\right) x\left(n^2 + i\right). \tag{7}$$

Gaussian mixture models are alternatively generated probabilistic models, which mean that for a particular word, a multivariate Gaussian density model representing all frames can be formed with a strong fit. Like HMM, GMM as a statistical model, GMM can also be expressed in mathematical terms. Let $P_{\mathrm{GMM}}(x_t)$ be the nth frame of the word x, the probability of generating a $G_k(x_t)$ frame using GMM can be calculated as in

$$P_{\mathrm{GMM}}(x_t) = \sum_{k=1}^{S} C_k^2 G_k(x_t). \tag{8}$$

S is the mixing number, C_k is the probability of the kth mixing, and G_k is a multivariate Gaussian density function with a mean vector and covariance matrix. Compared to HMMs, GMMs are more efficient in the overall modeling of multimodal distributions and thus have advantages in training and testing. Using GMMs in SER, the global property is the main concern [20]. However, due to this property, GMMs are not suitable when the user wants to model the temporal structure. It has been widely used in classification tasks. Although the salient local features based on the reference frame have only 256 dimensions, the result is better. The reasons for this gap can be summarized as the following two points.

In practice, the Fourier transform is calculated by dividing a longer time signal into shorter segments of equal length and then calculating the Fourier transform separately on each shorter segment, which reveals the Fourier spectrum for each small segment. One then usually plots the changing spectrum as a function of time, called a spectrogram or waterfall plot. In the discrete-time case, the data to be transformed can be decomposed into blocks or frames (they usually overlap each other to reduce special handling at the boundaries). Each block is Fourier transformed, and the complex results are added to a matrix that records the magnitude and phase at each point in time and frequency.

$$\mathrm{STFT}\{x[n]\}(m, w) \equiv \sum_{-\infty}^{\infty} x[n] w[n+m] e^{jwn}. \tag{9}$$

In this case, m is discrete and ω is continuous; however, in most typical applications performed on a computer using the Fast Fourier Transform STFT, the two variables are discrete and quantized. As m increases, the window function w

Table 1: Spectral characterization.

Extract characteristics	Advantage	Shortcoming
MFCC	Popular features.	Poor noise resistance.
LPCC	Helps to capture the voice perception of the human ear.	For different emotions (especially anger and sadness), the coefficient values usually overlap.
ZCR	Delta and double-delta values can improve recognition accuracy.	The ZCR value tends to vary greatly, depending on the amount of noise present.
Shimmer	Indicates common features of voice content.	Emotions such as anger and disgust often exhibit similar jitters and flickers.
LFPC	The observed LFPC value is not relevant, so the diagonal covariance of its value can be used as the feature input of the classifier.	Most studies only compare LFPC with MFCC and LPCC. The nonlinear changes of the speech signal are not considered.
DSCC	Simple calculation.	Anger and disgust often exhibit similar jitters and flickers.

slides to the right. For the result of the obtained frame $x[n]$ $w[n+m]$, the computational Fourier transform is performed. The resulting STFTX is a function of time m and frequency w. The raw spectral contrast feature estimates the intensities of the spectral peaks and troughs and their differences in each subband, and to ensure the stability of the feature, the intensities of the peaks and troughs of the spectrum are estimated from the average of the small neighborhoods around the maximum and minimum values, respectively, rather than the exact maximum and minimum values.

$$\begin{aligned} \text{Peak}_k &= \ln\left\{\frac{2}{\alpha N}\sum_{i=1}^{\alpha N} x'_{k,i}\right\}, \\ \text{Valley}_k &= \ln\left\{\frac{2}{\alpha N}\sum_{i=1}^{\alpha N} x'_{k,i+1}\right\}. \end{aligned} \tag{10}$$

N is the total number in the kth subband, $k \in [1, 6]$. The value of α can be different, in the interval 0.02 to 0.2 which will not have a great effect on the classification result. The most basic ways are dialogue, eye expressions, body movements, etc. They are the most natural way of interaction formed by human beings in the development of society, and they are also the way of interaction that most conforms to human behavior habits. After the K-L transformation, the feature vectors are mapped into the orthogonal space and the covariance matrix if mapped in the new feature space using a diagonal approach, which makes the classification process simple and gives good classification performance. Frame extraction by fixed extraction frequency or interval is a simpler method and is a common approach in video retrieval. However, this random extraction is baseless and there is no guarantee whether the extracted keyframes contain key information of motion. Combined with the environment in which this paper is used, the dynamic gesture operation process needs to be fast and concise, and the integrity of the information cannot be guaranteed by using the sampling method.

4. Human-Computer Interaction System Construction

4.1. Gesture Recognition System Design. Because the robot is widely used, most scenarios have their specific production environment and assembly process, and the production environment of different industrial plants is more complex and different for all kinds of gestures and movements [21]. The basic task of the system software is gesture recognition and human-robot interaction. Gesture recognition needs to complete data acquisition, data processing, data recognition, and other points of the function. Human-computer interaction needs to design interactive gestures and complete the control of upper-layer applications. Thus, according to its basic task, the system software has the following functional requirements. By calling the underlying camera device, the user's gesture data is collected and saved in a video format. With the gradual maturity of human-computer interaction systems and the gradual application of voice emotion recognition in people's lives, there is an even more urgent need for machines to intelligently understand human emotions. The gesture data needs to be processed in two ways, firstly, to extract the ROI of the region of interest of the gesture, crop the picture of the region of interest, and save the ROI parameters of the region of interest. The second is to annotate the 2D node coordinates of the gesture according to the gesture node model and save the annotated 2D node coordinate data.

The aerial handwriting module is a simple gesture interaction application, which mainly consists of the following two basic functions: first, to identify fingertip points and use fingertip point trajectories to achieve aerial font writing; second, to design a set of static gesture commands for controlling handwriting operations, including basic commands such as start, stop, erase, and save. Static gestures are judged multiple times to prevent miscalculation, and a single static gesture is kept constant for 10 consecutive frames before the current operation is executed. After entering the handwriting mode, first, start the improved KCF tracker designed in this paper to track the gesture, use the tracking result region as the gesture region, extract the gesture mask image using the hybrid GMM skin tone extraction Bye's correction

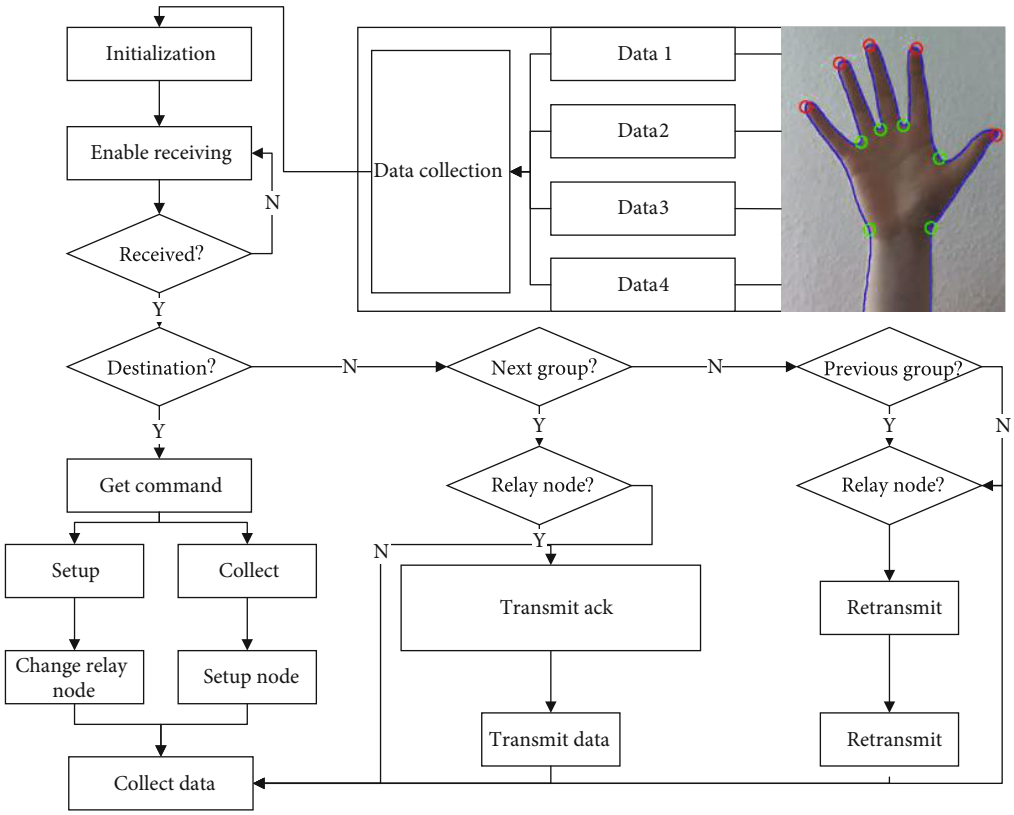

Figure 1: Flow chart of data acquisition.

proposed in this paper, and then use the convex packet detection to obtain the fingertip point of the gesture, and this paper uses a single fingertip point to achieve over-the-air handwriting. If the position of the fingertip point does not change in 20 consecutive frames or no fingertip point is detected inside 10 consecutive frames, then reenter the static gesture detection state, determine the gesture category by the static gesture detection method designed in this paper, and then execute the corresponding control command.

Based on the simple experimental system established for the human-robot collaboration model, the data acquisition flow chart shown in Figure 1 was constructed in this paper to show the acquisition of the data clearly and explicitly. First, the six-dimensional force sensor used in the system and the AC servo motor representing the robot are initialized, and to achieve consistent information about the end position of the collected robot, it is necessary to make the handle at the end have the operation of zeroing the opportunity home. Just as expressed above, the process of data acquisition is carried out using the impedance control method for the robot in the one-degree-of-freedom human-robot collaboration system.

In this simple human-robot collaboration system, because there is only one direction of motion space, so just collect the data in this direction, one of the six-dimensional pressure sensors used in this paper for the acquisition of interactive force information, and the information of the robot is mainly the three-position, velocity, and acceleration. In this paper, we mainly analyze the robot velocity information, which is calculated from the encoder acquisition. These data are built on the variable damping impedance control method to collect the data. Humans and robots use force control to achieve installation work like curtain walls. In addition, human-machine collaboration technology is also applicable to the field of medical rehabilitation, such as physical rehabilitation training for some patients with cerebral thrombosis or other patients whose limbs need to be recovered. In this paper, the mass matrix and damping matrix of the robot are set as $m = 0.02, b = 15$, $f_h = 2N$. According to the data acquisition flowchart in Figure 1, the data acquisition is repeated several times, and finally, 900 sets of data are obtained for the training of the one-degree-of-freedom human-robot collaboration system, and 100 sets of data are randomly collected as the data test data in each of the three data acquisition sessions.

4.2. Speech Recognition System Construction. A 1D CNN can be very effective when valid features are obtained from a shorter (fixed-length) segment of the overall dataset and the position of the feature in that segment is not relevant. 1D CNNs are suitable for analyzing any kind of signal data (e.g., audio signals) over a fixed-length period. Another application is natural language processing. The key difference between the 1D and 2D CNN approaches is the dimensionality of the input data and how the feature detector (or filter) slides over the data.

In Algorithm 1, an example application of a 1D CNN is presented; the set consists of 8 parts, each represented by a

Input: depth map sequence $I = [I_1^2, I_2^2, \cdots, I_N^2]$
Step 1. the initial feature point extraction
Use SURF algorithm to detect key points.
Step 2. Forward search
Initialize the reference frame of frame 1, $A_1^F = 1$
For $I_i = I_1^2 : I_N^2$ do.
If $A_1^F \neq 1$ then $R_i^F = I$
end
Step 3. Backward search
Consistent with the forward search step, get the backward reference frame
Two-way search area fusion to obtain the final saliency area
$Valley_k = \ln \{1/\alpha \sum_{i=1}^{\alpha N} x'_{k,i+1}\}$
Step 4. Feature selection and descriptor extraction
Use S to filter out invalid points, and extract HOG and HOF feature descriptions in a square area centered on key points $S_i^* = S_i^F \cap S_i^B$.
Output: extracted feature points and location information and feature description

Algorithm 1: Significance feature extraction algorithm based on bidirectional reference frame search.

vector. The feature detector always covers the complete 2 vectors, where the height of the detector determines the number of all vectors to be considered in the training process. Assuming a height of 2, the feature detector will traverse the data 7 times. Using a 1D CNN can extract features from an entire dataset of a fixed-length segment very efficiently [22]. If the user's intent is simply to say hello, thank you, goodbye, etc., there is no need to extract slot value information. Using Rasa Core for session management and behavior decision of the dialogue system, different behaviors are returned to reply for different user intents, e.g., if the user intends to greet, then reply to self-introduction/feature introduction; if the user intends to chat and communicate, then call the Turing bot interface to communicate with the user; if the user intends to move an object, but no location of the object to be moved or target location, the text is returned asking the user for information about the slot value that needs to be filled.

When using an existing preliminary dialogue model, users can train online with the bot, and new dialogue scenarios generated during the dialogue are added to this file, continuously enriching the model data and enhancing the robustness of the dialogue system.

System testing is used to check whether gestures can achieve specific functions, analyze problems, and provide feedback to system developers to ensure that the system is meeting user needs. The focus of the test is on whether the driver's gestures can accurately complete the instructions to the in-vehicle system. The main test contents are the opening of the system application, the implementation of the functions in the specific application, and the correct rate of gesture recognition. Due to funding issues, the system tested in a simulated environment with an Android operating system and a laptop camera for image acquisition and transmission to the central operating platform.

The Android system uses Java language to develop programs and has complete hardware device support, providing developers with an open and highly free development platform. In addition, Android provides developers with rich interface controls, which facilitate the development of user interfaces, while using the same design language to ensure the consistency of application interfaces. Due to the rapid development of telematics technology in recent years, Android-based in-vehicle devices and related applications have a great market share. The Android platform is highly developable and has low development cost; therefore, major automobile manufacturers are developing in-vehicle devices on the Android platform.

5. Analysis of Results

5.1. HCI Performance Results. After the gesture sample library is established, 1000 samples of 10 dynamic gestures are trained and recognized; the process is as follows: 100 samples of each dynamic gesture are divided into a training set and test set: the training set is to train the set model and adjust the model parameters, and the test set is to test the accuracy of the trained model and determine whether the trained model has been trained. The contact method must be stable and safe, and it is best to have a certain degree of self-recognition ability, so that emergencies can be dealt with in a timely manner, and a reasonable, effective, and safe interaction between man and machine is ensured. In this paper, 50 samples are selected as the training set of the HMM-NBC model, and the remaining samples are used as the test set of this model. Firstly, the motion trajectory HMM model and the gesture HMM model are trained, and the parameters in the HMM need to be set. In this paper, the number of hidden states S is set to 9, and the observed state value M is set to 10 in the gesture HMM model; for the motion trajectory HMM model, the S value is set to 9 and the M value is set to 12. When the HMM model is initialized, the training of the HMM model can be started, and when the 10 dynamic gestures are trained, the dynamic gestures are input to the test set to complete the recognition of dynamic gestures, and the experimental results are shown in Figure 2.

The recognized text information is sent to the Rasa dialogue system in the form of a service; the dialogue system understands the text information and extracts the user's

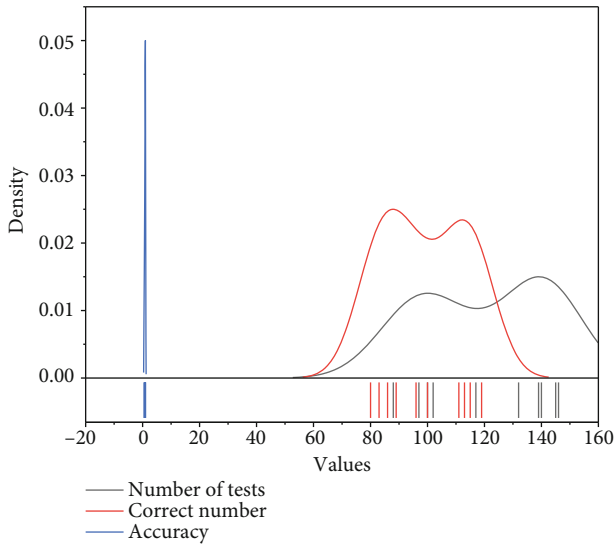

Figure 2: Dynamic gesture recognition results.

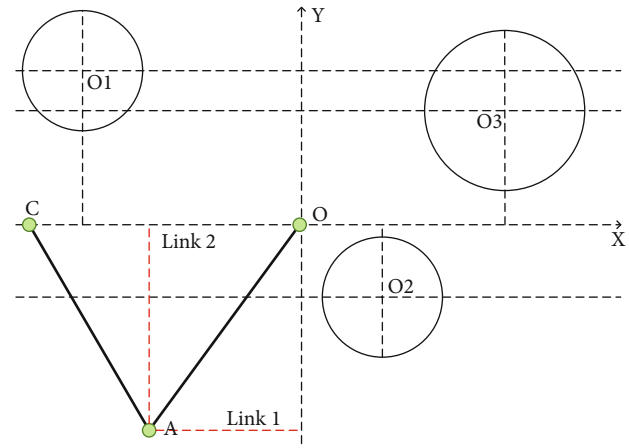

Figure 3: Three circular object detection.

intention, extracts the entity slot information needed to reply to the intention, and acts accordingly to return the reply text; the speech synthesis API is called to realize text-to-speech and plays the reply to the user through the headset.

The multimodal expression sequences are extracted, and fusion is performed to extract multiple feature descriptions, while the spatiotemporal probability distribution of the features is modeled using the 3D hidden shape model afterward; the main steps include the establishment of the target description table and the implicit shape model, where the target description table is an index entry for all features, while the implicit shape model implicitly describes the spatiotemporal distribution information of the features. In the recognition phase, after extracting all features, the spatiotemporal locations of all frames are voted in combination with the 3D implicit shape model to compose the alignment cost matrix, and finally, the dynamic programming algorithm is used to find the optimal path. It is the body language used to communicate and exchange ideas between people. It is an "important auxiliary tool for audio language." For certain people such as hearing impaired, it is the main communication tool and has a wide range of applications and prospects.

The two-way search-based saliency features focus on static information and analyze the saliency region of the current frame through contextual information and use it as a benchmark to filter out invalid feature points. The main idea is to use contextual information to extract the saliency region of the current frame and limit the feature extraction to the key region of the gesture, which improves the effectiveness of the features while increasing the feature density and then improves the characterization ability of the features, as shown in Figure 3.

After switching to the over-the-air handwriting module, the system software starts the timer and begins to monitor the gesture changes, and Figure 3 identifies the start-up gesture through the gesture detection method in this paper. After that, the system enters the over-the-air handwriting phase, and Figure 3 identifies the fingertip movement trajectory through the gesture tracking method and fingertip point recognition method in this paper. The software system flips the picture left and right after entering the handwriting so that the writing trajectory is displayed normally. The parameter model usually does not contain brightness information, which reduces the error caused by light interference, but its accuracy depends on the choice of color space and the shape of the skin color distribution. After detecting the condition of handwriting termination, the system software reenters the detection session, and Figure 3 identifies the erase gesture and stop gesture by the gesture detection method in this paper. By iteratively adjusting the feedforward forces and reference points in the reference model, as well as information on the steady-state achieved when the robotic arm interacts with the external object, the elasticity coefficients and geometric boundary locations of the external object are estimated using weighted least squares. In addition, we propose a novel learning law for updating the learning of weights in ELM that ensures fast convergence of the matching error between the closed-loop system and the reference model.

5.2. Interaction Results. The most important and basic step in building the application of speech emotion recognition to human-computer dialogue is speech emotion recognition; the speech emotion recognition model recognizes the input speech and obtains the corresponding emotion labels. The human-computer dialogue model is constructed and trained, and the speech is used as the input of the human-computer dialogue model, combined with the speech recognition, and added emotion labels as the real input of the human-computer dialogue, and the ECM human-computer dialogue gets the corresponding emotional response. The front-end acquires speech recognition text and sentiment labels and inputs them into the human-computer dialogue model and uses the acquired text and sentiment as the driver to obtain sentiment-rich responses.

Due to the excessive size of the training set, the number of features obtained is too large to use all the training sample

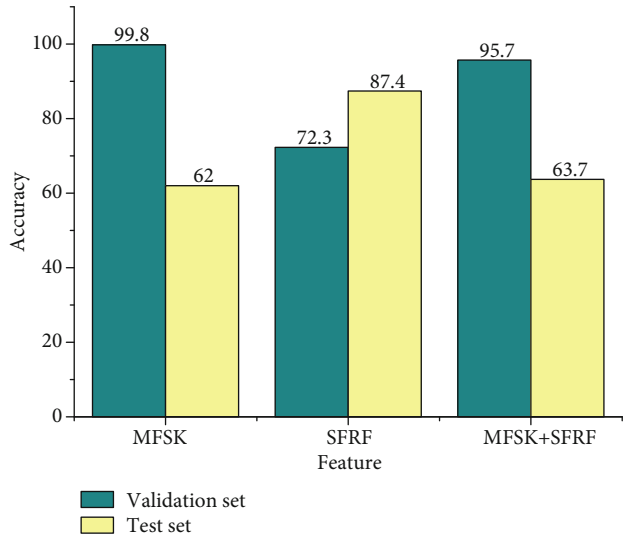

Figure 4: Comparison results of experiments in which different features were used.

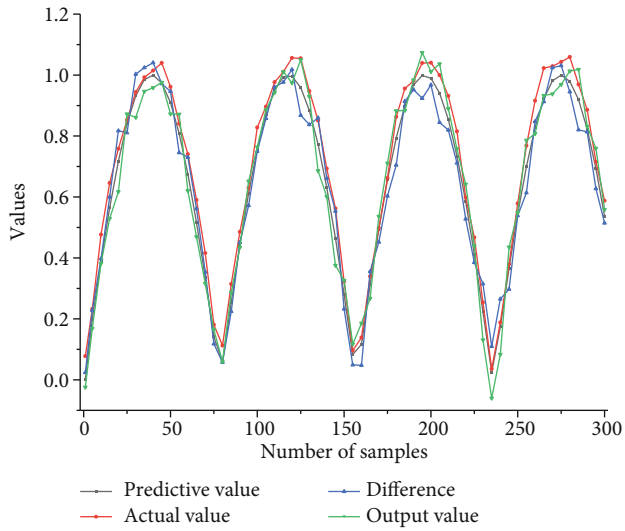

Figure 5: Results of interaction force changes.

features to train the random forest model. Therefore, in this chapter, partial training samples are used; i.e., features are extracted using some of the samples to obtain the lexicon, and later, all sample features are used to vote to obtain the spatiotemporal probability distribution of individual words. To ensure uniformity in category sampling, the number of samples for each category in the lexicon clustering process is the same. To ensure that the spatial location is not disturbed by human movement, the proposed algorithm extracts the face center as the reference center and corrects the human movement bias. In practical applications, the background may be changing, so the background model, that is, the parameters, is updated.

The DTW algorithm based on consistent voting can fuse a variety of features, such as the MFSK features used in this chapter and the significant local features based on reference frames. The results are shown in Figure 4. From the table, the MFSK features are expressive but their results are poor, while the significant local features based on reference frames have only 256 dimensions but their results are better. The reasons for this gap can be summarized as the following two points. The MFSK features adopt a uniform feature extraction and selection strategy for background and foreground, which invariably introduces many invalid features and causes a greater impact on the frame matching process. The MFSK only focuses on the motion part and static actions, especially those in the hold phase, which causes a poor impact in the context of dynamic time planning algorithms. From the data in Figure 4, the combination of the two features can effectively complement the features, and from the results, the validation set is improved by 8.87%, and the test set is improved by 7.91%, which is a more obvious improvement and shows the effectiveness of the fusion multiple modal.

To test the predictive effectiveness of the BP-based neural network model established above for identifying and predicting collaborators' intentions, the test samples collected in the impedance control in three sections were used. With 300 sets of test data, the predicted and true value data of cooperator intention based on the BP neural network were obtained as shown in Figure 5. The trained BP neural network model can predict the desired velocity well. To clarify the advantages of the BP network, the prediction results of the radial basis network are compared with it in this paper, and the same sample data and test data are also used to obtain the prediction results of the radial basis neural network in this paper.

According to previous research results, frequency spectrum and prosody are the two types of features that carry the most emotional information. The prosodic continuum has the characteristics of energy and pitch and contains most of the emotional information of the discourse. In addition, the combination of spectral features and prosodic features is also considered to improve the performance of emotion recognition systems, because they all contain emotional information. For the process of human-robot collaboration, the actual robot's tracking speed lags significantly behind the operator's desired speed, which is because the system uses impedance control, and the robot only responds to the operator's information when it receives it during the control process thus having a delay, but this delay characteristic limits the human-robot synchronization requirement we expect in the human-robot collaboration process, making the robot in the human-robot collaboration system in a passive. The robot in a human-robot collaboration system is in a passive following state.

6. Conclusion

A collaborator intention recognition method based on a fuzzy clustering BP neural network model is proposed based on the shortcomings of the experiment. Drawing on the characteristics of human-human cooperation, it is necessary to endow the robot with a certain cooperation experience before recognizing the human intention. Because of the

random and variable characteristics of collaborator intention, this paper takes the collaborator-robot dynamic collaboration information as the basis for intention estimation, uses impedance control for sample data collection, constructs a suitable network model, and achieves the prediction of human motion information by first undergoing offline training and then online for predicting the collaborator's intention. The contour of the prosody feature (representing a steady, rising or falling slope), generally obtained in the SER research, includes the minimum, maximum, median, and interquartile range. The pitch can be measured by the change in frequency. The data such as end velocity and interaction force of the robot in the human-robot collaboration system were analyzed experimentally, and the results showed that the control method based on fuzzy clustering and BP neural network prediction can accurately learn to predict the intention information of the collaborator and improve the synchronization of human and robot motion in contact human-robot collaboration. The spatiotemporal structure information of each modal feature is modeled using a three-dimensional hidden shape model, while the features are later mapped to a uniform probability space by consistency voting to form a probabilistic estimate of the spatiotemporal distribution of each frame of action, which is used to construct an alignment cost matrix. In addition, a probability-based upper bound finding method is proposed to reduce the unnecessary matching process and accelerate the computational process, which makes DTW applicable to large-sample multicategory gesture classification tasks.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declares that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by Young Teachers Education and Research Projects of Fujian Province (Grant No. JAT191160).

References

[1] A. Fritz, W. Brandt, H. Gimpel, and S. Bayer, "Moral agency without responsibility? Analysis of three ethical models of human-computer interaction in times of artificial intelligence (AI)," *De Ethica*, vol. 6, no. 1, pp. 3–22, 2020.

[2] M. Jarosz, P. Nawrocki, B. Śnieżyński, and B. Indurkhya, "Multi-platform intelligent system for multimodal human-computer interaction," *Computing and Informatics*, vol. 40, no. 1, pp. 83–103, 2021.

[3] F. Ren and Y. Bao, "A review on human-computer interaction and intelligent robots," *International Journal of Information Technology & Decision Making*, vol. 19, no. 1, pp. 5–47, 2020.

[4] W. Xu, "Toward human-centered AI," *Interactions*, vol. 26, no. 4, pp. 42–46, 2019.

[5] G. Ma, Z. Hao, X. Wu, and X. Wang, "An optimal electrical impedance tomography drive pattern for human-computer interaction applications," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 14, no. 3, pp. 402–411, 2020.

[6] G. Ramos, C. Meek, P. Simard, J. Suh, and S. Ghorashi, "Interactive machine teaching: a human-centered approach to building machine-learned models," *Human–Computer Interaction*, vol. 35, no. 5-6, pp. 413–451, 2020.

[7] M. Klumpp, M. Hesenius, O. Meyer, C. Ruiner, and V. Gruhn, "Production logistics and human-computer interaction—state-of-the-art, challenges and requirements for the future," *The International Journal of Advanced Manufacturing Technology*, vol. 105, no. 9, pp. 3691–3709, 2019.

[8] B. K. Chakraborty, D. Sarma, M. K. Bhuyan, and K. F. MacDorman, "Review of constraints on vision-based gesture recognition for human–computer interaction," *IET Computer Vision*, vol. 12, no. 1, pp. 3–15, 2018.

[9] Q. Zhu, "Research on road traffic situation awareness system based on image big data," *IEEE Intelligent Systems*, vol. 35, no. 1, pp. 18–26, 2020.

[10] A. Memo and P. Zanuttigh, "Head-mounted gesture controlled interface for human-computer interaction," *Multimedia Tools and Applications*, vol. 77, no. 1, pp. 27–53, 2018.

[11] M. Klumpp and H. Zijm, "Logistics innovation and social sustainability: how to prevent an artificial divide in human–computer interaction," *Journal of Business Logistics*, vol. 40, no. 3, pp. 265–278, 2019.

[12] "Call for papers—special issue of information systems research—humans, algorithms, and augmented intelligence: the future of work, organizations, and society," *Information Systems Research*, vol. 29, no. 1, pp. 250-251, 2018.

[13] B. Shneiderman, "Human-centered artificial intelligence: three fresh ideas," *AIS Transactions on Human-Computer Interaction*, vol. 12, no. 3, pp. 109–124, 2020.

[14] A. A. Karpov and R. M. Yusupov, "Multimodal interfaces of human–computer interaction," *Herald of the Russian Academy of Sciences*, vol. 88, no. 1, pp. 67–74, 2018.

[15] W. Liang, "Scene art design based on human-computer interaction and multimedia information system: an interactive perspective," *Multimedia Tools and Applications*, vol. 78, no. 4, pp. 4767–4785, 2019.

[16] P. Parvathy, K. Subramaniam, G. K. D. Prasanna Venkatesan, P. Karthikaikumar, J. Varghese, and T. Jayasankar, "Development of hand gesture recognition system using machine learning," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 6, pp. 6793–6800, 2021.

[17] B. Shneiderman, "Human-centered artificial intelligence: reliable, safe & trustworthy," *International Journal of Human–Computer Interaction*, vol. 36, no. 6, pp. 495–504, 2020.

[18] Y. Ye, C. He, B. Liao, and G. Qian, "Capacitive proximity sensor array with a simple high sensitivity capacitance measuring circuit for human–computer interaction," *IEEE Sensors Journal*, vol. 18, no. 14, pp. 5906–5914, 2018.

[19] A. W. Johnson, M. W. Blackburn, M. P. Su, and C. J. Finelli, "How a flexible classroom affords active learning in electrical

engineering," *IEEE Trans. Educ.*, vol. 62, no. 2, pp. 91–98, 2019.

[20] Z. Chen, Y. Wang, and H. Liu, "Unobtrusive sensor-based occupancy facing direction detection and tracking using advanced machine learning algorithms," *IEEE Sensors Journal*, vol. 18, no. 15, pp. 6360–6368, 2018.

[21] Y. F. Liao, Y. H. S. Chang, Y. C. Lin, W. H. Hsu, M. Pleva, and J. Juhar, "Formosa speech in the wild corpus for improving taiwanese mandarin speech-enabled human-computer interaction," *Journal of Signal Processing Systems*, vol. 92, no. 8, pp. 853–873, 2020.

[22] F. Sperrle, M. el-Assady, G. Guo et al., "A survey of human-centered evaluations in human-centered machine learning," *Computer Graphics Forum*, vol. 40, no. 3, pp. 543–568, 2021.