





Deep Learning for Intelligent Human–Computer Interaction

Zhihan Lv ^{1,*}, Fabio Poiesi ², Qi Dong ³, Jaime Lloret ⁴ and Houbing Song ⁵¹ Department of Game Design, Faculty of Arts, Uppsala University, SE-62167 Uppsala, Sweden² Technologies of Vision, Digital Industry Center, Fondazione Bruno Kessler, Via Sommarive 18, 38123 Trento, Italy³ Amazon AWS AI, Seattle, WA 98125, USA⁴ Instituto de Investigacion para la Gestion Integrada de Zonas Costeras, Universitat Politecnica de Valencia, 46022 Valencia, Spain⁵ Security and Optimization for Networked Globe Laboratory (SONG Lab), Embry-Riddle Aeronautical University, Daytona Beach, FL 32114, USA

* Correspondence: zhihan.lyu@speldesign.uu.se

Abstract: In recent years, gesture recognition and speech recognition, as important input methods in Human–Computer Interaction (HCI), have been widely used in the field of virtual reality. In particular, with the rapid development of deep learning, artificial intelligence, and other computer technologies, gesture recognition and speech recognition have achieved breakthrough research progress. The search platform used in this work is mainly the Google Academic and literature database Web of Science. According to the keywords related to HCI and deep learning, such as “intelligent HCI”, “speech recognition”, “gesture recognition”, and “natural language processing”, nearly 1000 studies were selected. Then, nearly 500 studies of research methods were selected and 100 studies were finally selected as the research content of this work after five years (2019–2022) of year screening. First, the current situation of the HCI intelligent system is analyzed, the realization of gesture interaction and voice interaction in HCI is summarized, and the advantages brought by deep learning are selected for research. Then, the core concepts of gesture interaction are introduced and the progress of gesture recognition and speech recognition interaction is analyzed. Furthermore, the representative applications of gesture recognition and speech recognition interaction are described. Finally, the current HCI in the direction of natural language processing is investigated. The results show that the combination of intelligent HCI and deep learning is deeply applied in gesture recognition, speech recognition, emotion recognition, and intelligent robot direction. A wide variety of recognition methods were proposed in related research fields and verified by experiments. Compared with interactive methods without deep learning, high recognition accuracy was achieved. In Human–Machine Interfaces (HMIs) with voice support, context plays an important role in improving user interfaces. Whether it is voice search, mobile communication, or children’s speech recognition, HCI combined with deep learning can maintain better robustness. The combination of convolutional neural networks and long short-term memory networks can greatly improve the accuracy and precision of action recognition. Therefore, in the future, the application field of HCI will involve more industries and greater prospects are expected.

Keywords: human–computer interaction; deep learning; speech recognition; gesture recognition; emotion recognition



Citation: Lv, Z.; Poiesi, F.; Dong, Q.; Lloret, J.; Song, H. Deep Learning for Intelligent Human–Computer Interaction. *Appl. Sci.* **2022**, *12*, 11457. <https://doi.org/10.3390/app122211457>

Academic Editor: Antonio Fernández-Caballero

Received: 10 October 2022

Accepted: 8 November 2022

Published: 11 November 2022

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

With the progress of science and technology, many pioneers of technology are trying to combine voice, vision, text, and other information, that is, multimodal information, to promote the upgrade of Human–Computer Interaction (HCI) technology. Multimodal interaction has also become a hot topic in academia and industry [1]. Multimodal technology will not be limited to speech and visual recognition but will gradually change the

whole world in this revolution. For example, lip recognition, speech recognition, speech translation, speech synthesis, and several industry-leading multimodal interaction basic technologies have been applied in various industries. Gesture interaction technology, as a command, is transformed into a language that can be recognized by computers by capturing the movements of human hands and limbs. It has become another important method of HCI after keyboards, mice, and touch screens [2–4]. In terms of intelligent hardware, the mainstream method in the industry is processing signals by microphone arrays and eliminating noise by hardware. However, when the environment is complex and noisy, there is still a large bottleneck in speech recognition [5–7]. The next generation of revolutionary HCI technology may not impact the whole industry, such as the emergence of graphical interfaces and touch technology, but may use data-driven intelligence to realize the potential revolution of HCI [8]. The vigorous development of artificial intelligence has greatly promoted the intelligence of machines, and the in-depth study of the interaction between humans and machines has promoted new gesture interaction technology and automatic speech recognition technology [9–11].

HCI is the product of interdisciplinary research. The concept of HCI was first proposed in 1975 and the professional name appeared in 1980 [12–14]. With the popularization of the concept of HCI, research on HCI is increasing daily. Human–machine interaction is simply “the way people and machines deal with each other” [15]. With the rise of deep learning technology, the research process of HCI has further accelerated [16–18]. Human–robot fingertip communication has gradually shifted from command communication to emotional communication, and there are also some difficulties and challenges in the evolution of this interaction. For example, voice and gesture, as an entry method of virtual reality, are invading our life and are crucial to the application of interaction [19]. Different from the physical world, humans and machines in the virtual world are no longer limited to the objective laws of the physical world, and the logic of HCI is completely different [20–22]. HCI in the virtual world can develop a high-dimensional perspective of information for people and a broader dimension of receiving information. At the same time, it can also expand our access to information and experience through artificial ways. However, the virtual world has higher requirements for gesture interaction and voice interaction. The main problems at present are as follows. The first is how the dialog robot used in speech recognition can effectively recognize environmental noise and real interactive sounds and how the machine can better understand human language through machine learning and artificial intelligence technology [23,24]. The second is that the problem of gesture recognition lies in how to accurately identify which of the continuous movements are unconscious and which are truly conscious interactive movements. The third is which functions in the HCI system are more suitable for gesture recognition. The fourth is how deep learning technology can further improve the accuracy of action capture and recognition of interactive gestures. As a result of these problems, it is no longer humans adapting to machines in the future dialog robot product form but machines adapting to humans. Dialog robot products based on artificial intelligence technology will gradually become mainstream [25–27].

Regarding the existing problems, nearly 1000 studies were screened in this work based on keywords related to HCI and deep learning, such as “intelligent HCI”, “speech recognition”, “gesture recognition”, and “natural language processing”, through the Google academic and literature database Web of Science. Then, nearly 500 studies of research methods were selected and approximately 100 studies were finally selected as the research content of this work after five years (2019–2022) of year screening. The application status of intelligent HCI in deep learning in various industries is studied, such as gesture recognition, speech interaction, and natural language processing. In this work, the understanding of speech interaction and gesture interaction in a Virtual Reality (VR) environment in HCI is summarized and analyzed, as well as the application of natural language processing in search and chat bots. This work focuses on the improvement of dynamic gesture

understanding by deep learning technology, which can provide a reference for the future development of HCI.

2. Adoption Status of Deep Learning in Intelligent HCI

HCI mainly analyzes the exchange of information between human actions and computers. HCI is a comprehensive research field associated with cognitive psychology, ergonomics, multimedia, and VR [28]. The information exchange of HCI relies on interactive devices, including human–computer interactive devices and computer–human interactive devices [29–31]. HCI devices include the keyboard, mouse, joystick, joystick, data suit, position tracker, data glove, and pressure pen. Computer–human interaction devices include printers, plotters, monitors, helmet-mounted monitors, and speakers. The progression process of HCI involves voice interaction technology, image recognition, Augmented Reality (AR), and VR, as well as somatosensory interaction technology, which has become popular in recent years [32–34]. Among the four types of technologies, voice interaction is the one with the highest input efficiency and the most natural interaction mode, which can easily broaden the adoption scenarios of products. Image recognition is applied in the field of automatic driving and security for road condition recognition and human face recognition. AR and VR provide immersive experiences, not only for interaction but also for display and movement [35,36]. People can directly use their body movements to interact with the surrounding devices or environment via motion sensing without using any complex control devices so that people can interact in an immersive way. HCI is changing with the progression of science and technology (Figure 1).

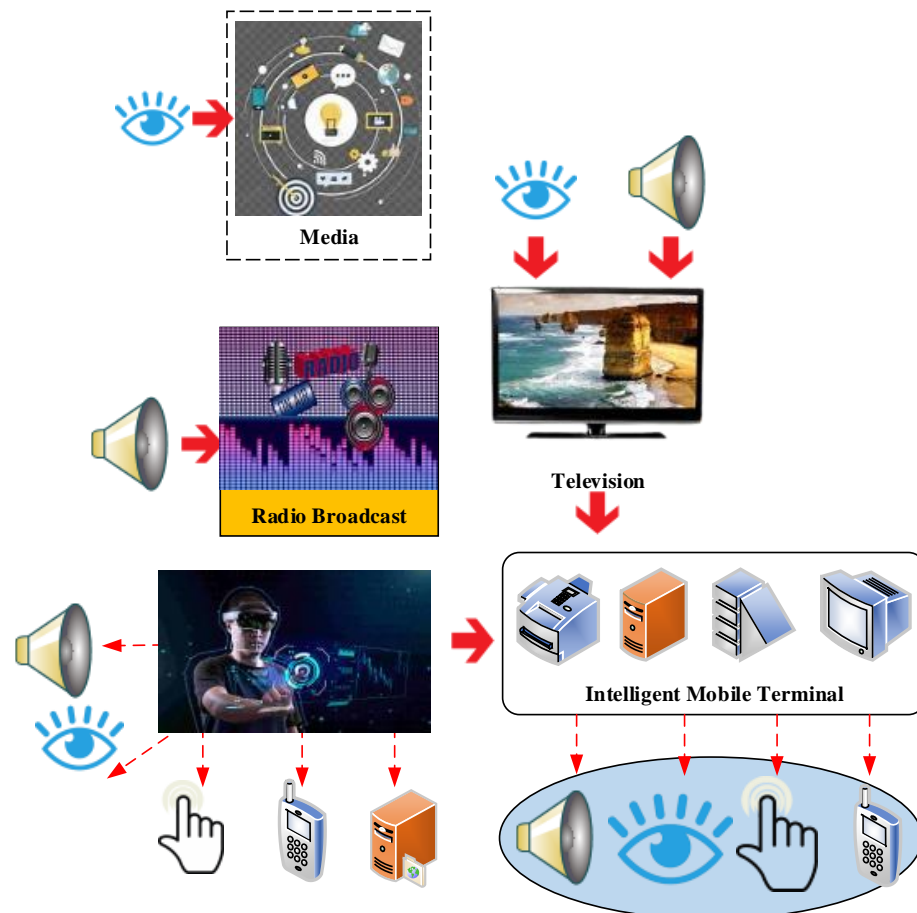


Figure 1. The history of HCI.

Deep learning is a new research direction in machine learning. In recent years, it has achieved relatively successful development in image recognition and retrieval, language information processing, and speech recognition [37–39]. Other approaches include context-aware systems, behavior synthesis in user modeling studies or embedded dialog agents, and natural speech processing, all of which rely on deep learning to support human interactions with intelligent systems. Deep learning adoptions are based on building models that mimic the neural connections of the human brain, which processes images, sound, and text signals when data features are described hierarchically through multiple transformation stages, and then data interpretation is provided [40]. It is a technology that enables machine learning, which today is mostly neural networks. Neural networks are inspired by the human brain and are the interconnections between neurons [41–43]. The adoption of deep learning in HCI can not only improve the accuracy of speech recognition and image recognition but also enhance the realism of interaction [44]. Language understanding is a technique that explores the language problem of HCI. Unlike speech recognition, which converts speech signals into text or corresponding commands, language understanding involves making machines understand human language. People use the computer in the language they are most accustomed to and no longer need to spend time and energy learning various computer languages, therefore language understanding technology is relatively difficult. Sensors have gradually been widely used due to the trend of environmental digitization brought by the Internet of Things (IoT) technology [45–47]. Content media, objects in the environment, and humans themselves are all going through a digitalized process. Interaction design is particularly important and how to create and form a natural HCI will become an important proposition. Whether it is system adoptions, intelligent devices, and scenes in the environment, it will tend to be a more natural, easy, and humanized HCI, which appears when the user has an appropriate help guide and therefore does not require the user to have rigid memory function or immediate operation understanding. The research framework of intelligent HCI design is shown in Figure 2.

In the field of advertising and information transmission, touch screens are an important form of HCI. Digital signage integrating audio, video, and image information displays have become an important tool of HCI. In this process, its design, information output, and user interaction also become more flexible. Shen et al. (2019) [48] pointed out that media has changed the way people communicate with their friends. Whether it is a self-service machine, transportation information display screen, or shopping mall marketing display screen, the HCI requirements inspired by different scenarios have similar demands for hardware product solutions. Shen et al. (2019) [49] applied a text mining method called two-level conceptual link analysis. In the traditional HCI, the keyboard is an indispensable part, which also causes certain limitations to the adoption scenarios. Embedded computer hardware, as important basic hardware for the construction of HCI scenes, also presents many possibilities. Obviously, intelligent HCI combined with deep learning is deeply applied in gesture recognition, speech recognition, emotion recognition, and natural language processing. A variety of recognition methods are proposed and verified through experiments, which can achieve high recognition accuracy. Therefore, applying deep learning in HCI design can broaden the application prospects.

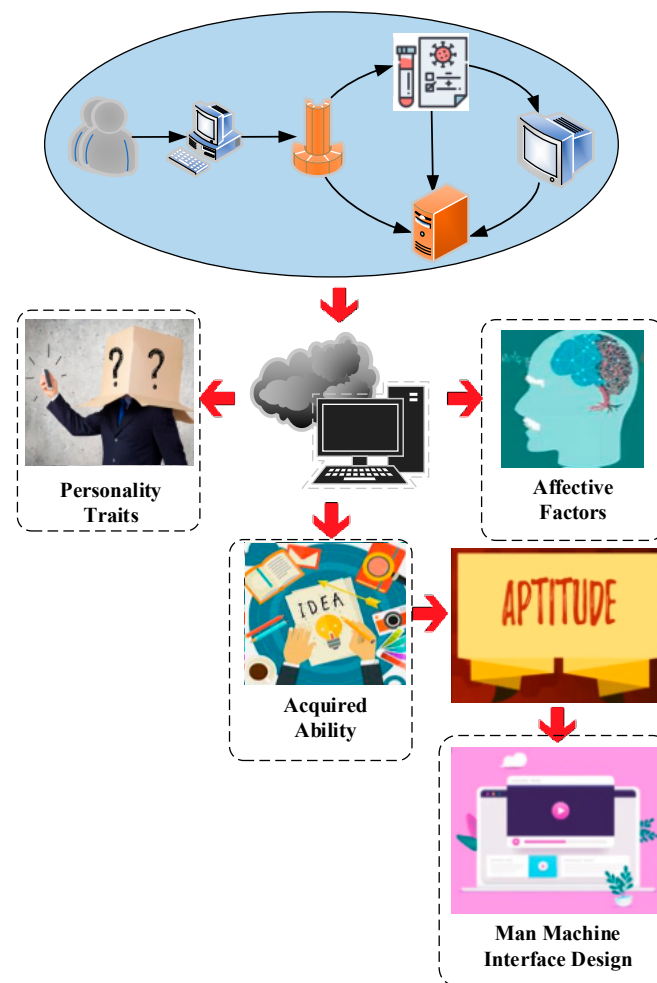


Figure 2. Research framework of intelligent HCI design.

3. Application of Deep Learning in HCI Intelligent Systems

HCI refers to the exchange of information between people and computers, including the computer providing information to people through output or display devices and people inputting relevant information to the computer through input devices. Multimodal simulation is the concrete 3D virtual realization of the situational environment and co-existing agent and the most prominent content represented by communicative behavior in discourse. Pustejovsky and Krishnaswamy (2021) [50] believed that embodiments act as an imperative part of the design and modeling of systems developed for HCI. This work describes VoxWorld, a simulation platform for building HCIs. The platform supports a multimodal dialog system that communicates through language, gesture, action, facial expression, and gaze tracking in a task-oriented interactive environment. With the continuous development of sensor technology, the acquisition cost of depth images is decreasing. Gesture recognition under depth images and red–green–blue (RGB) images has gradually become a research direction in pattern recognition. However, most of the current deep gesture image processing methods are relatively simple, ignore the relationship and influence between the two modes, and fail to make full use of the related factors between different modes. To solve the above problems, Duan et al. (2021) [51] optimized the effect of depth image information processing by considering the independent features and related features of multimodal data and constructed an adaptive weight algorithm for the fusion of different features. Simulation results showed that the proposed method was superior to the traditional deep gesture image processing method and the gesture recognition rate was higher. The proposed method also achieved higher recognition accuracy than that of other

advanced gesture recognition methods, which verified the feasibility and robustness of the proposed method. These two studies indicate that multimodal image acquisition through deep learning can improve the accuracy of gesture recognition in HCI systems.

The same application effect is also reflected in the context-aware system. For example, Wang et al. (2018) [52] used deep learning as a data-driven technology for continuous human motion analysis and human-machine cooperation demand prediction in future intelligent manufacturing to improve the planning and control of robots in completing shared tasks. The feasibility of engine assembly was verified by numerical examples, which met the requirements. Similarly, Wang et al. (2020) [53] proposed a context-aware citation recommendation model under an end-to-end memory network. The model uses bidirectional long short-term memory (Bi-LSTM) to learn the representation of paper and citation context. Furthermore, the superior performance of the model was proven by experiments on different datasets. In addition to context-aware intelligent HCI systems, as user modeling research suggests, deep learning is also widely used in user modeling based on historical interaction matrices and recommendation systems under matching function learning. The existing deep learning-based recommendation methods usually use the user's historical interaction terms to perform static user preference modeling. Wang et al. (2022) [54] adopted the time-aware deep learning framework in their research to model dynamic user preferences via an attention mechanism and predict matching scores based on deep learning. It significantly and consistently outperformed the existing time-aware and deep learning-based recommendation methods in the top-k recommendation.

Since HCI covers a wide range, the research literature has rich and multidisciplinary content, with limited studies showing the big picture of the field. Such analyses provide researchers with a better understanding of the field, revealing current issues, challenges, and potential research gaps. Gurcan et al. (2021) [55] discussed the research trend of the development stage of HCI research in the past 60 years. The results revealed 21 major themes that delineate the future of HCI research. The topics were analyzed by extending the found topics beyond the snapshot, considering their stage of development, number, and acceleration to provide a panoramic view showing trends increasing and decreasing over time. In this context, the shift of HCI research from machine-oriented systems to human-oriented systems indicates its future direction toward up-context sensing adaptive systems. Chhikara et al. (2020) [56] combined joint learning with emotion analysis to create an advanced, simple, safe, and efficient HCI system for emotion monitoring. In this study, facial expressions and voice signals were combined to find macro expressions and create an emotion index. The index is monitored to determine users' mental health. The data collected from users are monitored to analyze the users' mental health and provide counseling solutions in the trough period, which has achieved a good treatment effect for users. In the field of artificial intelligence, HCI technology and its related intelligent robot technology are indispensable and interesting contents. Regarding the software algorithms and hardware systems, the above techniques study and attempt to build a natural HCI environment. Ren and Bao (2020) [57] provided an overview of HCI and intelligent robots in their study. This study highlighted existing technologies for listening, speaking, reading, writing, and other senses that are widely used in human interaction and capable of providing solutions to some of the big challenges of HCI and intelligent robotics. Hence, the performance of deep learning methods in different HCI intelligent systems is obviously better than that of the unused systems.

4. Development Status of Intelligent Voice Interaction System

AI-based human-computer voice interaction technology makes previously tedious work simple and easy to operate and greatly simplifies some steps of people's daily lives. From the birth of the earliest personal voice assistant to the present representative HCI products, such as intelligent speakers, these AI products are not only symbolic representatives of scientific and technological progress but also improve people's quality of life. In the beginning, when using cell phones, people just made calls and sent messages, but now peo-

ple can also communicate with cell phones by voice. With traditional speakers, people just wanted to listen to music to please them. AI speakers in the 21st century can not only meet the needs of users to listen to music but can also have all kinds of conversations with users. Voice interaction is an important and convenient method in an HCI system. It accesses the user's input information starting from the whole interaction system, including voice, face, and multimodal emotion-related information. The input information can be understood in the dialog system, and the output can be generated later through this dialog part. Finally, text can also be displayed by speech synthesis, among which the most important part is the speech part and the dialog system part, which is the whole process of speech interaction (Figure 3).

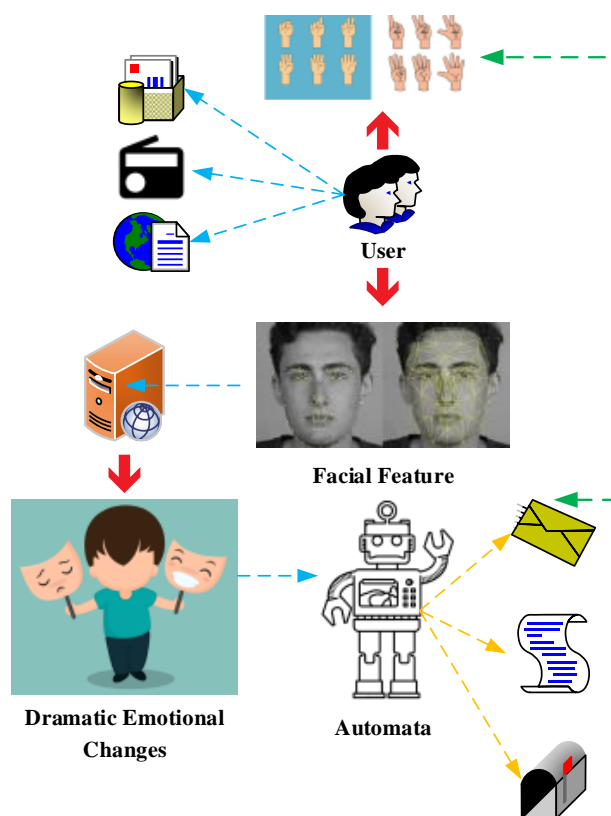


Figure 3. Man-machine voice interaction process.

Traditional HCI methods are unable to meet peoples' research and development needs for artificial intelligence and the level of interaction between people and existing artificial intelligence products in their lives is not deep enough. For example, for intelligent speakers, the appearance of this kind of product to a certain extent improves peoples' quality of life, but it can only meet the needs of users and machines in a single round of interaction. To further improve the level of HCI, researchers adopted deep learning in it. For example, deep learning is combined with traditional methods and applied to human-computer speech interaction systems to realize a variety of deep HCI mechanisms.

Speech recognition has made significant progress in the last decade. Increasing progress was made in end-to-end automatic speech recognition (ASR), which transcribes speech to text without any pretrained alignment. The ASR architecture based on hybrid connectionism temporal categorization and attention takes advantage of both. The performance of the hybrid time classification and attention ASR system is comparable to that of the hidden Markov model ASR system. However, deploying a hybrid temporal classification and attention system for online speech recognition is still a problem. Miao et al. (2020) [58] introduced an online mixing time classification and attention end-to-end ASR.

This architecture replaces all the offline components of the traditional online hybrid time classification and attention ASR architectures with corresponding streaming components. Experiments showed that the proposed online hybrid time classification and attention model outperforms state-of-the-art competitors by enhancing the real-time detection of HCI service.

To address the issue of the difficulty of recognizing interactions caused by different languages and accents, Liao et al. (2020) [59] established the Formosa Speech Project, an ancient name for Taiwan given by the Portuguese, to collect a large scale of Taiwanese Mandarin to promote the development of Taiwan-specific speech project technology and held the Formosa Plastic Speech Recognition Challenge to promote the corpus. The performance of the existing Taiwan-specific voice project system was also evaluated. In addition to the language used, the most important thing for speech recognition is the understanding of context. One of the most important components of context is the emotion in the speaker's voice. Emotion recognition provides a prior for human decision processing, interaction, and cognitive processes, making it possible to input human-like features to the HMI, such as empathy and responding with appropriate emotions in the text-to-speech engine. Speech emotion recognition refers to extracting the emotional state of the speaker. The main purpose of emotional speech recognition is to adjust the system response to detect the frustration or annoyance of the voice. Ho et al. (2020) [60] proposed a multi-modal approach for speech emotion recognition via multilevel multi-head fusion attention and a recursive neural network. For the audio functionality, the mel-frequency cepstrum coefficient is determined from the raw signal using the OpenSMILE toolbox, and then the text information is embedded using a pre-trained model of the bidirectional encoder representation from the converter. These features are fed in parallel to the self-attention mechanistic base RNN to exploit the context of each timestamp, using multi-head attention techniques to fuse all representatives to predict emotional states. The experimental results on three databases show that the combination of interactive emotion action capture, a multimodal emotion line dataset, and multimodal opinion emotion and emotion intensity achieves better performance than the single model, which also shows the important role of recurrent neural networks in speech emotion recognition.

Similarly, to study the context in speech recognition, Hazer-Rau et al. (2020) [61] proposed a multimodal dataset for the study of sentiment computing obtained in the HCI environment. Experimental movement and interaction scenarios were designed and implemented based on the generic paradigm of gamification for inducing dialog-based HCI-related emotional and cognitive load states. Based on the HCI scenario in the study, 16 sensor patterns were recorded based on the multimodal emotion corpus uulmMAC of the University of Ulm, resulting in the final uulmMAC dataset of 57 subjects and 95 recorded sessions. It was found that the evaluation of reported subjective feedback showed significant differences between the series, very consistent with the induced state, and the questionnaire analysis showed stable results. In summary, the uulmMAC database is a valuable contribution to affective computing and multimodal data analysis, captured in mobile interaction scenarios that approximate real HCI. It consists of massive subjects and allows investigation across time and space. It verifies and checks for quality issues through subjective feedback and can be used in emotional computing and machine learning adoptions.

Despite some achievements in context recognition, speech recognition is still challenging due to the difficulty of adapting to new languages, dealing with changes in speech datasets, and overcoming distortion factors. Deep learning systems can overcome these challenges by using training algorithms, such as gradient descent optimization, using depth maps with multiple processing layers, and using high-level abstractions in datasets. Dokuz and Tufekci (2021) [62] proposed four strategies to select small batch samples to represent the variation of each feature in the speech recognition task in the dataset to improve the model performance of deep learning-based speech recognition. Experimental results showed that the proposed strategy was more effective than the standard small-batch

sample selection strategy. Similarly, Sun et al. (2020) [63] studied the effectiveness of three methods to improve the performance of acoustic models in low-resource environments. They are monophonic and Tritone learning, as well as functional combinations. The three methods were applied to the network architecture and the results were compared with the baseline. In the proposed hybrid Markov model, the task of Mandarin speech recognition was significantly improved at the phoneme level compared with the neural network method. This also shows that neural networks are not excellent in all speech recognition tasks and there is still much room for development.

Speaking recognition is mostly aimed at adults but automatic speech recognition of children's speech was considered to be a more challenging problem than adult speech. This is due to things such as great acoustic speech variability, including mispronunciation due to ongoing biological changes in growth, vocabulary and language skill progression. Further challenges arise with spontaneous speech from conversational interactions. Kumar et al. (2020) [64] proposed a speech recognition model that uses linguistic information from interactions to adjust children's speech. Two approaches were proposed to exploit this context, namely, lexical repetition and semantic response generation. For the latter, a sequence-to-sequence model is used, which learns to predict target sub-utterances in the context of given adult utterances, incorporating the long-term context into the model by propagating the unit state during the session. Automatic robot activity understanding is imperative in HCI. Existing manipulator control methods, such as position control and visual control methods, cannot achieve autonomous learning. Reinforcement learning can process the interaction between robot control and the environment but it should relearn the control when the position of the target object changes. Thus, Li et al. (2021) [65] proposed a quality model for end-to-end manipulator control using a deep reinforcement learning scheme. Specifically, a strategy search algorithm was designed to realize automatic manipulator learning. To avoid relearning the manipulator, a CNN control scheme was designed to keep the manipulator robust. Extensive experiments have verified the effectiveness of the proposed method.

Combined with the above research results, the voice interaction technology adopted in the HCI intelligent system is comprehensively analyzed and the improvement effect after deep learning and artificial intelligence is combined (Table 1).

Table 1. Development status of deep learning technology in assisting voice interaction in intelligence systems.

Author and Year	Research Scope	The Research Methods	Results	Analysis
Miao et al. (2020)	Speech recognition	Online hybrid CTC/Attention end-to-end ASR architecture.	Compared with the offline CTC/attention model, the online CTC/attention model proposed in this study improves the real-time factors of HCI services and maintains a moderate degradation of its performance.	It takes advantage of the advantages of CTC and attention, which is a significant advance for end-to-end speech automation architecture.

Table 1. Cont.

Author and Year	Research Scope	The Research Methods	Results	Analysis
Liao et al. (2020)	Voice HCI	Collect a large-scale Taiwan Province Putonghua pronunciation and corpus.	The evaluation results showed that the Taiwan-specific Mandarin speech recognition system achieved a Chinese character error rate of 8.1 percent.	We think that the specific Mandarin speech recognition system in Taiwan Province is necessary to improve the performance of man-machine interaction of Putonghua speech in Taiwan Province.
Ho et al. (2020)	Speech emotion recognition	Multilevel multi-head fusion attention mechanism and recurrent neural network.	Experimental results on three databases show that the multimodal speech emotion recognition method has better performance than using a single model.	Recognizing human emotions from speech requires characteristic audio and text features before the data can be fed into appropriate deep-learning algorithms.
Hazer-Rau et al. (2020)	Affective computing in speech interaction	uulmMAC database	The uulmMAC database has made a valuable contribution to the field of effective computing and multimodal data analysis.	Affective computing datasets including classification, feature analysis, multimodal fusion, or intertemporal survey can improve the efficiency of affective computing.
Dokuz and Tufekci (2021)	Speech recognition system	Mini-batch gradient descent	Compared with the standard small-batch sample selection strategy, the proposed strategy performs better.	The deep learning system makes the speech recognition system better adapt to the new language by training algorithms.
Sun et al. (2020)	Speech recognition	Hybrid Hidden Markov Models-phoneme-level neural networks	It applies to all widely used network structures today. The average relative character error rate is reduced by 8.0%.	Acoustic model performance can be improved without the use of data augmentation or transfer learning methods.
Kumar et al. (2020)	Speech recognition for children	Lexical repetition and semantic response generation	The context adaptation model results in a significant improvement over the baseline.	It is applicable to improve the performance of children's speech recognition by using information transmission from adult interlocutors.

From Table 1, researchers have made remarkable achievements in speech recognition tasks and natural language generation tasks for the advantages of strong learning ability and good adaptability of deep learning algorithms. With the use of deep learning techniques, speech recognition systems have achieved greater success, and HCI has become more common. In Human-machine Interfaces (HMIs) with voice support, context plays an

important role in improving user interfaces. Whether it is voice search, mobile communication, or children's speech recognition, HCI combined with deep learning can maintain better robustness.

5. Human Gesture Recognition Based on Deep Learning

A gesture is a form of non-verbal communication that can be used in several fields, such as communication between deaf and mute people, robot control, HCI, home automation, and medical applications. Gesture-based studies employ a number of different techniques, including those based on instrumented sensor technology and computer vision. In other words, gestures are divided into many directions, such as posture and gesture, as well as dynamic and static, or a mixture of both. Oudah et al. (2020) [66] listed the performance of these methods, computer vision technology to deal with similarities and differences, hand segmentation technology used, classification algorithm and shortcomings, number and types of gestures, datasets used, and detection range and camera types used. The use of gesture recognition contains many complex technical difficulties. Gestures are often used by people to convey their thoughts and feelings. For example, hearing-impaired groups always rely on sign language to communicate with each other. However, most normal people do not understand the language and face difficulties in communicating with deaf and mute people. Therefore, the development of automated sign language recognition systems can help facilitate this communication and close the gap.

The structured style of sign language gestures helps facilitate non-verbal communication between the deaf and the hearing impaired. Sign language recognition problems are classified into two categories, namely, static gesture recognition, which focuses on finger spelling, and dynamic recognition, which is related to the recognition of isolated words and continuous sentences. Many continuous sign language recognition systems utilize extended versions of the isolated word framework to recognize entire sentences. With the rapid development of computer science and related fields, the way humans interact with computers has evolved into a more natural and ubiquitous form. Various techniques were developed to capture a user's facial expressions as well as body movements and postures to serve two types of applications. The captured information becomes a "snapshot" of the user so that the computer can better understand the user's intention or emotional state. The user applies natural motion instead of using dedicated input devices to send commands for system control or to interact with digital content in a virtual environment. Gesture interpretation must be carried out quickly and with high accuracy in the vision-based gesture interaction between humans and computers.

Human posture estimation is a challenging task in computer vision. It refers to the process of inferring posture in an image, aiming to determine the position or spatial position of a person's body key points from a given image or video. The estimation principle is shown in Figure 4.

There has been significant progress in addressing the problems and challenges related to human pose estimation aided by deep learning and publicly available datasets. In this survey, Pareek and Thakkar (2020) [67] discussed the characteristics of various machine learning and deep learning techniques of HAR and the public datasets used for HAR and revealed the advantages and disadvantages of action representation, reduction and action analysis methods, as well as the challenges and future directions of HAR. Munea et al. (2020) [68] described the methods used in human pose estimation and then listed some applications and the drawbacks faced in pose estimation. This has established a new research idea for gesture interaction understanding. Therefore, an increasing number of scholars are committed to studying how to optimize human gesture recognition using deep learning, artificial intelligence, and other related technologies in the application process of HCI systems. For example, as the hand is the key to natural HCI, researchers have made many efforts to integrate our hands into the interaction cycle to obtain a more convenient and comfortable interaction experience. For example, Tsai et al. (2020) [69] proposed a low-cost HCI system with a gesture recognition function. The system uses a

variety of visualization techniques, as well as skin and motion detection to capture the region of interest from the background region and proposes a linking element labeling algorithm to identify the centroid of the object. To identify the exact region of the gesture, the arm region is removed with the help of the convex hull algorithm. The simulation results show that despite some interference in the simulation environment, the recognition rate is still very high. The principle of image-based human recognition is shown in Figure 5.

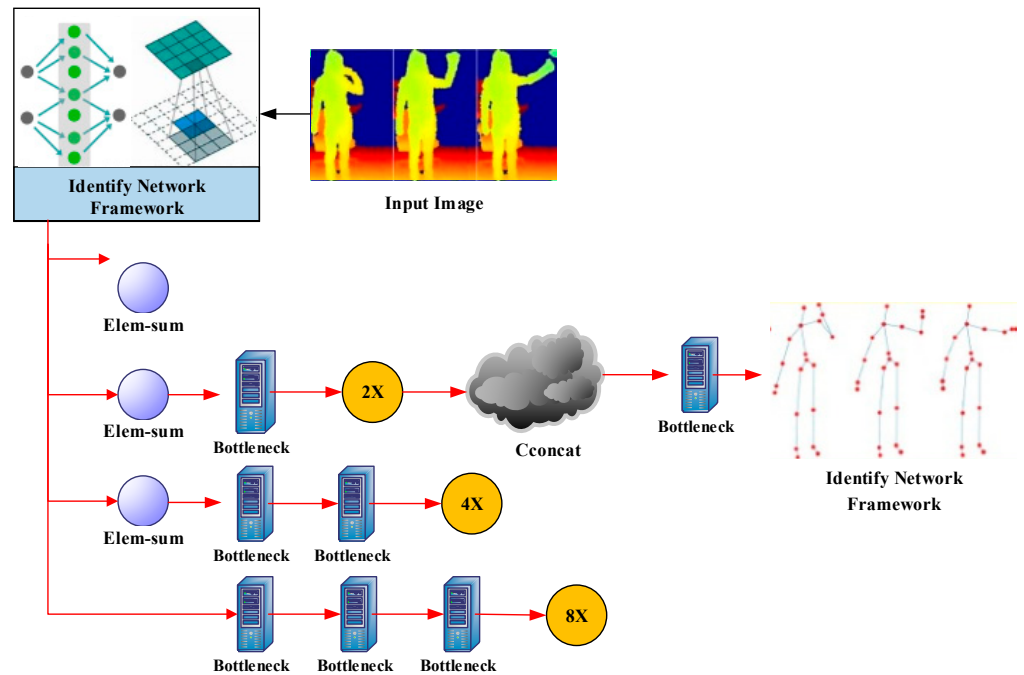


Figure 4. Principle of human posture estimation.

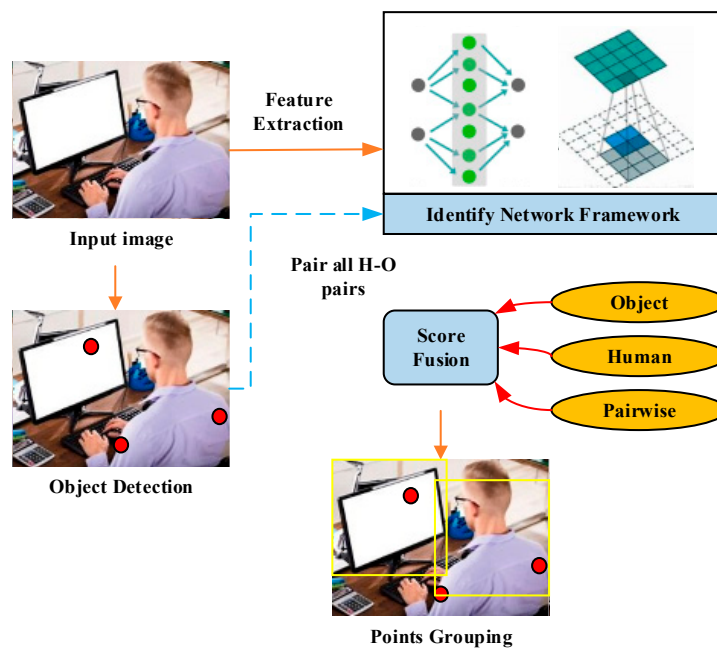


Figure 5. Image-based human interaction detection framework.

In this applied research, the expected goal is basically achieved through continuous efforts, and the original design of the dual-flow model is improved according to the research. To realize human action and gesture recognition, modeling based on the human

skeleton is the main method. Skeleton-based human action recognition has become an active research area and the key to this task is to fully explore spatial and temporal features. However, most graph neural network-based methods use a fixed adjacency matrix defined by datasets, which can only capture the structural information provided by joints directly connected through bones while ignoring the dependencies between unconnected distant joints. In addition, the fixed adjacency matrix makes the network unable to extract multilevel semantic features. Yang et al. (2020) [70] proposed a pseudograph CNN with time and channel attention. Fixed normalized adjacent matrices are replaced by learnable matrices. Hence, the matrix can learn the dependencies between joined joints. Learnable matrices help networks capture multilevel features in spatial domains. In addition, because frames and input channels containing prominent features play an imperative role in distinguishing actions from other actions, a mixed focus on timing and channels is recommended. Similarly, there are many interactive methods of human action and gesture recognition based on bone modeling.

As sensor technology and artificial intelligence make their progress, video gesture recognition technology in the background of big data makes HCI more natural and flexible, which brings a richer interactive experience for teaching, vehicle control, and video games. To perform robust recognition under the conditions of illumination variation, background clutter, fast movement and partial occlusion, Sun et al. (2020) [71] proposed a multilevel feature fusion algorithm based on a dual-stream convolutional neural network, which mainly consists of three steps. First, the Kinect sensor acquires red–green–blue depth images to build a gesture database. Moreover, data augmentation is performed on both the training and test sets. Then, the multistage feature fusion model of the dual-stream convolutional neural network is established and trained. Experimental results show that the proposed network model can stably track and recognize complex backgrounds, such as similar skin colors, illumination variations, and occlusions. Compared with the single-channel model, the average detection accuracy is improved by 1.08% and 3.56%, respectively.

Furthermore, unimodal human behavior recognition on RGB or bone has been extensively studied. Each of these approaches has its own strengths and limitations, as they portray action from different perspectives. Characteristics of different patterns can complement each other to describe actions. Therefore, it makes sense to use the complementarity of the two models to identify actions. However, existing multimode approaches fail to take full advantage of the complementarity of RGB and skeleton modes. Li et al. (2020) [72] proposed a skeleton-guided multimodal network for human behavior recognition. The proposed method makes full use of the complementarity of the two modes at the level of semantic features. From a technical point of view, a bootstrap block is introduced, which is a key component of the skeleton bootstrap multimodal network, and two related operation schemes are discussed. Experimental results show that the proposed method achieves the most advanced performance compared with the existing methods. The same conclusion appeared in the work of Afza et al. (2020) [73]. Here, an action recognition technology based on feature fusion and optimal feature selection was implemented. First, the color transformation of hue-saturation-intensity (HIS) was carried out to improve the contrast of video frames. Then, the motion features were extracted by an optical flow algorithm. Furthermore, a new parallel method named the length control feature was extracted and fused with shape and texture features. The new weighted entropy variance was applied to combination vectors, and the best vector was selected for classification. The multimodal skeleton guidance network is better than single-modal recognition.

Most existing methods using convolutional neural networks and long short-term memory have achieved promising performance for skeleton-based action recognition. However, these methods are limited in their ability to explore rich information about spatiotemporal relationships. Graph convolutional networks achieve the latest results of skeleton-based behavior recognition by extending convolutional neural networks to graphs. However, due to the lack of effective feature aggregation methods, such as maximum pooling in CNNs, existing graph convolutional network-based methods can only learn local

information between adjacent joints. Moreover, it is difficult to obtain high-level interaction features, such as the interaction between the five parts of the human body. Moreover, subtle differences in confounding actions are often hidden in specific channels of key joint features, and this discriminative information has rarely been exploited in previous methods. Chen et al. (2020) [74] proposed a graph convolution network based on a structural graph pool scheme and joint channel attention module. The scheme for the pool of structural maps aggregates human skeletal maps based on prior knowledge of human typology. This pooling scheme not only leads to more global representations but also reduces the number of parameters and the computational cost. The joint channel attention module learns to selectively focus on discriminative joints of the bone and give different degrees of attention to different channels. This novel attention mechanism enhances the ability of the model to classify confounding behaviors. Zhu et al. (2020) [75] proposed a new spatiotemporal model with an end-to-end bidirectional LSTM-CNN, which uses a hierarchical spatiotemporal dependence model to explore the rich spatiotemporal information in skeleton data.

Although convolutional neural networks have achieved great success in object recognition in still images, the improvement of convolutional neural networks over traditional methods for identifying actions in videos is slight because raw videos usually have more redundant or irrelevant information than still images. This point was proven by Yang et al. (2020) [76], where a spatial-temporal attentive convolutional neural network (STA-CNN) was proposed, which selects the discriminative time period and automatically pays attention to the information space region. The STA-CNN model integrates the temporal attention mechanism and spatial attention mechanism into a unified CNN to identify actions in videos. The novel time attention mechanism automatically mines distinguishable time fragments from long and noisy videos. First, the spatial attention mechanism uses instantaneous motion information in optical flow characteristics to locate the significant regions of motion. Then, the training is performed by auxiliary classification loss with a global average pooling layer to focus on discriminant non-moving regions in video frames. The STA-CNN model delivers state-of-the-art performance on two of the most challenging datasets, namely, UCF-101 and HMDB-51.

Human behavior recognition has become the focus of the wider adoption of computer vision. Recognizing the ambiguity of movement comes not only from the difficulty of defining body part movements but also from real-world problems, such as camera movements, dynamic backgrounds, and harsh weather. To assess the performance of these methods quantitatively and qualitatively, common datasets representing various operations under multiple conditions and constraints are recorded. Jegham et al. (2020) [77] summarized it according to the types of problems solved by existing methods. In addition, existing datasets introduced for the field of human behavior recognition were compared. Human motion recognition in video is a difficult task because of its complex background, geometric transformation, and massive data. Computer vision-based video action recognition is widely used in video surveillance, behavior detection, HCI, medical-aided diagnosis, and motion analysis. However, video action recognition may be affected by many factors, such as background and illumination. Dual-stream convolutional neural networks are trained with video spatiotemporal models and fused at the output. The multisegment dual-stream convolutional neural network model trains spatiotemporal information from video, extracts its features and fuses them, and then determines the category of video action. Qiao et al. (2021) [78] used the Google Xception model and transfer learning in their study and took the Xception model trained on ImageNet as the initial weight, which solved model underfitting caused by insufficient video behavior datasets and can effectively reduce the influence of various factors in the video. This approach also greatly improves accuracy and reduces training time. More importantly, the kinetics400 dataset was used for pretraining to compensate for the lack of datasets, which greatly improved the accuracy of the model. The human action recognition process in the video is shown in Figure 6.

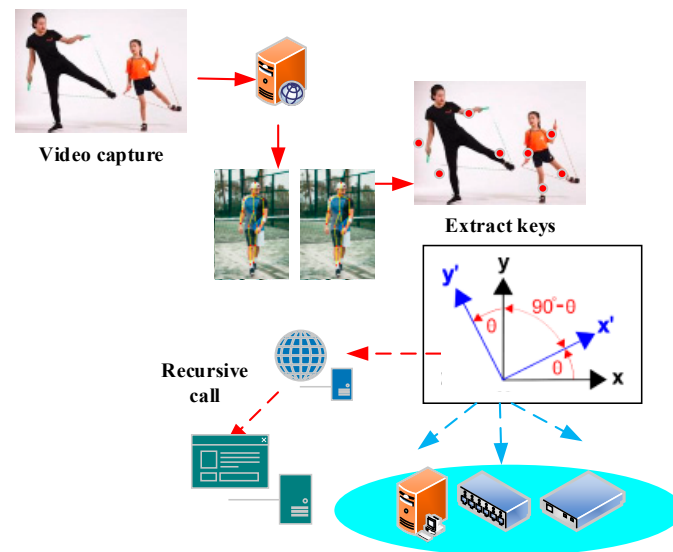


Figure 6. Human motion recognition process in the video.

To achieve human action recognition in videos, Vishwakarma (2020) [79] developed an effective algorithm that can use a single decisive posture to identify human behavior in videos. To accomplish the task, the optical flow was utilized to extract the deterministic posture and the feature was extracted by the double transform of wavelet. The double transform was accomplished by the Gabor wavelet transform and Ridgelet transform. Gabor wavelet transforms generate eigenvectors by computing first-order statistics of various proportions and directions of input poses, which are robust to translation, scaling, and rotation. The Ridgelet transform was utilized to calculate the direction-dependent shape features of human behavior. The fusion features provide a powerful algorithm. The validity of the algorithm was measured on the public Weizmann, Ballet, Movement, and UT Interaction datasets, which reported accuracies of 96.66%, 96%, 92.75%, and 100%, respectively. Comparison of accuracy with related advanced techniques showed the excellent performance of the research. Similarly, Tran et al. (2020) [80] proposed a new method for real-time fingertip detection and gesture recognition using a depth camera and a 3D convolutional neural network. The system can accurately and reliably extract the position of fingertips and recognize gestures in real-time. They demonstrated the accuracy and robustness of the interface by evaluating the gesture recognition of various gestures. In addition, a tool was developed to manipulate computer programs to show the possibility of using gesture recognition. Experimental results showed that the proposed system has high gesture recognition accuracy. Therefore, it is deemed a good method for the future of manual HCI-based gesture interfaces. Chen et al. (2020) [81] conducted an in-depth review of data gloves and vision-based sensor systems and adopted corresponding modeling methods. It was explained that these methods based on computer vision are very promising in hand pose estimation.

HAR is a hot topic in academia and among other stakeholders. Today, it has a wide range of uses and can be used in many practical adoptions, such as health, assisted living, and elderly care. Both visual and sensor-based data are available for HAR. Visual data include video images and skeleton images, while sensor-based data are obtained as digital data from accelerometers, gyroscopes, and other devices. The classification tools and data types used are critical to HAR performance. Ozcan and Basturk (2020) [82] used stacked autoencoders (SAEs) to perform activity identification according to the sensor data. If structural optimization is left to the user experience, using SAE to find results with near-optimal accuracy can be a challenging process. It aimed to increase the accuracy of HAR classification methods via heuristic optimization algorithms. Therefore, the structural parameters of SAE were optimized using a newly developed hybrid algorithm including

particle swarm optimization and artificial swarm optimization algorithm in the internal structure. Each algorithm performed 30 runs and the results were analyzed in detail by statistical methods. The SAE supported by the hybrid algorithm gave the minimum error and was the most robust algorithm.

This point was also verified by Khan et al. (2020) [83], who proposed a novel HAR system under the use of directional gradients and histograms of deep features to incorporate traditional handcrafted features. Initially, human contours were extracted with the help of saliency-based methods. In the first stage, motion and geometric features are extracted from the selected channel, while the Chi-square distance between the extracted minimum distance feature and the threshold-based minimum distance feature is calculated in the second stage. Then, the extracted deep CNN and the handmade features were fused together to generate a result vector. In addition, to address the curse of dimensionality, an entropy-based feature selection technique was proposed to identify the most discriminant features for classification using multiclass support vector machines. All simulations were performed on publicly available benchmark datasets, including Weizmann, YouTube, UCF Sports, and UT-Interaction. There was also a comparative assessment. Compared with a few existing methods, the proposed model showed excellent performance.

Gesture recognition has attracted the attention of many researchers due to its wide application in robotics, gaming, virtual reality, sign language, and HCI. Sign language is a structured form of hand gesture and the most effective way to communicate with hearing impairment. Coupled with the current variety of computer interfaces, users can interact with virtual worlds in various ways. There are three major challenges, namely hand segmentation, hand shape feature representation, and gesture sequence recognition to develop an efficient sign language recognition system to recognize dynamically isolated gestures in virtual environments. Traditional sign language recognition methods use color-based hand segmentation algorithms to segment hands, handcrafted features are extracted for hand shape representation, and hidden Markov models are used for sequence recognition. Seinfeld et al. (2021) [84] identified a set of concepts related to different user representations and conducted a multidisciplinary review of the multisensory and cognitive factors behind the control and subjective experience of user representations. Aly and Aly (2020) [85] proposed a new framework for signature-independent sign language recognition using a variety of deep learning architectures under hand semantic segmentation, hand shape feature representation, and deep recurrent neural networks. The recently developed combined method called DeepLabv3 and semantic segmentation is trained using a set of pixel-labeled hand images to extract the hand region from each frame of the input video, using a deep Bidirectional Long Short-Term Memory (BiLSTM) recursive neural network to recognize the extracted sequence of feature vectors. The BiLSTM network contains three BiLSTM layers and a fully connected softmax layer. The performance of the proposed method was evaluated using a challenging Arabic sign language database containing 23 isolated words captured from three different users. Experimental results show that the proposed framework outperforms state-of-the-art signer-independent test strategy methods.

In addition to the field of virtual reality, one of the main mobile edge computing technologies in healthcare monitoring systems is human motion recognition. Built-in multi-functional sensors make smartphones a ubiquitous platform for acquiring and analyzing data, allowing smartphones to perform human motion recognition. The task of identifying human activities using accelerometers built into smartphones has been well addressed, but in practice, these traditional methods fail to identify complex and real-time human activities with multimodal and high-dimensional sensor data. Wan et al. (2020) [86] designed an architecture based on a smartphone inertial accelerometer for human action recognition in their study. As participants perform typical daily activities, the smartphone collects sensory data sequences, extracts efficient features from the raw data, and then captures the user's physical behavior data through multiple three-axis accelerometers. It was concluded that hand posture estimation is a great academic and technical challenge because of the struc-

ture and dexterous movement of the human hand. Driven by advances in hardware and artificial intelligence, various data glove prototypes and computer vision-based methods have been proposed in recent years for accurate and fast hand pose estimation. However, existing reviews either focus on data wearables or visual methods or are based on specific types of cameras, such as depth cameras.

Combined with the above research results, the gesture interaction technology adopted in the HCI intelligent system is comprehensively analyzed, and the improvement effect after deep learning and artificial intelligence is combined (Table 2).

Table 2. Development status of deep learning technology in assisting gesture interaction in intelligent systems.

Author and Year	Research Scope	The Interaction Technology Adopted	Results	Summary and Analysis
Pareek and Thakkar (2020)	Human behavior recognition	Public datasets and Deep learning	Action recognition technology and applications of human behavior recognition are reviewed.	Content-based video HCI, education, healthcare, and abnormal activity detection can achieve better results on the basis of effective datasets.
Tsai et al. (2020)	Gesture recognition	Multiple visual techniques and connected component labeling algorithm	A low-cost HCI system with gesture recognition function is established, and the recognition rate is very high.	To perform gesture interpretation quickly and with high accuracy, a reliable labeling algorithm is needed.
Yang et al. (2020)	Human action recognition	Graph Convolutional network and Human skeleton modeling	Performance comparable to state-of-the-art methods is achieved on NTU-RGB+D and HDM05 datasets.	Graph convolutional neural networks can help solve the dependence relationship between unconnected distant joints and improve recognition accuracy.
Sun et al. (2020)	Gesture recognition algorithm	Dual stream convolutional neural network and Kinect sensor	The multilevel feature fusion model of dual stream convolutional neural network is established and trained. For gesture tracking and recognition in complex backgrounds, the average detection accuracy increased by 1.08% and the average accuracy increased by 3.56%.	Sensor technology, artificial intelligence, and big data technology make the HCI of video gesture recognition more natural and flexible.

Table 2. Cont.

Author and Year	Research Scope	The Interaction Technology Adopted	Results	Summary and Analysis
Li et al. (2020)	Gesture recognition	Skeleton-guided multimodal network	In this way, skeleton features can guide RGB features in action recognition, to enhance the important RGB information closely related to actions.	The single-mode human behavior recognition mode of RGB or bone is integrated and complementary to describe the action, and the recognition performance can be optimized.
Afza et al. (2020)	Gesture recognition	Sparse activation function and feature fusion and weighted entropy-variance	The recognition rate is 97.9%, 100%, 99.3%, and 94.5% in four famous action datasets, respectively.	The action recognition technology based on feature fusion and best feature selection has high recognition rate.
Chen et al. (2020)	Gesture recognition	Graph convolutional network	The new graph convolution network based on structure graph pooling scheme and joint channel attention module reduces the number of parameters and computational cost.	An effective feature aggregation method is one of the keys to skeleton-based action recognition. Attention mechanisms can enhance the model's ability to classify confusing behaviors.
Zhu et al. (2020)	Gesture recognition	Two-way LSTM-CNN	The new spatiotemporal model of end-to-end bidirectional low frequency modulation (BiLSTM-CNN) is effective on NTU RGB+D, SBU interaction, and UTD-MHAD datasets.	Efficient and low-cost human bone capture systems rely on the complementary performance of neural networks.
Yang et al. (2020)	Gesture recognition	Spatiotemporal attention convolutional neural networks	The spatiotemporal attention mechanism automatically mines discriminative temporal fragments from long, noisy videos. State-of-the-art performance was achieved on datasets UCF-101 (95.8%) and HMDB-51 (71.5%).	Convolutional neural networks alone can achieve high accuracy in object recognition of delicate images, but the improvement effect of motion recognition in video is not obvious.

Table 2 shows that there are many research achievements on human action recognition in still images and videos. In particular, action recognition combined with convolutional neural networks and long short-term memory networks can greatly improve the accuracy and precision of action recognition. However, it also faces some technical difficulties, therefore, it is necessary to synthesize or improve the traditional neural network to achieve a better recognition effect. Therefore, in the future, more researchers will conduct multidimensional analysis and propose more solutions.

6. Natural Language Processing and HCI

Human–computer dialog is one of the most natural ways of HCI. Its development has influenced and promoted the progress of speech recognition and synthesis, natural language understanding [87–89], dialog management, and natural language generation. Due to the limitations of interaction efficiency and ergonomics, it is difficult for gesture interaction and other methods to become the mainstream HCI mode in the short term. However, products with voice interaction capability have been widely studied since their application [90–92]. In terms of interactive means, users can give orders, play music, control their home, and perform other tasks only through dialog with related products, which can truly free their hands and improve the happiness index of life. Natural language understanding (NLU) can enable the computer to understand the user’s language to make further decisions or complete interactive actions, which is an important task for products with voice interaction capabilities to handle [93–95], such as machine translation, man–machine dialog robots, and smart homes [96]. It is popular to adopt deep learning to solve some problems in natural language processing, and the performance will be better compared with traditional machine learning methods [97]. However, deep learning still belongs to the category of machine learning. Many concepts of machine learning are common in deep learning, such as datasets, loss functions, overfitting, and other basic concepts. Many scholars have made efforts in this field.

For example, in the research of Yunanto et al. (2019) [98], to build educational games to realize artificial intelligence, the implementation method of establishing a Non-Player Character (NPC) based on natural language processing was implemented so that NPC can automatically answer questions about English. The average score of educational games with this NPC was higher than 75% of users. The presence of NPCs in educational games can increase user interest. With this natural language processing technique, the popularity ranking of the educational game genre can be increased.

Dialog robots are used in many systems. If a robot accepts a task instruction in natural language, it must first decode the instruction to understand the user’s intention. Therefore, Pramanick et al. (2022) [99] introduced a system named Talk-to-Resolve (TTR), which enables the robot to solve the deadlock by visually observing the scene to initiate coherent dialog and communication with the coach. Using the observed scene and the given instructions to calculate the robot’s next action together can greatly improve the accuracy of the conversation. While robots should observe their surroundings in different ways, natural gestures and spoken words are the most convenient ways for humans to interact with robots. Only when the robot can understand this type of interaction will it be possible to achieve a true human–machine dialog. This point is reflected in the study of El-Komy et al. (2022) [100]. In this study, smartphones with visual, language, and intelligence functions were used to help visually impaired people avoid obstacles, and voice output was used to remind them, which is undoubtedly good news for blind people. Recupero and Spiga (2020) [101] proposed a method to allow NAO humanoid robots to perform user-spoken natural language commands, define an action robot ontology, perform machine reads on the input text given by the user (natural language), and attempt to identify the action commands to be performed by the robot. This research is a large step forward for human–machine conversations and for machines’ understanding of natural language.

To achieve a real natural dialog with emotion in HCI, emotion processing and recognition are needed [102–104]. Li et al. (2019) [105] proposed a method of combining prosody valence and text emotion through decision-level fusion, which reduced fatal recognition errors and thus improved user experience. According to the distribution inferred from person-to-person conversation data, the parameters estimated by the recognition function were used for prediction. The evaluation of ten participants showed that the system enhanced by the module can effectively carry out natural conversations. Similarly, Jia (2021) [106] proposed a novel sentiment classification framework. The framework can identify semantic emotional terms and emoticons. This strategy resulted in multi-emotion and polarity classifications that were 3% to 4% more accurate than the next best-performing

baseline classifier. This is very helpful for enhancing the interaction between humans and chatbots and for emotion classification. However, even without speech recognition errors, robots may face difficulties interpreting natural language commands [107,108]. Therefore, Marge and Rudnicky (2019) [109] proposed a robust method to deal with poor communication between humans and robots in the task-oriented oral dialog. When a robot encounters a communication problem, it can look back at its interaction history to consider how it solved similar situations. The research helps robots solve problems in time when they encounter difficult instructions that are difficult to interpret.

One of the additional chatbot developments is to help people use named entity recognition in the text to book flights, track sentences to detect user intent and respond when the context of the conversation domain is limited. Permatasari and Maharani (2021) [110] used NLU in their study to analyze and design chatbot interactions, aiming to make the robot understand the user's meaning and provide the best and correct response. It turns out that dialog managers using reinforcement learning can bring low costs to computation in chatbots. Therefore, the combination of natural language understanding and reinforcement learning is very helpful for robot humanization.

Ghiță et al. (2020) [111] introduced a social robot framework that is designed in a modular and powerful way for the assisted care scenario. The framework includes robot services for navigation, human detection and recognition, multilingual natural language interaction and dialog management, as well as activity recognition and general behavior composition. In addition, the dialog was widely used for verbal interaction between humans and robots, such as auxiliary robots in hospitals. However, the robot is usually limited by the scheduled conversation, so it is difficult to understand the new words of the new target. Rofi'ah et al. (2021) [112] discussed conversations in Bahasa Indonesia about entertainment, motivation, emergency situations, and ways to help with knowledge growth. In emergency situations, patients were able to request a robot to call a nurse. Reinforcement learning methods to overcome the limitations of robot knowledge were adopted to achieve the new dialog goals of the patient assistant.

The above application of natural language processing in HCI is summarized in Table 3.

Table 3. Development status of human–computer interaction in natural language processing.

Author and Year	Research Scope	The Interaction Technology Adopted	Results	Summary and Analysis
Yunanto et al. (2019)	Educational games and NPCS	Natural language processing	The average score of an educational game with this NPC is higher than 75% of users.	The presence of intelligent NPCS in educational games can increase user interest.
Recupero and Spiga (2020)	Human–computer dialog and Natural language processing	Natural language and speech interaction	For each action the robot can perform, a corresponding element is simulated in the ontology to understand human natural language.	Robots are being given the ability to read natural language more intelligently, a huge step forward in understanding human movement and speech.
Li et al. (2019)	Man–machine dialog and sentiment analysis	Multimodal sentiment analysis and Natural language processing	The user experience is improved by reducing low-level identification errors.	To achieve a truly natural human–computer conversation, it needs to combine sentiment analysis.

Table 3. Cont.

Author and Year	Research Scope	The Interaction Technology Adopted	Results	Summary and Analysis
Marge and Rudnicky (2019)	Speech recognition and HCI	TeamTalk and Nearest neighbor algorithm	A recovery strategy is selected for virtual robots that encounter unexplained instructions	Information from the robot's path planner and its surroundings can help the robot detect and recover from miscommunication in a conversation.
Jia (2021)	The man-machine dialog	Emotion classification and Language processing	This strategy makes the multi-emotion and polarity classification 3% to 4% more accurate than the next best-performing baseline classifier.	It makes sense that emoticons should be considered in sentiment classification schemes.
Yunanto et al. (2019)	Educational games and NPCs	Natural language processing	The average score of an educational game with this NPC is higher than 75% of users.	The presence of intelligent NPCs in educational games can increase user interest.
Recupero and Spiga (2020)	Human-computer dialog and Natural language processing	Natural language processing and speech interaction	For each action the robot can perform, a corresponding element is simulated in the ontology to understand human natural language.	Robots are being given the ability to read natural language more intelligently, a huge step forward in understanding human movement and speech.
Ghiță et al. (2020)	Social robot	Natural language processing and robot operating system and voice interaction	It focuses on the quantitative evaluation of each functional module, discussing their performance and possible improvements in different settings.	Social robots can provide economic efficiency and growth in areas such as retail, entertainment, and active and assisted living.
Rofi'ah et al. (2021)	Dialog robot	Reinforcement learning and voice interaction	Reinforcement learning approaches that overcome the knowledge limitations of robots achieve new dialog goals for patient assistants.	The hospital's assistive robot uses reinforcement learning to help it grow its database of knowledge conversations, making the robot more understanding.

7. Summary of the Application Status of Deep Learning in HCI

Based on the above studies on the application of deep learning in different fields of HCI systems, the algorithms (methods) and datasets used in these studies are summarized in Table 4.

Hence, for the HCI system, deep learning, regardless of speech recognition, emotion recognition, or human-computer dialog, makes the established system more intelligent and can greatly enhance the ability of the machine model to identify and classify and analyze confusing behaviors. HCI creates a new generation of social information technology by understanding the relationship among information technology, human life, and social development to achieve the goal of constantly challenging the limits of human potential. Deep learning provides technical support for HCI, which makes it no longer a simple dialog

between humans and machines but also contains emotional information, thus making HCI develop on a deeper level.

Table 4. Summary of algorithms and datasets used by deep learning technology in intelligent HCI systems.

Author and Year	Research Scope	Algorithm (Model)	Dataset
Hazer-Rau et al. (2020)	Emotional computing	Affective computing and multimodal data analysis methods	uulmMAC
Iio et al. (2020)	Social robot	The question–answer–response dialog model	A collection of conversations from the nursing home site
Li (2021)	Gesture recognition	HCI model of manipulator operated by manipulator	Subject site collection
Calvo et al. (2021)	Speech recognition and interaction	Mobile and personal voice assistant platforms	Questionnaire survey results and on-site evaluation
Tao and Busso (2020)	Speech recognition and interaction	Multitask learning and automatic audiovisual speech recognition systems	An audio–visual corpus
Duan et al. (2021)	Gesture recognition	Weight adaptive algorithm combining different features	Gesture image dataset
Wang et al. (2020)	Context awareness	Context-aware citation recommendation model based on end-to-end memory network	Three real datasets
Miao et al. (2020)	Speech recognition	Online hybrid based on connectionist temporal classification/attention end-to-end automatic speech recognition architecture	LibriSpeech
Ho et al. (2020)	Speech emotion recognition	Multimodal speech emotion recognition method based on multilevel multi-head fusion attention mechanism and recurrent neural network	CMU-MOSEI, IEMOCA and MELD
Chen et al. (2020)	Motion recognition	A novel graph convolution network based on structure graph pooling scheme and joint channel attention module	NTU-RGB+D, Kinetics-M, and SYSU-3D
Zhu et al. (2020)	Motion recognition	A new spatiotemporal model of end-to-end bidirectional Low-frequency modulation (BiLSTM-CNN)	NTU RGB+D, SBU Interaction and UTD-MHAD
Yang et al. (2020)	Motion recognition	Spatiotemporal concern convolutional neural network model	UCF-101 (95.8%) and HMDB-51 (71.5%)
Aly and Aly (2020)	Gesture recognition	Hidden Markov model and Color-based hand segmentation algorithm	Hand graphics collection based on DeepLabv3
Li et al. (2019)	Man–machine dialog	An algorithm that combines prosody valence with text emotion through decision-level fusion	A survey of the subject’s experience

Table 4. Cont.

Author and Year	Research Scope	Algorithm (Model)	Dataset
Marge and Rudnicky (2019)	Man–machine dialog	Nearest neighbor learning algorithm	Crowd-sourced data and user experience data
Jia (2021)	Dialog robot	Word2vec and vector arithmetic and improved k-means similarity calculation	Emotional dictionary
Permatasari and Maharani (2021)	Robot	Support vector machine and feature extraction combination algorithm	Chatbot dialog collection

8. Challenges of Deep Learning in Intelligent HCI

HCI is the interactive relationship between the system and users, which uses dialog language between people and computers that completes the process of information exchange between man and computer in a certain interactive way. The human–computer interface refers to the part visible to the user, on which the user communicates with the system and performs operations. In HCI, the natural interaction behavior of human beings and the state change of physical space are multichannel, imprecise, and very unstable modes. As a cognitive subject, it is a great challenge for computers to understand human natural interactions, intentions, and questions and to provide accurate feedback. At present, there is still much room to improve the accuracy and real-time performance of natural perception technology. Human physiology and psychological changes can affect the state of interaction at any time. In the era of the Internet of Everything, to compensate for the limitations of science and technology, experience design has never been as important as it is now. The core of HCI design is gradually developing toward the direction of intelligence, humanization, and scenarization. With so many smart devices, so many screens, and so many notifications, there is so much information overload that users cannot even digest it. The more information users acquire, the more anxious they become. Target-driven business competition results in every device and software competing for users' limited time.

Speculation about the next revolutionary method of HCI is the focus of the industry. It was described as the closest thing to natural human interaction, which includes voice interaction, gesture recognition in multiple scenes, brain–computer interaction by connecting the human brain and computer, holographic operation in different scenes, and full interface without a touch screen. These interaction modes have relatively high requirements on technology and product hardware, and it may be difficult to give full play to their advantages under the current technical constraints. However, as future technologies and products develop to a certain stage, new mainstream HCI will emerge in them. Voice interaction is viewed as one of the primary traffic entry points for users in many future scenarios. Therefore, seeking reliable and effective far-field speech technology breakthroughs has become an urgent demand in the current industry and academia. Multichannel microphone array technology was proven to significantly improve the quality of speech recognition. When the number of signal acquisition channels is large enough, additional multichannel synchronization technology needs to be developed. In addition, at present, there are few integrated multiple microphones in consumer electronics and few relevant research results, which also increases the difficulty in the progression of this hardware solution. In speech recognition systems, the design of microphone array signal processing algorithms should be emphasized.

Far-field speech recognition is mainly faced with echo interference, indoor reverberation, multisource interference, and non-stationary noise interference. To solve the echo interference problem, echo cancellation technology should be adopted to remove the sound played by the device from the signal received by the microphone. The technology is already well established on handheld mobile devices, including open-source software such as speex and webrtc. However, to achieve a greater echo suppression effect, the two schemes use

abundant nonlinear processing methods. The speech recognition engine is very sensitive to the nonlinear processing of speech signals. Therefore, if direct near-field echo cancellation is used in the field of far-field speech recognition, the effect is not good. Deep learning outperforms traditional models in machine learning tasks. A deep neural network is an effective solution because of the ability to automatically learn the time correlation of time series. However, choosing the most convenient deep neural network and its parameterization is a complex task that requires considerable professional knowledge. Therefore, the applicability of the existing architecture to different prediction tasks needs to be further explored. The success of the deep Web is remarkable. They have made visual research very popular, dramatically increased the interaction between academia and industry, allowed visual technology to be applied to many disciplines, and produced many important results. However, the limitation of the depth of the network is also more prominent; it cannot be interpreted and is vulnerable to fraud, which is too dependent on annotation data features. Hence, many researchers have issued the call of “Deep Learning is dead”, calling for attention to Deep Learning outside of the other methods, therefore, Deep Learning faces a great challenge in the future.

9. Conclusions

With the progression of science and technology, the HCI method has developed from traditional print media to intelligent media. Gesture control, voice control, dialog robots, and other VR, AR, and AI interactive devices have emerged in an endless stream, bringing earth-shaking changes to people’s lives. Gesture control has several advantages over traditional touch screens. Gesture control is an alternative to voice control, especially in public areas. Virtual reality glasses allow users to immerse themselves in an artificial three-dimensional world. Virtual, augmented, and mixed reality is utilized for entertainment, gaming and industry 4.0, which also allow for remote control. As a result, mankind was able to expand its field of experience and action. Machines will continue to become better at interpreting signals, such as when an autonomous car responds correctly to hand signals from a traffic policeman. Caregiving robots assess the needs of those who cannot express their feelings. The more complex contributions that the machines make, the more imperative it becomes for them to communicate effectively with their users. Speech recognition and human movement recognition based on deep learning have improved the accuracy and realism of HCI. HCI is just the beginning. In the future, more data from different sensors will be combined to capture and control complex processes.

Funding: The statement information related to funding has been added to the text.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Jarosz, M.; Nawrocki, P.; Śnieżyński, B.; Indurkha, B. Multi-Platform Intelligent System for Multimodal Human-Computer Interaction. *Comput. Inform.* **2021**, *40*, 83–103. [\[CrossRef\]](#)
2. Prathiba, T.; Kumari, R.S.S. Content based video retrieval system based on multimodal feature grouping by KFCM clustering algorithm to promote human–computer interaction. *J. Ambient. Intell. Humaniz. Comput.* **2021**, *12*, 6215–6229. [\[CrossRef\]](#)
3. Wang, Z.; Jiao, R.; Jiang, H. Emotion Recognition Using WT-SVM in Human-Computer Interaction. *J. New Media* **2020**, *2*, 121–130. [\[CrossRef\]](#)
4. Fu, Q.; Lv, J. Research on Application of Cognitive-Driven Human-Computer Interaction. *Am. Sci. Res. J. Eng. Technol. Sci.* **2020**, *64*, 9–27.
5. Ince, G.; Yorganci, R.; Ozkul, A.; Duman, T.B.; Köse, H. An audiovisual interface-based drumming system for multimodal human–robot interaction. *J. Multimodal User Interfaces* **2020**, *15*, 413–428. [\[CrossRef\]](#)
6. Raptis, G.; Kavvetsos, G.; Katsini, C. MuMIA: Multimodal Interactions to Better Understand Art Contexts. *Appl. Sci.* **2021**, *11*, 2695. [\[CrossRef\]](#)
7. Wang, J.; Cheng, R.; Liu, M.; Liao, P.-C. Research Trends of Human–Computer Interaction Studies in Construction Hazard Recognition: A Bibliometric Review. *Sensors* **2021**, *21*, 6172. [\[CrossRef\]](#)
8. Wu, D.; Zhang, J.; Zhao, Q. Multimodal Fused Emotion Recognition About Expression-EEG Interaction and Collaboration Using Deep Learning. *IEEE Access* **2020**, *8*, 133180–133189. [\[CrossRef\]](#)

9. Lai, H.; Chen, H.; Wu, S. Different Contextual Window Sizes Based RNNs for Multimodal Emotion Detection in Interactive Conversations. *IEEE Access* **2020**, *8*, 119516–119526. [[CrossRef](#)]
10. Yadav, S.K.; Tiwari, K.; Pandey, H.M.; Akbar, S.A. A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions. *Knowledge-Based Syst.* **2021**, *223*, 106970. [[CrossRef](#)]
11. Mosquera-DeLaCruz, J.H.; Loaiza-Correa, H.; Nope-Rodríguez, S.E.; Restrepo-Girón, A.D. Human-computer multimodal interface to internet navigation. *Disabil. Rehabil. Assist. Technol.* **2021**, *16*, 807–820. [[CrossRef](#)] [[PubMed](#)]
12. Nayak, S.; Nagesh, B.; Routray, A.; Sarma, M. A Human–Computer Interaction framework for emotion recognition through time-series thermal video sequences. *Comput. Electr. Eng.* **2021**, *93*, 107280. [[CrossRef](#)]
13. Yang, T.; Hou, Z.; Liang, J.; Gu, Y.; Chao, X. Depth Sequential Information Entropy Maps and Multi-Label Subspace Learning for Human Action Recognition. *IEEE Access* **2020**, *8*, 135118–135130. [[CrossRef](#)]
14. Panjaitan, M.I.; Rajagukguk, D.M. Development of computer-based photoshop learning media using computer based interaction method. *J. Sci.* **2020**, *8*, 37–41.
15. Liu, X.; Zhang, L. Design and Implementation of Human-Computer Interaction Adjustment in Nuclear Power Monitoring System. Mi-croprocessors and Microsystems. *Microprocess. Microsyst.* **2021**, 104096. [[CrossRef](#)]
16. Yuan, J.; Feng, Z.; Dong, D.; Meng, X.; Meng, J.; Kong, D. Research on Multimodal Perceptual Navigational Virtual and Real Fusion Intelligent Experiment Equipment and Algorithm. *IEEE Access* **2020**, *8*, 43375–43390. [[CrossRef](#)]
17. Dybvik, H.; Erichsen, C.K.; Steinert, M. Demonstrating the feasibility of multimodal neuroimaging data capture with a wearable electroencephalography + functional near-infrared spectroscopy (eeg+fnirs) in situ. *Proc. Des. Soc.* **2021**, *1*, 901–910. [[CrossRef](#)]
18. Hu, Y.; Li, Z. Research on Human-Computer Interaction Control Method in the Background of Internet of Things. *J. Interconnect. Networks* **2022**, *22*, 2143015. [[CrossRef](#)]
19. Fox, J.; Gambino, A. Relationship Development with Humanoid Social Robots: Applying Interpersonal Theories to Human–Robot Interaction. *Cyberpsychol. Behav. Soc. Netw.* **2021**, *24*, 294–299. [[CrossRef](#)]
20. Henschel, A.; Hortensius, R.; Cross, E.S. Social cognition in the age of human–robot interaction. *Trends Neurosci.* **2020**, *43*, 373–384. [[CrossRef](#)]
21. Sebo, S.; Stoll, B.; Scassellati, B.; Jung, M.F. Robots in groups and teams: A literature review. *Proc. ACM Hum.-Comput. Interact.* **2020**, *4*, 176. [[CrossRef](#)]
22. Lei, X.; Rau, P.-L.P. Should I Blame the Human or the Robot? Attribution within a Human–Robot Group. *Int. J. Soc. Robot.* **2021**, *13*, 363–377. [[CrossRef](#)]
23. Iio, T.; Yoshikawa, Y.; Chiba, M.; Asami, T.; Isoda, Y.; Ishiguro, H. Twin-Robot Dialogue System with Robustness against Speech Recognition Failure in Human-Robot Dialogue with Elderly People. *Appl. Sci.* **2020**, *10*, 1522. [[CrossRef](#)]
24. Pan, S. Design of intelligent robot control system based on human–computer interaction. *Int. J. Syst. Assur. Eng. Manag.* **2021**, 1–10. [[CrossRef](#)]
25. Ma, G.; Hao, Z.; Wu, X.; Wang, X. An optimal Electrical Impedance Tomography drive pattern for human-computer interaction applications. *IEEE Trans. Biomed. Circuits Syst.* **2020**, *14*, 402–411. [[CrossRef](#)]
26. Li, X. Human–robot interaction based on gesture and movement recognition. *Signal Process. Image Commun.* **2020**, *81*, 115686. [[CrossRef](#)]
27. Robert, L.P., Jr.; Bansal, G.; Lütge, C. ICIS 2019 SIGHCI workshop panel report: Human–computer interaction challenges and opportunities for fair, trustworthy and ethical artificial intelligence. *AIS Trans. Hum.-Comput. Interact.* **2020**, *12*, 96–108. [[CrossRef](#)]
28. Shu, Y.; Xiong, C.; Fan, S. Interactive design of intelligent machine vision based on human–computer interaction mode. *Microprocess. Microsyst.* **2020**, *75*, 103059. [[CrossRef](#)]
29. Luria, M.; Sheriff, O.; Boo, M.; Forlizzi, J.; Zoran, A. Destruction, Catharsis, and Emotional Release in Human-Robot Interaction. *ACM Trans. Hum.-Robot Interact.* **2020**, *9*, 22. [[CrossRef](#)]
30. Demir, M.; McNeese, N.J.; Cooke, N.J. Understanding human-robot teams in light of all-human teams: Aspects of team interaction and shared cognition. *Int. J. Hum.-Comput. Stud.* **2020**, *140*, 102436. [[CrossRef](#)]
31. Johal, W. Research Trends in Social Robots for Learning. *Curr. Robot. Rep.* **2020**, *1*, 75–83. [[CrossRef](#)]
32. Jyoti, V.; Lahiri, U. Human-Computer Interaction based Joint Attention cues: Implications on functional and physiological measures for children with autism spectrum disorder. *Comput. Hum. Behav.* **2020**, *104*, 106163. [[CrossRef](#)]
33. Suwa, S.; Tsujimura, M.; Ide, H.; Kodate, N.; Ishimaru, M.; Shimamura, A.; Yu, W. Home-care Professionals’ Ethical Perceptions of the Development and Use of Home-care Robots for Older Adults in Japan. *Int. J. Hum.-Comput. Interact.* **2020**, *36*, 1295–1303. [[CrossRef](#)]
34. Gervasi, R.; Mastrogiacomo, L.; Franceschini, F. A conceptual framework to evaluate human-robot collaboration. *Int. J. Adv. Manuf. Technol.* **2020**, *108*, 841–865. [[CrossRef](#)]
35. Pretto, N.; Poiesi, F. Towards gesture-based multi-user interactions in collaborative virtual environments. In Proceedings of the 5th International Workshop LowCost 3D-Sensors, Algorithms, Applications, Hamburg, Germany, 28–29 November 2017; pp. 203–208.
36. Pani, M.; Poiesi, F. Distributed data exchange with Leap Motion. In *International Conference on Augmented Reality, Virtual Reality and Computer Graphics*; Springer: Cham, Switzerland, 2018; pp. 655–667.
37. Cao, Y.; Geddes, T.A.; Yang, J.Y.H.; Yang, P. Ensemble deep learning in bioinformatics. *Nat. Mach. Intell.* **2020**, *2*, 500–508. [[CrossRef](#)]

38. Wang, G.; Ye, J.C.; De Man, B. Deep learning for tomographic image reconstruction. *Nat. Mach. Intell.* **2020**, *2*, 737–748. [[CrossRef](#)]
39. Yu, K.; Tan, L.; Lin, L.; Cheng, X.; Yi, Z.; Sato, T. Deep-Learning-Empowered Breast Cancer Auxiliary Diagnosis for 5GB Remote E-Health. *IEEE Wirel. Commun.* **2021**, *28*, 54–61. [[CrossRef](#)]
40. Panwar, H.; Gupta, P.; Siddiqui, M.K.; Morales-Menendez, R.; Singh, V. Application of deep learning for fast detection of COVID-19 in X-rays using nCOVnet. *Chaos Solitons Fractals* **2020**, *138*, 109944. [[CrossRef](#)]
41. Ma, W.; Liu, Z.; Kudyshev, Z.A.; Boltasseva, A.; Cai, W.; Liu, Y. Deep learning for the design of photonic structures. *Nat. Photonics* **2021**, *15*, 77–90. [[CrossRef](#)]
42. Wang, S.; Zha, Y.; Li, W.; Wu, Q.; Li, X.; Niu, M.; Wang, M.; Qiu, X.; Li, H.; Yu, H.; et al. A fully automatic deep learning system for COVID-19 diagnostic and prognostic analysis. *Eur. Respir. J.* **2020**, *56*, 2000775. [[CrossRef](#)]
43. Minaee, S.; Kalchbrenner, N.; Cambria, E.; Nikzad, N.; Chenaghlu, M.; Gao, J. Deep Learning-based Text Classification: A Comprehensive Review. *ACM Comput. Surv.* **2021**, *54*, 1–40. [[CrossRef](#)]
44. Yuan, Q.; Shen, H.; Li, T.; Li, Z.; Li, S.; Jiang, Y.; Xu, H.; Tan, W.; Yang, Q.; Wang, J.; et al. Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sens. Environ.* **2020**, *241*, 111716. [[CrossRef](#)]
45. Calvo, I.; Tropea, P.; Viganò, M.; Scialla, M.; Cavalcante, A.B.; Grajzer, M.; Gilardone, M.; Corbo, M. Evaluation of an Automatic Speech Recognition Platform for Dysarthric Speech. *Folia Phoniatr. Logop.* **2021**, *73*, 432–441. [[CrossRef](#)]
46. Tao, F.; Busso, C. End-to-End Audiovisual Speech Recognition System with Multitask Learning. *IEEE Trans. Multimedia* **2020**, *23*, 1–11. [[CrossRef](#)]
47. Bhatt, S.; Jain, A.; Dev, A. Continuous Speech Recognition Technologies—A Review. In *Recent Developments in Acoustics*; Springer: Singapore, 2021; pp. 85–94. [[CrossRef](#)]
48. Shen, C.-W.; Luong, T.-H.; Ho, J.-T.; Djailani, I. Social media marketing of IT service companies: Analysis using a concept-linking mining approach. *Ind. Mark. Manag.* **2019**, *90*, 593–604. [[CrossRef](#)]
49. Shen, C.-W.; Chen, M.; Wang, C.-C. Analyzing the trend of O2O commerce by bilingual text mining on social media. *Comput. Hum. Behav.* **2019**, *101*, 474–483. [[CrossRef](#)]
50. Pustejovsky, J.; Krishnaswamy, N. Embodied Human Computer Interaction. *KI-Künstl. Intell.* **2021**, *35*, 307–327. [[CrossRef](#)]
51. Duan, H.; Sun, Y.; Cheng, W.; Jiang, D.; Yun, J.; Liu, Y.; Liu, Y.; Zhou, D. Gesture recognition based on multi-modal feature weight. *Concurr. Comput. Pract. Exp.* **2021**, *33*, e5991. [[CrossRef](#)]
52. Wang, P.; Liu, H.; Wang, L.; Gao, R.X. Deep learning-based human motion recognition for predictive context-aware human-robot collaboration. *CIRP Ann.* **2018**, *67*, 17–20. [[CrossRef](#)]
53. Wang, J.; Zhu, L.; Dai, T.; Wang, Y. Deep memory network with Bi-LSTM for personalized context-aware citation recommendation. *Neurocomputing* **2020**, *410*, 103–113. [[CrossRef](#)]
54. Wang, R.; Wu, Z.; Lou, J.; Jiang, Y. Attention-based dynamic user modeling and Deep Collaborative filtering recommendation. *Expert Syst. Appl.* **2022**, *188*, 116036. [[CrossRef](#)]
55. Gurcan, F.; Cagiltay, N.E.; Cagiltay, K. Mapping Human–Computer Interaction Research Themes and Trends from Its Existence to Today: A Topic Modeling-Based Review of past 60 Years. *Int. J. Hum.-Comput. Interact.* **2021**, *37*, 267–280. [[CrossRef](#)]
56. Chhikara, P.; Singh, P.; Tekchandani, R.; Kumar, N.; Guizani, M. Federated Learning Meets Human Emotions: A Decentralized Framework for Human–Computer Interaction for IoT Applications. *IEEE Internet Things J.* **2020**, *8*, 6949–6962. [[CrossRef](#)]
57. Ren, F.; Bao, Y. A review on human-computer interaction and intelligent robots. *Int. J. Inf. Technol. Decis. Mak.* **2020**, *19*, 5–47. [[CrossRef](#)]
58. Miao, H.; Cheng, G.; Zhang, P.; Yan, Y. Online Hybrid CTC/Attention End-to-End Automatic Speech Recognition Architecture. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 1452–1465. [[CrossRef](#)]
59. Liao, Y.-F.; Chang, Y.-H.S.; Lin, Y.-C.; Hsu, W.-H.; Pleva, M.; Juhar, J. Formosa Speech in the Wild Corpus for Improving Taiwanese Mandarin Speech-Enabled Human-Computer Interaction. *J. Signal Process. Syst.* **2020**, *92*, 853–873. [[CrossRef](#)]
60. Ho, N.-H.; Yang, H.-J.; Kim, S.-H.; Lee, G. Multimodal Approach of Speech Emotion Recognition Using Multi-Level Multi-Head Fusion Attention-Based Recurrent Neural Network. *IEEE Access* **2020**, *8*, 61672–61686. [[CrossRef](#)]
61. Hazer-Rau, D.; Meudt, S.; Daucher, A.; Spohrs, J.; Hoffmann, H.; Schwenker, F.; Traue, H.C. The uulmMAC Database—A Multimodal Affective Corpus for Affective Computing in Human-Computer Interaction. *Sensors* **2020**, *20*, 2308. [[CrossRef](#)]
62. Dokuz, Y.; Tufekci, Z. Mini-batch sample selection strategies for deep learning based speech recognition. *Appl. Acoust.* **2021**, *171*, 107573. [[CrossRef](#)]
63. Sun, X.; Yang, Q.; Liu, S.; Yuan, X. Improving Low-Resource Speech Recognition Based on Improved NN-HMM Structures. *IEEE Access* **2020**, *8*, 73005–73014. [[CrossRef](#)]
64. Kumar, M.; Kim, S.H.; Lord, C.; Lyon, T.D.; Narayanan, S. Leveraging Linguistic Context in Dyadic Interactions to Improve Automatic Speech Recognition for Children. *Comput. Speech Lang.* **2020**, *63*, 101101. [[CrossRef](#)] [[PubMed](#)]
65. Li, X.; Zhong, J.; Kamruzzaman, M. Complicated robot activity recognition by quality-aware deep reinforcement learning. *Futur. Gener. Comput. Syst.* **2021**, *117*, 480–485. [[CrossRef](#)]
66. Oudah, M.; Al-Naji, A.; Chahl, J. Hand Gesture Recognition Based on Computer Vision: A Review of Techniques. *J. Imaging* **2020**, *6*, 73. [[CrossRef](#)]
67. Pareek, P.; Thakkar, A. A survey on video-based Human Action Recognition: Recent updates, datasets, challenges, and applications. *Artif. Intell. Rev.* **2021**, *54*, 2259–2322. [[CrossRef](#)]

68. Munea, T.L.; Jembre, Y.Z.; Weldegebriel, H.T.; Chen, L.; Huang, C.; Yang, C. The Progress of Human Pose Estimation: A Survey and Taxonomy of Models Applied in 2D Human Pose Estimation. *IEEE Access* **2020**, *8*, 133330–133348. [[CrossRef](#)]
69. Tsai, T.-H.; Huang, C.-C.; Zhang, K.-L. Design of hand gesture recognition system for human-computer interaction. *Multimedia Tools Appl.* **2020**, *79*, 5989–6007. [[CrossRef](#)]
70. Yang, H.; Gu, Y.; Zhu, J.; Hu, K.; Zhang, X. PGCN-TCA: Pseudo Graph Convolutional Network With Temporal and Channel-Wise Attention for Skeleton-Based Action Recognition. *IEEE Access* **2020**, *8*, 10040–10047. [[CrossRef](#)]
71. Sun, Y.; Weng, Y.; Luo, B.; Li, G.; Tao, B.; Jiang, D.; Chen, D. Gesture recognition algorithm based on multi-scale feature fusion in RGB-D images. *IET Image Process.* **2020**, *14*, 3662–3668. [[CrossRef](#)]
72. Li, J.; Xie, X.; Pan, Q.; Cao, Y.; Zhao, Z.; Shi, G. SGM-Net: Skeleton-guided multimodal network for action recognition. *Pattern Recognit.* **2020**, *104*, 107356. [[CrossRef](#)]
73. Afza, F.; Khan, M.A.; Sharif, M.; Kadry, S.; Manogaran, G.; Saba, T.; Ashraf, I.; Damaševičius, R. A framework of human action recognition using length control features fusion and weighted entropy-variances based feature selection. *Image Vis. Comput.* **2021**, *106*, 104090. [[CrossRef](#)]
74. Chen, Y.; Ma, G.; Yuan, C.; Li, B.; Zhang, H.; Wang, F.; Hu, W. Graph convolutional network with structure pooling and joint-wise channel attention for action recognition. *Pattern Recognit.* **2020**, *103*, 107321. [[CrossRef](#)]
75. Zhu, A.; Wu, Q.; Cui, R.; Wang, T.; Hang, W.; Hua, G.; Snoussi, H. Exploring a rich spatial-temporal dependent relational model for skeleton-based action recognition by bidirectional LSTM-CNN. *Neurocomputing* **2020**, *414*, 90–100. [[CrossRef](#)]
76. Yang, H.; Yuan, C.; Zhang, L.; Sun, Y.; Hu, W.; Maybank, S.J. STA-CNN: Convolutional Spatial-Temporal Attention Learning for Action Recognition. *IEEE Trans. Image Process.* **2020**, *29*, 5783–5793. [[CrossRef](#)] [[PubMed](#)]
77. Jegham, I.; Ben Khalifa, A.; Alouani, I.; Mahjoub, M.A. Vision-based human action recognition: An overview and real world challenges. *Forensic Sci. Int. Digit. Investig.* **2020**, *32*, 200901. [[CrossRef](#)]
78. Qiao, H.; Liu, S.; Xu, Q.; Liu, S.; Yang, W. Two-Stream Convolutional Neural Network for Video Action Recognition. *KSII Trans. Internet Inf. Syst.* **2021**, *15*, 3668–3684.
79. Vishwakarma, D.K. A two-fold transformation model for human action recognition using decisive pose. *Cogn. Syst. Res.* **2020**, *61*, 1–13. [[CrossRef](#)]
80. Tran, D.-S.; Ho, N.-H.; Yang, H.-J.; Baek, E.-T.; Kim, S.-H.; Lee, G. Real-Time Hand Gesture Spotting and Recognition Using RGB-D Camera and 3D Convolutional Neural Network. *Appl. Sci.* **2020**, *10*, 722. [[CrossRef](#)]
81. Chen, W.; Yu, C.; Tu, C.; Lyu, Z.; Tang, J.; Ou, S.; Fu, Y.; Xue, Z. A Survey on Hand Pose Estimation with Wearable Sensors and Computer-Vision-Based Methods. *Sensors* **2020**, *20*, 1074. [[CrossRef](#)]
82. Ozcan, T.; Basturk, A. Human action recognition with deep learning and structural optimization using a hybrid heuristic algorithm. *Clust. Comput.* **2020**, *23*, 2847–2860. [[CrossRef](#)]
83. Khan, M.A.; Sharif, M.; Akram, T.; Raza, M.; Saba, T.; Rehman, A. Hand-crafted and deep convolutional neural network features fusion and selection strategy: An application to intelligent human action recognition. *Appl. Soft Comput.* **2020**, *87*, 105986. [[CrossRef](#)]
84. Seinfeld, S.; Feuchtner, T.; Maselli, A.; Müller, J. User Representations in Human-Computer Interaction. *Hum.-Comput. Interact.* **2021**, *36*, 400–438. [[CrossRef](#)]
85. Aly, S.; Aly, W. DeepArSLR: A Novel Signer-Independent Deep Learning Framework for Isolated Arabic Sign Language Gestures Recognition. *IEEE Access* **2020**, *8*, 83199–83212. [[CrossRef](#)]
86. Wan, S.; Qi, L.; Xu, X.; Tong, C.; Gu, Z. Deep Learning Models for Real-time Human Activity Recognition with Smartphones. *Mob. Netw. Appl.* **2020**, *25*, 743–755. [[CrossRef](#)]
87. Qiu, X.; Sun, T.; Xu, Y.; Shao, Y.; Dai, N.; Huang, X. Pre-trained models for natural language processing: A survey. *Sci. China Technol. Sci.* **2020**, *63*, 1872–1897. [[CrossRef](#)]
88. Maulud, D.H.; Zeebaree, S.R.; Jacksi, K.; Sadeeq, M.A.M.; Sharif, K.H. State of art for semantic analysis of natural language processing. *Qubahan Acad. J.* **2021**, *1*, 21–28. [[CrossRef](#)]
89. Otter, D.W.; Medina, J.R.; Kalita, J.K. A Survey of the Usages of Deep Learning for Natural Language Processing. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 604–624. [[CrossRef](#)]
90. Sullivan, F.R.; Keith, P.K. Exploring the potential of natural language processing to support microgenetic analysis of collaborative learning discussions. *Br. J. Educ. Technol.* **2019**, *50*, 3047–3063. [[CrossRef](#)]
91. Narechania, A.; Srinivasan, A.; Stasko, J. NL4DV: A Toolkit for Generating Analytic Specifications for Data Visualization from Natural Language Queries. *IEEE Trans. Vis. Comput. Graph.* **2020**, *27*, 369–379. [[CrossRef](#)]
92. Alexakis, G.; Panagiotakis, S.; Fragkakis, A.; Markakis, E.; Vassilakis, K. Control of Smart Home Operations Using Natural Language Processing, Voice Recognition and IoT Technologies in a Multi-Tier Architecture. *Designs* **2019**, *3*, 32. [[CrossRef](#)]
93. Gu, Y.; Tinn, R.; Cheng, H.; Lucas, M.; Usuyama, N.; Liu, X.; Naumann, T.; Gao, J.; Poon, H. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthc.* **2021**, *3*, 1–23. [[CrossRef](#)]
94. Feder, A.; Keith, K.A.; Manzoor, E.; Pryzant, R.; Sridhar, D.; Wood-Doughty, Z.; Eisenstein, J.; Grimmer, J.; Reichart, R.; Roberts, M.E.; et al. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Trans. Assoc. Comput. Linguist.* **2022**, *10*, 1138–1158. [[CrossRef](#)]
95. Kang, Y.; Cai, Z.; Tan, C.-W.; Huang, Q.; Liu, H. Natural language processing (NLP) in management research: A literature review. *J. Manag. Anal.* **2020**, *7*, 139–172. [[CrossRef](#)]

96. Zhang, W.E.; Sheng, Q.Z.; Alhazmi, A.; Li, C. Adversarial attacks on deep-learning models in natural language processing: A survey. *ACM Trans. Intell. Syst. Technol.* **2020**, *11*, 1–41. [[CrossRef](#)]
97. Zeng, Z.; Deng, Y.; Li, X.; Naumann, T.; Luo, Y. Natural language processing for EHR-based computational phenotyping. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2018**, *16*, 139–153. [[CrossRef](#)]
98. Yunanto, A.A.; Herumurti, D.; Rochimah, S.; Kuswardayan, I. English Education Game using Non-Player Character Based on Natural Language Processing. *Procedia Comput. Sci.* **2019**, *161*, 502–508. [[CrossRef](#)]
99. Pramanick, P.; Sarkar, C.; Banerjee, S.; Bhowmick, B. Talk-to-Resolve: Combining scene understanding and spatial dialogue to resolve granular task ambiguity for a collocated robot. *Robot. Auton. Syst.* **2022**, *155*, 104183. [[CrossRef](#)]
100. El-Komy, A.; Shahin, O.R.; Abd El-Aziz, R.M.; Taloba, A.I. Integration of computer vision and natural language processing in multimedia robotics application. *Inf. Sci. Lett.* **2022**, *11*, 9.
101. Recupero, D.R.; Spiga, F. Knowledge acquisition from parsing natural language expressions for humanoid robot action commands. *Inf. Process. Manag.* **2020**, *57*, 102094. [[CrossRef](#)]
102. Nistor, A.; Zadobrischi, E. The Influence of Fake News on Social Media: Analysis and Verification of Web Content during the COVID-19 Pandemic by Advanced Machine Learning Methods and Natural Language Processing. *Sustainability* **2022**, *14*, 10466. [[CrossRef](#)]
103. Wang, D.; Su, J.; Yu, H. Feature Extraction and Analysis of Natural Language Processing for Deep Learning English Language. *IEEE Access* **2020**, *8*, 46335–46345. [[CrossRef](#)]
104. Sun, B.; Li, K. Neural Dialogue Generation Methods in Open Domain: A Survey. *Nat. Lang. Process. Res.* **2021**, *1*, 56. [[CrossRef](#)]
105. Li, Y.; Ishi, C.T.; Inoue, K.; Nakamura, S.; Kawahara, T. Expressing reactive emotion based on multimodal emotion recognition for natural conversation in human–robot interaction. *Adv. Robot.* **2019**, *33*, 1030–1041. [[CrossRef](#)]
106. Jia, K. Chinese sentiment classification based on Word2vec and vector arithmetic in human–robot conversation. *Comput. Electr. Eng.* **2021**, *95*, 107423. [[CrossRef](#)]
107. Korpusik, M.; Glass, J. Deep Learning for Database Mapping and Asking Clarification Questions in Dialogue Systems. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2019**, *27*, 1321–1334. [[CrossRef](#)]
108. Chang, Y.-C.; Hsing, Y.-C. Emotion-infused deep neural network for emotionally resonant conversation. *Appl. Soft Comput.* **2021**, *113*, 107861. [[CrossRef](#)]
109. Marge, M.; Rudnicky, A.I. Miscommunication Detection and Recovery in Situated Human–Robot Dialogue. *ACM Trans. Interact. Intell. Syst.* **2019**, *9*, 1–40. [[CrossRef](#)]
110. Permatasari, D.A.; Maharani, D.A. Combination of Natural Language Understanding and Reinforcement Learning for Booking Bot. *J. Electr. Electron. Inf. Commun. Technol.* **2021**, *3*, 12–17. [[CrossRef](#)]
111. Ghiță, A.; Gavril, A.F.; Nan, M.; Hoteit, B.; Awada, I.A.; Sorici, A.; Mocanu, I.G.; Florea, A.M. The AMIRO Social Robotics Framework: Deployment and Evaluation on the Pepper Robot. *Sensors* **2020**, *20*, 7271. [[CrossRef](#)]
112. Rofi'ah, B.; Fakhurroja, H.; Machbub, C. Dialogue management using reinforcement learning. *TELKOMNIKA Telecommun. Comput. Electron. Control* **2021**, *19*, 931–938. [[CrossRef](#)]