

Received 14 May 2022, accepted 30 June 2022, date of publication 4 July 2022, date of current version 28 July 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3188264

SURVEY

Robotics Utilization in Automatic Vision-Based Assessment Systems From Artificial Intelligence Perspective: A Systematic Review

RAFEEF FAUZI NAJIM ALSHAMMARI^{1,2}, HASLINA ARSHAD³,
ABDUL HADI ABD RAHMAN⁴, AND O. S. ALBAHRI⁵

¹Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), Bangi, Selangor 43600, Malaysia

²College of Science, University of Kerbala, Karbala 56001, Iraq

³Institute of IR4.0, Universiti Kebangsaan Malaysia (UKM), Bangi, Selangor 43600, Malaysia

⁴Center for Artificial Intelligence Technology, Universiti Kebangsaan Malaysia (UKM), Bangi, Selangor 43600, Malaysia

⁵Computer Techniques Engineering Department, Mazaya University College, Nasiriyah 64001, Iraq

Corresponding author: O. S. Albahri (osamahsh89@gmail.com)


This work was supported by the Universiti Kebangsaan Malaysia (UKM) under Grant TAP-K009683 and Grant TT-2021-001.

ABSTRACT This study presents a systematic review of the automatic vision-based assessment systems and models from related fields over the last 6 years. Many studies focus on automatic vision-based assessment in the area of robotics, artificial intelligence and virtual reality or augmented reality (AR) to enhance the communication between humans and machines for general purposes. Three reliable databases, IEEEExplore, Science Direct and Web of Science, were used to obtain relevant studies on the given topic. Several stages of filtering and scanning were applied according to the inclusion/exclusion criteria to filter the obtained 3505 papers from 2015 to 2020. Finally, 97 papers met the criteria. They were classified into four main categories based on their field by following the scientific taxonomy: computer-vision-based, robotics-based, AR-based and hybrid-based categories, accounting for 42.26% ($n = 41/97$ papers), 48.45% ($n = 47/97$ papers), 6.18% ($n = 6/97$ papers) and 4.12% ($n = 4/97$ papers), respectively. Subsequently, a deep and critical analysis of this multifield systematic review highlighted new research opportunities, motivations, challenges and recommendations that need attention to integrate interdisciplinary studies. Thus, automatic vision-based assessment, which is a field requiring automated solutions, tools and methods, enhances the ability of the assistive technology and facilitates the interaction of individuals with machines. Many studies have been conducted on the automatic vision-based assessment systems and their subtypes to promote accurate communication and performance evaluation in human-machine interaction. This study can provide researchers with useful guides and valuable information for future research. This study also addresses the ambiguity of automatic vision-based assessment models in interdisciplinary fields.

INDEX TERMS Artificial intelligence, assessment, engagement, evaluation, robotics.

I. INTRODUCTION

Robotics has been a fast-developing field that has attracted considerable interest in education [1]. Nowadays, researchers focus their efforts on developing robots that can assist people with various activities [2]. Developing intelligent robots with consistent behaviours, robustness and flexibility in service provisioning and the ability to adapt to varied circumstances

The associate editor coordinating the review of this manuscript and approving it for publication was Kostas Kolomvatsos .

and requirements is essential. Socially assistive robotics is a relatively recent research field that seeks to develop robots that can assist users. Socially assistive robots (SARs) aim to provide continuous support for individuals by adopting appropriate emotional, cognitive and social cues for assistance purposes [2]. The design of robots based on artificial intelligence (AI) is regarded as a link between robotics and AI. AI and humanoid robots are not science fiction but realistic inventions rapidly used in innovation studies and interdisciplinary applications in education, hospitality and healthcare [1].

AI includes the development of computer systems that perform tasks automatically without human intervention [3]. Numerous AI applications have been used in people's lives. Many AI-based applications support humans in many scenarios. The incorporation of AI in robots aids people in productivity improvement by making work safe and saving important time and training purposes [3]. AI technologies have been implemented through deep learning algorithms to comprehend human behaviour thoroughly [4]. Additionally, deep learning of artificial neural networks is required to enable robots to achieve complex tasks [5]. Multiple AIs are integrated into robotics, allowing robots to perform tasks, such as semantic recognition, emotional recognition, data mining and analysis [6].

The developments in machine learning and AI have achieved remarkable progress in computer vision and its applications [7]. Computer vision and machine learning, which addressed some of the standard image processing problems, such as image recognition, present new unique image processing problems, such as object detection and object segmentation [8]. Integrating a robot with various computer vision capabilities, such as face detection and recognition, body tracking and gesture detection, proposes a fully functional intelligent robot by utilising multiple algorithms and performance parameters [8]. A robust emotional awareness is essential in completing the assisting task in human-robot interaction (HRI). Facial expression, one of the most prevalent nonverbal techniques that humans employ to convey internal emotional states, plays an integral part in interpersonal relationships. An essential aspect of HRI is the capacity of a robot to recognise a person's emotional state automatically [9]. Over the last 2 decades, several algorithms for facial recognition have been developed with high accuracy. Adding visual intelligence to an active tracking robot improves its capabilities, such as the intelligence of detecting objects in the images and reading texts from them [8]. Furthermore, gesture recognition and human behaviour analysis, including viewing and understanding their surroundings and the humans that inhabit them, are extremely valuable for robotic systems [10].

The use of social robots in education has become increasingly popular in recent years. The demand for a teaching assistant robot (TA robot) that can work on various tasks to support teachers' requirements in the classroom has greatly increased, particularly when the tasks are difficult to achieve only by the teachers themselves [11]. As teaching assistants, robots can assist learners in answering their questions and grading their responses in natural and responsive communications. The robot is embodied with an emotion recognition framework to evaluate students through facial expressions and improve their learning effectiveness [5]. The majority of current research attempts to enhance the contemporary educational method by incorporating real-world scenarios and studies [12], [13].

Education is the most important element in achieving a sustainable life because it allows individuals to learn new knowledge and academic skills. In recent years, robots have

been frequently utilised for education as robots teaching second language, such as the Chinese language [1], education robots assisting therapy sessions for children with communication disorders [14] and robots motivating children to learn Science, Technology, Engineering and Mathematics (STEM) concepts [15].

Motivation plays a vital role in students' learning process. Moreover, robots can stimulate the curiosity and imagination of students. Robots with human characteristics are attractive to individuals. During a robot-student interaction, robots become learning partners of students, enabling students to understand scientific knowledge or to practise social skills. Moreover, robots can guide students' behaviour by pronouncing proper words and allowing students to regard these robots as real learning companions [6]. Students argue that the classical classroom is boring and not engaging [1]. Considerable research indicated that robots are engaging and motivating for learners. The employment of educational robots is considered beneficial in increasing learners' motivation and performance during the learning process [16]. Moreover, the TA robot can evaluate student performance to support instructors in their classes [17].

The research interest in developing assistive technology to conduct real-time activities, including automated vision-based assessment, is growing. Considerable research has been performed to develop methods for the automatic detection and classification of human behaviour in various environments, such as domestic [18], education [5], [17], [19] and healthcare [20], [21]. Automated vision-based assessment is important for human-machine interactions requiring a highly accurate classifier [10]. This study aims to conduct a systematic review and determine the major concerns and challenges of the automated vision-based assessment in three domains: robotics, AI and virtual reality (VR)/augmented reality (AR) or mixed reality.

This study presents a systematic analysis of researchers' achievements and summarises previous works in response to the need for automated vision-based assessment systems. Furthermore, the study provides a clear overview by specifying the motivation for utilising various techniques and methods in the automated vision-based assessment and the challenges and recommended solutions for solving the existing problems in further research. It also suggests a taxonomy of literature and clarifies clusters of various aspects in this relevant research area. A systematic review protocol is initially utilised to cover all academic research in automatic vision-based assessment from all the scientific perspectives. This study provides additional ideas, which can present a valued chance for the researchers to solve the current research gaps by using new approaches.

This study includes seven sections with corresponding subsections. Section 2 presents the information sources, study phases, search strategy, inclusion and exclusion criteria, article analysis process and the article search results of the automated vision-based assessment. Section 3 includes the taxonomy and results of the automated vision-based

assessment studies. Moreover, Section 4 presents the discussion of this study, which includes the datasets, motivation, challenges and recommendations of the studies in our survey. Additionally, Section 5 presents the statistical findings of the reviewed papers. Finally, Section 6 concludes the paper.

II. SYSTEMATIC REVIEW PROTOCOL

A. INFORMATION SOURCES

Systematic search approaches were performed to collect relevant articles for the current topic by searching three databases: Web of Science (WOS), Science Direct (SD) and IEEEExplore digital library. These databases facilitate and formulate a basic and complicated search query and keep track of various journals and conference papers in computer science, robotics and AI.

B. STUDY PHASES

The procedure of the study involved an extensive search for relevant publications and depended on three phases. The first phase involved the review of the titles and abstracts of articles to eliminate unrelated or duplicated papers. The second phase required the full reading of screened papers from the first phase. Finally, the related papers were organised into groups in Mendeley to facilitate the referencing process of our review.

C. SEARCH STRATEGY

A comprehensive literature search was implemented by running our search query via the search boxes in the WOS, SD and IEEE databases from 2015 to 2020. The present study performed a Boolean search strategy using various keywords related to automated vision-based assessment approaches. In all indices, the searches were performed by entering the search terms as the string ‘Robot’ OR ‘Robotics’ followed by the keywords ‘Student Assessment’ OR ‘Teaching Assistant’ OR ‘Student Evaluation’ OR ‘Performance Assessment’ OR ‘Performance Evaluation’. We also utilised ‘Artificial Intelligence’ OR ‘Artificial Intelligence’ OR ‘Machine Learning’ to limit the scope of our searches into automated vision-based assessment concepts that relate to computer science and AI. Figure (1) clarifies the details of our search query. Articles and conference papers were selected using the advanced search options in each search engine. We considered the recent scientific research in the English language published in journals and conferences relevant to our article.

D. INCLUSION CRITERIA

Some criteria that should be included in the research results were specified in the search setting. The article should be an English journal and a conference paper. The main focus of the search includes the studies that are review or survey, evaluation, benchmarking and development of an application for automated vision-based assessment in the fields of AI, robotics and VR/AR. The majority of the reviewed papers

were published within the last 5 years (2015–2020). Figure. 1 depicts all articles that met the inclusion criteria.

E. EXCLUSIVE CRITERIA

After the omission of duplicate articles, all papers that failed to meet the eligibility criteria were excluded, as illustrated in Figure 1. The exclusion criteria are as follows:

- The articles are written in a language other than English.
- The publication is a book chapter or belongs to other types of papers.
- The articles that are not related to the scope of our study, i.e. articles that do not focus on automation assessment methods based on AI, robotics or AR.

F. ARTICLE ANALYSIS PROCESS

The list of all the included papers from the three databases was organised in an Excel file. The authors have read the articles in their entirety, highlighted a substantial number of comments on the reviewed publications and classified these articles to build the taxonomy for this study. The key findings have been summarised, described and tabulated. The important information, including the list of reviewed articles, source indices, objectives, motivation, challenges, methodologies, datasets utilised, evaluation criteria, validation approaches and recommendations, was saved in Word and Excel formats.

G. ARTICLE SEARCH RESULTS

The papers that meet our inclusion criteria were included in our study, as shown in Fig. (1). The first initial search included 3505 papers, and 572, 1129 and 1804 of which were from WOS, SD and IEEEExplore, respectively. The three databases had approximately 49 duplicate publications, resulting in 3456 studies. In addition, 3246 papers were omitted from the scope after reviewing their titles and abstracts, resulting in 210 papers. Finally, 97 papers were chosen based on the full-text reading. These publications were carefully read to construct a complete study subject map.

III. AUTOMATED VISION-BASED ASSESSMENT TAXONOMY AND RESULTS

This research established a generic taxonomy for automated vision-based assessment, including several groups. These groups were discovered during a preliminary literature review. The goal of the systematic review is to provide an overview of the assessment model based on automated vision. All the separated papers were read, grouped into their related classes and saved into an Excel file. The grouping process of all papers followed the formation of scientific taxonomy. Then, the final papers from the three databases were categorised into four: computer-vision-based (41/97), robotics-based (47/97), AR-based (6/97) and hybrid-based (4/97). The taxonomy study led to the identification of patterns and the specification of main categories and subcategories.

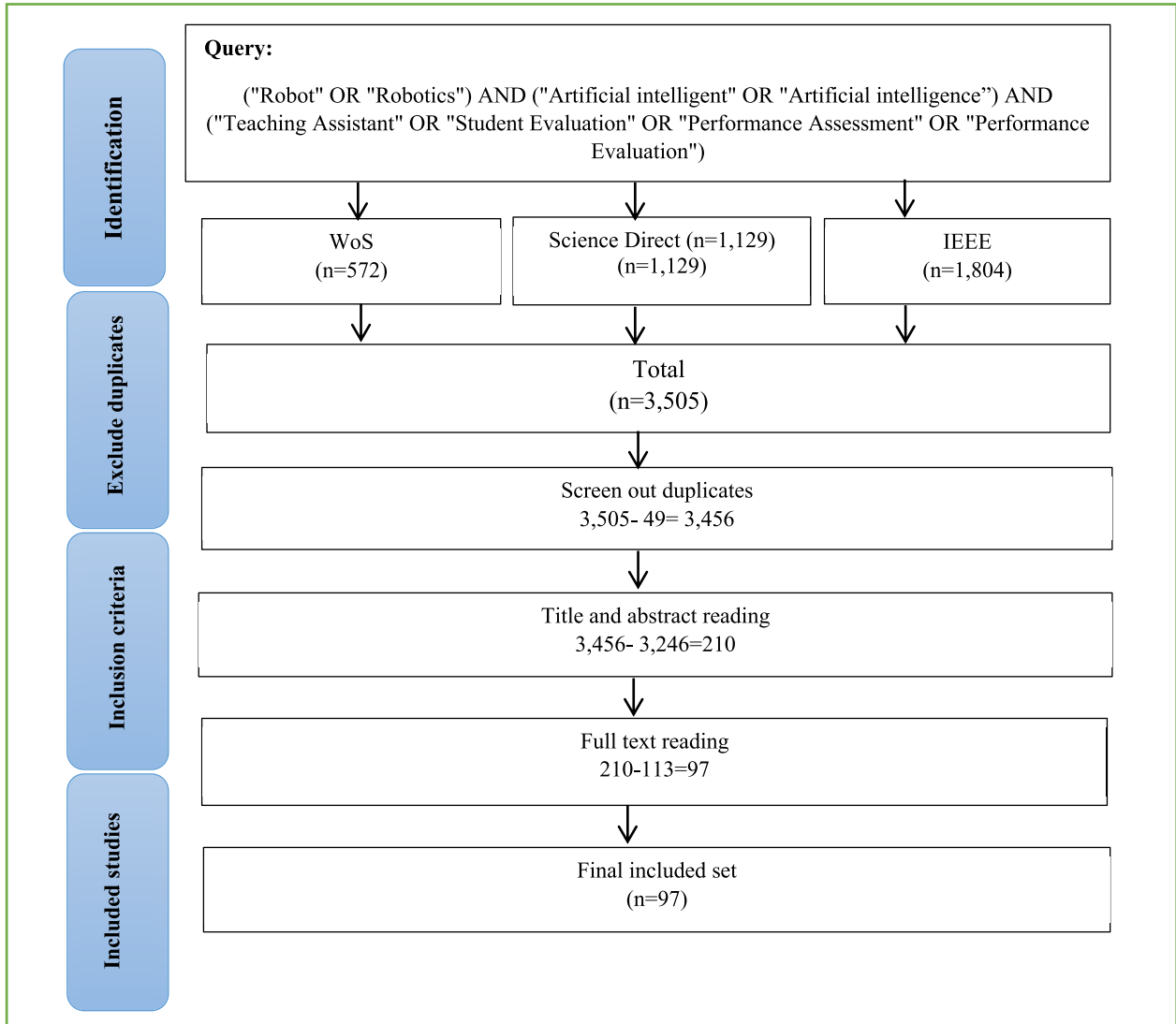


FIGURE 1. Systematic technique for identification, exclusion and inclusion of related articles.

Figure 2 illustrates the research taxonomy for automated vision-based assessment.

A. COMPUTER-VISION-BASED CATEGORY

Computer vision applications provide extensive possibilities for human emotional recognition in computer science. The emotions and feedback of individuals are transmitted by their nonverbal cues, including facial expressions, body movements and hand gestures. Recognising human emotions and behaviours automatically aids in teaching social intelligence to machines. Out of 97 papers included in this study, 41 works were under this category. The works on computer vision focus on behaviour analysis. The behaviour subcategory was divided into three sections: face-based (34/41), body-based (2/41) and gesture-based (2/41).

B. COMPUTER-VISION-BASED CATEGORY

Computer vision applications provide extensive possibilities for human emotional recognition in computer science. The

emotions and feedback of individuals are transmitted by their nonverbal cues, including facial expressions, body movements and hand gestures. Recognising human emotions and behaviours automatically aids in teaching social intelligence to machines. Out of 97 papers included in this study, 41 works were under this category. The works on computer vision focus on behaviour analysis. The behaviour subcategory was divided into three sections: face-based (34/41), body-based (2/41) and gesture-based (2/41).

1) BEHAVIOUR-BASED

Numerous studies on the automatic detection of human behaviours and information in images and video content were conducted. The emphasis is on discovering common behavioural patterns in the available literature. Automated recognition of human emotional states is a challenging task. Computer vision algorithms distinguish various affective states through facial expressions, hand gestures and body

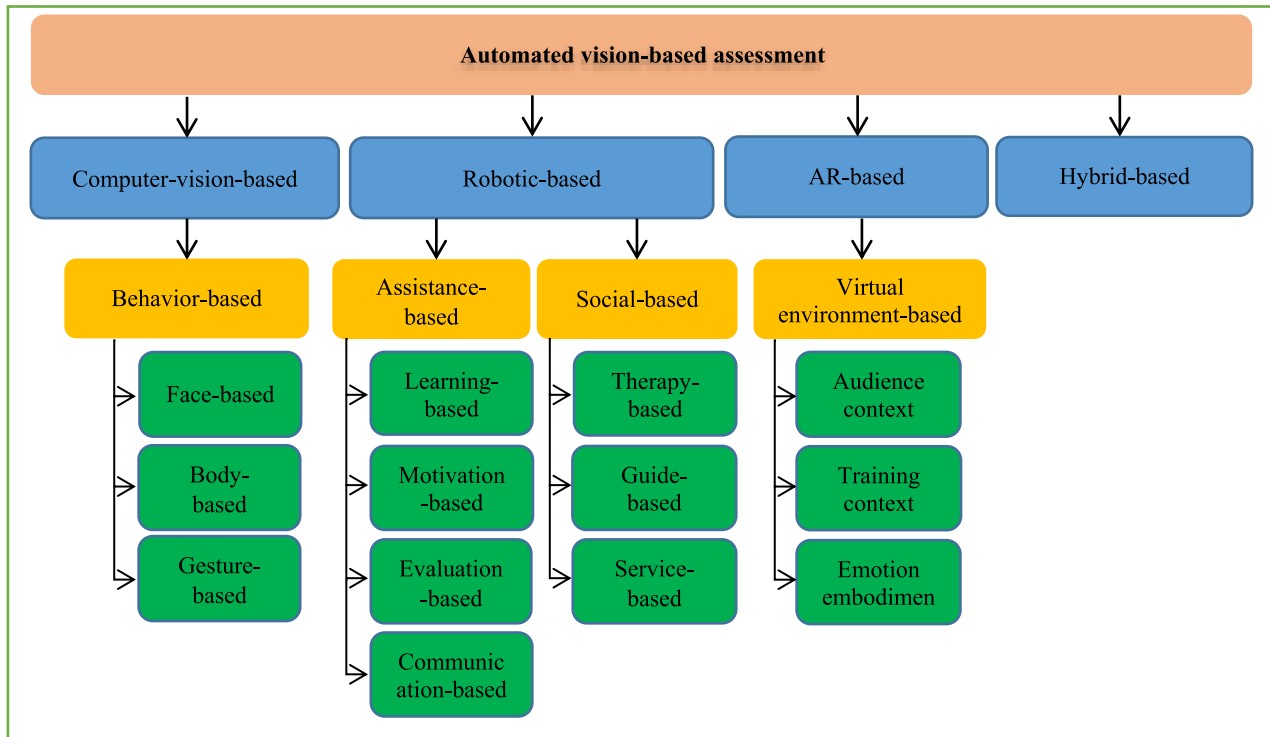


FIGURE 2. Automated vision-based assessment model taxonomy of the research literature.

postures. These strategies demand a suitable database for training, testing and validation [22]. Thus, a study [22] established a database for recognising human affective states with object location by utilising three aspects: facial expression, body posture and hand gestures. In [23] proposed an automated methodology for student behaviour detection. This methodology can be employed in student presentations. The model was developed based on computer vision approaches and machine learning techniques in combination with video content to aid in student assessment. The study utilised pattern recognition methods to investigate and analyse human behaviours and discover relationships between human behaviours and personal modalities [23]. A new class of support vector machines (SVMs) for multiclass classification has been developed to recognise human activities in unconstrained scenarios and address face image and standard classification issues by leveraging geometric information contained in the training data. Direct solutions to the optimisation problem have been proposed in linear and nonlinear settings [24].

As a part of computer vision, which deals with the vision of human body language, the behaviour-based subcategory can be divided into three sections:

2) FACE-BASED

Automatic emotion detection may enhance the human computer interaction (HCI) experience by responding to the mental state of the individuals, which is at the core of

advanced technology. Paul Ekman and his colleagues defined the fundamental human emotions. Their research influenced the current state of emotion analysis. Their work included classifications of happiness, sadness, anger, fear, surprise, disgust and neutral [25]. Automatic analysis of the human emotional states can be implemented with computer vision (CV) algorithms and approaches from frontal face images. One of the most popular classification issues is video facial expression detection and recognition, which is expected to become increasingly important in the fields of robotics and autonomous mobility.

The innovative approach described in [26] employs novel video pre-processing techniques in conjunction with a convolutional neural network (CNN) to analyse human emotions based on facial expression recognition (FER). In [27], the authors suggested a system for imitating face cues using a CNN. They implemented it on a robotic platform to enable the robot to mimic its user's emotional states when this robot looks into the user's face. Moreover, a study [28] developed a CNN-based model for real-time emotion-score recognition. This model captures the learners' learning image via a webcam, judges their learning emotions in real time and sends feedback to the lecturer, enabling real-time emotion feedback in distant education [28]. The model was aimed at improving teacher-learner interaction, thereby improving learning efficiency and helping personalise education.

An automatic model was developed to recognise academic confusion in online learning based on facial expressions to

improve the learning effect [29]. An automated class attendance assessment system was offered as a solution for student attendance regulation based on facial recognition and deep one-shot learning, and its method was evaluated in a real-world scenario using various situations and image capturing equipment [30]. Another study adopted a facial recognition technology to regulate attendance [31]. The research investigated the types of facial expression in teaching (FET) and the applicability of CNN algorithm modifications to address four potential teaching difficulties, including low-resolution images, head posture variability, occlusion and lighting condition fluctuation [31]. An automated uncertainty identification system based on facial expressions was developed by utilising the annotated facial-based uncertainty corpus database [32]. This work aimed to investigate the uncertainty discriminators from facial expressions in a learning environment using the Facial Action Coding System (FACS) to improve the study process and learning efficiency [32].

A novel deep convolution network-based emotion analysis framework was introduced to aid in detecting and diagnosing mental states by analysing users' facial expressions from facial images [33]. The system can interpret face images and analyse the temporal evolution of emotions using a novel approach that extracts deep features from AlexNet's Fully Connected Layer 6 and employs a standard linear discriminant analysis classifier to produce the final classification result [33]. A unique three-dimensional (3D) CNN architecture, called the Multiscale Spatiotemporal Network (MSN), was developed to represent accurately the face information associated with sad behaviours captured in images or videos and facilitate automated depression diagnosis [34]. Additionally, an automated criminal identification system (CIS) that utilises deep learning-based CNN techniques was demonstrated to combine facial recognition, emotion detection, suspicious suit detection and age and gender identification of suspects [35].

Human factors are essential to developing an effective user-centred adaptive system. The influence of the internal state, particularly prior knowledge, on mutual gaze convergence between two persons was demonstrated during a lecture assignment with and without prior knowledge interaction. The main goal was to examine the differences in gaze behaviour between the two scenarios and determine whether prior knowledge directly impacts human gaze behaviour [36]. A recognition classifier for FER in assessing teaching skill (ATS) system was developed using CNN, modified CNN [31] and SVM [37]. The system of intelligent education, namely, Smart-E, aims to identify and comprehend learners' emotional states during the learning process by detecting changes in their facial expressions and blinking frequency [37]. It assists in providing the basis for the analysis of teachers' teaching effects. A novel approach was employed to detect the emotion based on the lip structure over a period and train the machine to understand nonverbal communication in the form of emotions based on Recurrent Neural Network (RNN) approaches [38]. For specific attention to facial pose

and illumination invariance, a unique deep CNN architecture was developed to pre-train as a stacked convolutional auto encoder (SCAE) using an improved version of the Greedy Layer-Wise (GLW) algorithm model for emotion recognition in unconstrained contexts [39], [40].

The feature extraction and feature selection approaches were performed in considerable research to develop facial expression recognition systems. Numerous studies proposed the construction of fully automated or semiautomatic detection and classification methods. The study [41] described a semi supervised emotion detection system using reduced features and a novel approach for feature selection. A feature sparseness-based regularisation strategy was presented to learn deep features with sufficient capacity for generalisation [42]. A novel methodology, called task-dependent multitask multiple kernel learning (TD-MTMKL), was utilised to identify the absence and presence of various action units (AUs) concurrently [43]. A novel image pre-processing technique that combines optimised pre-processing filters with a CNN-SVM classifier was implemented [44]. A unique Monogenic Directional Pattern (MDP) was proposed to speed up the selection of the optimum kernel [45]. The commonly used kernel for dimension reduction is a Generalised Discriminant Analysis (GDA) based on a unique pseudo-Voigt kernel (PVK) [45]. A new spatiotemporal restricted boltzmann machines (RBM)-based model was developed to learn the correlations (or transformations) quickly between image pairings associated with distinct facial emotions [46]. A novel contribution was made by combining back-propagation neural network (BPNN), discrete cosine transform (DCT) and histograms of oriented gradient (HOG) to complete the task of face classification via feature extraction [47]. SVMs were employed to train the data using Gabor wavelets and classification in [48]. Additionally, a unique approach to facial recognition was developed based on an extreme learning machine (ELM) to impose the l-norm on the hidden weight matrix and identify the active neurons, resulting in a simple network topology [48].

A coarse-to-fine architecture was suggested for robust face alignment to integrate the stacked hourglass network (SHN) with salient region attention refinement. Additionally, a novel CNN, called the SHN, was used to perform the heat map regression and predict initial facial landmarks [49]. The proposed model is novel because it incorporates a module for multiscale region learning and attention to salient regions [49].

Deep CNN was employed to achieve automatic facial expression recognition in [31], [50]–[54] and build intelligent systems that identify the seven basic human emotions. Moreover, a complete study was proposed to apply HOG descriptors to automatic facial expression identification [9]. This article discusses a group of HOG parameters that can characterise facial expression features well [9]. Furthermore, an automatic facial expression recognition chip for an assistant robot based on the local binary pattern (LBP) features and K-nearest neighbour (KNN) regression was developed,

introduced and tested on an embedded device, such as an FPGA development board [25]. The LBP algorithm was integrated with advanced image processing techniques, such as contrast adjustment, bilateral filtering, histogram equalisation and image blending, to create a novel method for human face recognition. The researchers aimed to enhance the LBP codes and, consequently, the face recognition system's overall accuracy by addressing many factors affecting face recognition accuracy [55]. Additionally, a system for reliable face recognition utilising SVM classification was developed to learn each image subset from low-quality Kinect-collected images. The collected images demonstrated various head poses, illumination, face emotions, sunglass disguise and hand-occlusive occlusions that were utilised to assess the proposed approach [56].

3) BODY-BASED

Body motion tracking is a vital computer vision research area that can provide data for body movement analysis, scene interpretation and behaviour identification. The tracking of human body movement has a broad application prospect based on various fields, such as intelligent surveillance, robot visual navigation, motion recognition and HCI. A study [57] explored in detail the computer vision tracking technology-based model of a human body in sports. This study aimed to examine the usage of many unlabelled samples and aid the training classifier in increasing its performance; the examination was followed by an analysis of the human body in sports [57]. A comprehensive semantics-based emotion detection approach was proposed [58]. This technique uses semantic rules to identify and recognise human behaviour under stressful conditions. It focuses on four human behaviours under stress: relaxation, hands-on-forehead (tensed), loitering and fidgeting in sitting and standing postures [58].

4) GESTURE-BASED

Given the technological advances, considerable effort was expended to develop intuitive interfaces for human-machine interaction, such as dynamic gesture recognition. At present, the most prevalent technique enables dynamic gesture detection by utilising multimodal data, such as colour, depth and skeletal information [59]. RGB cameras are commonly found in nearly every public place; thus, they can be used for gesture recognition without additional equipment. A novel star technique was created for a dynamic gesture recogniser that relies entirely on colour information extracted from each input video, which is an RGB image. This technique can also encode additional temporal information. This dynamic gesture classifier comprises an ensemble of trained CNNs fused together using a soft-attention method [59]. A gesture recognition approach was proposed to recognise gestures accurately and rapidly against a complex background [60]. The suggested approach was developed using a deep convolutional network, which consists of three modules: a basic network module for extracting feature information, squeeze-and-excitation networks for increasing

the affinity of feature channels and a feature pyramid attention module for fusing context information with varying scales [60].

The studies included in our survey evaluated classification methods using evaluation criteria. This approach is the most frequently used method for evaluating various classifiers in AI, and the efficacy of classification approaches in the computer-vision-based category can be clarified from this aspect. The classifiers' performance was measured using various metrics, including accuracy, confusion matrix, precision, recall (sensitivity), F-score and error rate (E-training and E-validation). All of these parameters are summarised in Table 1. The following list provides brief definitions of these evaluation criteria.

1. **Accuracy:** It is a crucial assessment metric because it quantifies the performance of the algorithms or the importance of their behaviour. It primarily assesses the quality of a classification system and how its value varies according to the selected datasets [61].
2. **Confusion matrix:** This matrix is used to describe the performance of classification models. Its values represent the classification class predicted and implemented by the classification models. The confusion matrix compares the discrepancy between incorrect and correct predictions to the actual test sample findings. The binary classification model generates a confusion matrix with four expected outcomes: true positives (TPs), true negatives (TNs), false positives (FPs) and false negatives (FNs) [61].
3. **Precision:** It measures the exactness of the classifier. It is equal to the total of the TPs over the total of TPs added to the total of FPs. Precision provides the proportion of subjects that are accurately recognised as positive outcomes. High precision means few FPs [61].
4. **Recall (sensitivity):** It is defined as the rate of sensitivity or TP. It represents the capacity of a test to detect positive outcomes. It also measures the capacity of classification models to specify instances of a particular class from a dataset [61].
5. **Time complexity:** It is an important matrix for evaluating and benchmarking classification models. It is critical when determining a system's performance efficiency. It is also one of the difficulties encountered by researchers because the best classification model obtains the lowest time complexity whilst maintaining a high level of accuracy [61].
6. **F-score:** It refers to a classification model's overall performance by calculating the harmonic mean of its precision and recall [61].
7. **Error rate:** It is a primary criterion for rating and benchmarking classification systems. It quantifies the classification model's errors within the dataset group. The optimal classification model is determined by the error rate outcomes because the optimal outcome of a classifier is stated in terms of the error rate measure on the sets of the training and validation. Thus, a small error rate indicates an accurate categorisation model [61].

TABLE 1. Evaluation matrices adopted in the reviewed articles under the computer-vision-based category.

Ref	Reliability									Time complexity	Error Rate	
	Accuracy	Precision	Specificity	Recall	Confusion Matrix				Behaviour of parameters		E-Training	E-validation
					TP%	TN%	FP%	FN%	F-score			
[27]	*				*	*	*	*				
[28]	*										*	
[22]	*	*		*					*			
[23]	*											
[29]	*	*		*	*	*	*	*	*		*	*
[30]	*				*	*	*	*	*		*	*
[36]	*									*		
[32]	*											
[9]	*	*		*					*	*	*	
[25]	*									*		
[57]	*											
[48]	*									*	*	*
[58]	*				*		*	*		*		
[26]	*									*	*	*
[45]	*				*	*	*	*		*	*	*
[46]	*									*	*	*
[55]	*				*	*	*	*		*	*	*
[56]	*									*	*	*
[47]	*				*	*	*	*		*	*	*
[33]	*									*	*	*
[59]	*	*		*	*	*	*	*	*	*	*	*
[24]	*	*								*	*	*
[60]	*			*					*	*	*	*
[41]	*				*	*	*	*		*	*	*
[42]	*		*	*	*	*	*	*		*	*	*
[43]		*		*					*	*	*	*
[44]	*									*	*	
[50]	*									*	*	
[34]	*									*	*	
[35]	*									*	*	*
[38]	*	*	*	*	*	*	*	*	*	*	*	
[39]	*									*	*	
[51]	*	*	*	*					*	*	*	*
[52]	*				*	*	*	*		*	*	*
[53]	*				*	*	*	*				*
[54]	*				*	*	*	*			*	*
[37]	*									*	*	
[40]	*				*	*	*	*		*	*	
[49]	*										*	
Total	38	8	3	9	15	14	15	15	9	28	30	20
%	97.4%	20.5%	7.7%	23%	38.5%	35.9%	38.5%	38.5%	23%	71.8%	76.9%	51.3%

The assessment and benchmarking criteria and sub-criteria utilised to evaluate various studies according to the category of computer-vision based classification methods are presented in Table 1. These criteria include accuracy, precision, specificity, recall (sensitivity), confusion matrix (TP%, TN%, FP% and FN%), time complexity, F-score and error rate (e-training and e-validation), accounting for 97.4%, 20.5%, 7.7%, 23%, (38.5%, 35.9%, 38.5% and 38.5%), 23%, 71.8% and (76.9% and 51.3%) of studies, respectively. Most studies

focus on accuracy, indicating the highest percentage in the table. The second-highest percentage is presented by error rate, particularly in E-training. The best classifier model has low error rates. The third-highest percentage is obtained by time complexity. Table 1 shows that no study implemented all the evaluation matrices. Additionally, the variability in the percentage of studies that adopted these criteria demonstrates the difficulty in evaluating and benchmarking computer vision classification models using specific criteria.

C. ROBOTICS-BASED CATEGORY

Robots are the product of innovative technologies developed in various fields. Research in this field has a promising future. Robots can be categorised into intelligent assistant robots and social robots based on their appearance and main functions. This category has 47 articles ($n = 47/97$). Its subcategories include assistant-based ($n = 29/47$) and social-based ($n = 18/47$), which are explained below.

1) ASSISTANT-BASED

Assistant robots are the result of the creation of new and advanced technologies that are flexible and intelligible; these robots have considerable potential for educational applications [6]. A study [16] investigated the essential needs of an educational robot amongst six different user groups (pre-schoolers, primary schoolers, high-school students, college students, adults and the elderly) to propose future directions for the development of educational robots. From a large-scale perspective, the present research discusses some aspects of assistant robots to obtain clear and scientific knowledge, which is relevant in developing educational robots, such as learning-, motivation-, evaluation- and communication based educational robots. In addition, [62] reviewed the relevant literature and explored the use of various types of robots to facilitate the learning process for children of all ages. They highlighted the education–robotics symbiosis model, which can be implemented successfully at multiple levels of the educational system [62]. Assistant robots have performed various roles in varied contexts, including learning, motivation, evaluation and communication, as discussed further below sections.

a: LEARNING-BASED EDUCATIONAL ROBOT

In the era of technological advancement, the integration of education and robotics is essential for education, skill acquisition and training and development of the future workforce. Thus, the broad context of using various robots seeks to facilitate teaching and learning for students of all ages. The primary aim of this part is to emphasise the critical nature of incorporating robotic systems within the education sector. A literature review on robot research was presented and categorised its application in education [63]. Furthermore, a literature review and focus group interview with experts were conducted to incorporate teacher perspectives into the design of robots as primary school teaching assistants [64]. Another study [3] was presented to collect data on an assessment indicator and determine the educational effects of AI-based robot design using the revised Bloom's taxonomy. The study clarified a comprehensive overview of the robot design by evaluating the required competencies based on an AI via an examination of a six-dimensional questionnaire completed by 10 experts.

A study [65] demonstrated a robot that serves as a teacher's assistant and co-learns with the students. Teachers can obtain feedback from the robot regarding whether or not the students

are making progress. The objective of the researchers is to assist students in acquiring mathematical concepts according to their learning ability and performance [65]. In [66], the researcher developed a custom-built humanoid robot for educational and assistive purposes. The case study of [1] investigated and designed the learning autonomous programmable robot because of the needs of Chinese language teachers. A TA robot was developed to aid teachers in describing the rules of collaborative games and demonstrating the games, thereby engaging students successfully in collaborative activities [11]. Additionally, a tool for designing TA robots was proposed; this tool, based on the reuse of knowledge chunks (KCs), allows the creation and reuse of workflows via the scenario editor in PRactical INTEGENT aPplicationS (PRINTEPS) to clarify the robots' action workflow and HRI [67]. Two primary school teachers created a workflow for the TA robot by using the suggested technique, and then each of them cooperated with the robot to teach a science lesson [67]. In the concept of learning the STEM educational methods, a new educational robot called ROBOBO was designed to provide students with engaging and natural interactive experiences [15]. In [68], a survey of current scientific knowledge on the development of number cognition and mathematical skills in children and artificial cognitive systems was performed. This survey emphasised the tight association of numerical cognition with the human body, thereby motivating researchers from other disciplines to apply cognitive developmental robotics (CDR) approach for the computational modelling of such a key component of human intelligence [68].

b: MOTIVATION-BASED EDUCATIONAL ROBOT

Many articles focused on employing a humanoid robot for motivating and engaging students in the learning process. In [69], the researchers aimed to motivate children to learn math through the robot's guidance. This study [69] proposed a new technique to teach mathematics to primary school pupils by utilising social robot assistance and contemporary pedagogical learning theories. The study of [3] focused on integrating robots into traditional classrooms through a digital platform. The teaching robot verifies whether students' motivation and learning effectiveness are enough, which can be enhanced by teaching sustainable learning. Robots become the teachers' assistants and can detect students' emotions through the proposed digital system during their learning; thus, the robot can provide instant feedback regarding their emotions [5]. In [70], a motivational interview provided by a social robot was used to elicit qualitative data from participants, including their assessment of the robot's usability during the interaction and its effect on their motivation. In paper [72], the smart doll was implemented with a high level of engagement to the children interacting with it across all ages to demonstrate advanced vision capabilities that can lead to novel, engaging products. A comprehensive environment was offered to children (1–5 years old) and early education teachers; this environment includes a set of robotic

assistants designed as stuffed toys, a smartphone application, a knowledge base of educational activities and an expert system to capture children's attention naturally [72]. Robots enhance engaging learning experiences whilst providing instant and intuitive feedback [73]. Scott *et al.* investigated whether a basic course focused on programming personal robots in preparation for a Robot Olympics on the final day of the course could interest students in programming practice and enhance their code quality [73]. The robots' ability to encourage, motivate humans and improve their performance was investigated [19]. A reinforcement learning (RL) system was devised for TA robots in which Baxter robot acts as a learning facilitator whilst learning about human nature, motives and goals. Humans and robotics work together as feedback learners by exchanging behaviours, policies and expectations to achieve a particular mission [19]. A social robot was used to play games with children for a long period and motivate them during education. A user-modelling module with the adaption ability for the complexity of a game based on the child's ability level was built to provide children with the appropriate difficulty [74]. This module used a Bayesian rating system in determining the child's skill and the difficulty level of game items to personalise the game's progression [74].

c: EVALUATION-BASED EDUCATIONAL ROBOT

This subclass clarifies the abilities of an assistant robot, which include assessing students' emotional states automatically. A robot that can capture the children's facial features and analyse fatigue state based on multi-cue fusion was designed to improve their learning efficiency and relieve children's fatigue state during the learning process. Furthermore, the robot communicates with children through multiple channels, including voice, image and sound [75]. In [17], a TA robot that can detect students' raising of hands was developed. This TA robot can also check a student's program by analysing the completion of a programming task to assist them efficiently. A study [76] compared the interpersonal perceptions of a social robot evaluator and a human evaluator to evaluate the performance in the interview.

d: COMMUNICATION-BASED EDUCATIONAL ROBOT

Several studies were focused on developing an assistant robot for communication purposes. A graphical user interface for a socially interactive robot that enables the robot to execute various tasks, such as facial expressions, gestures and speech, was designed for face-to-face interaction with children [77]. The robot was equipped with a remotely controlled interface that allows it to exhibit various facial expressions and motions by controlling its face and body components to study its interaction capability with children and improve the connection between a child and his or her parent or therapist [77]. As of now, one of the major objectives of AI is to comprehend and communicate with people naturally. The social robotic head 'Furhat' was described to determine the possibility of utilising it as an aid for teachers in enhancing

communication skills in the acquisition of native language and second language with individuals who have various communication disabilities [78]. In addition, a work [79] provided a background on how to develop the brain and mind of a social and emotional robot by providing definitions, descriptions and methods and using cases. Humanoid FACE robot expressing emotions was designed and implemented when communicating with humans, with satisfying results in the reported use cases. A study [80] explored whether a robot may learn how to understand and communicate with one or more humans in a room using advanced machine learning algorithms. A robot was trained to read the level of the users' engagement at the start of the communication to achieve a completely autonomous robotic platform that can adapt itself to its users in many scenarios [80]. Moualla *et al.* proposed the emergence of Artificial Aesthetic in robots by allowing a mobile robot to learn artwork appreciation through neural network analysis. The robot acquires its artificial taste knowledge from the communication with people. The robot Berenson is a new type of visitor; the indications of its preferences are given by capturing its expressions in front of artworks in the museum [81].

2) SOCIAL-BASED ROBOT

This subcategory is focused on the potential of social robots to aid people and improve their quality of life in the future. Some of the widespread utilisation of the robotics system in society include therapy, guide and service sectors, which are explained in detail in the following sections.

a: THERAPY-BASED SOCIAL ROBOT

Some researchers focus on a social robot for therapy and diagnosis diseases, such as those for children with autism and the elderly. They aim to achieve results from diagnostic and therapeutic interventions that are faster and more meaningful than those achieved through conventional methods are.

Numerous approaches incorporating the latest advances in machine learning (ML) and deep learning (DL) have been developed to provide a personalised intelligent framework for the automatic detection of children's affective states and engagement during a robot-assisted autism therapy when diagnosing children with autism spectrum disorder (ASD) [82]–[85]. The first attempt to embed ML strategies into robot–children interactions aimed to assess the behaviours of children with ASD before introducing a metric to measure the therapy's effectiveness [84]. A study [86] presented a robot-assisted framework with an artificial reasoning module to aid therapists in ASD diagnosis. The proposed system automatically analyses the child's face and behaviours from a video to summarise the responses, which assist clinicians in reducing the time required to diagnose ASD [86]. A new approach was presented to engage children aged 1–5 years to participate in therapies with the aid of a robotic assistant [14]. A robotic aide provides tactile, aural and visual stimuli to children receiving speech-language therapy (SLT). The robot was developed to

stimulate children to engage in physical and rehabilitation activities and contribute to the children's attention span extension [14]. Furthermore, a robotic system was developed for children with ASD to predict participants' response to joint attention (RJA) performance during human–human interaction (HHI) using their head pose patterns during HRI in a semisupervised machine learning framework using transductive SVM (TSVM) techniques [87]. Interactive games between a social robot and a child were proposed to increase therapeutic effectiveness whilst minimising family costs [85]. The therapeutic objectives for children with autism spectrum condition (ASC) include improved social communication and interaction abilities, joint attention, response inhibition and cognitive flexibility [85].

The study [20] presented a framework for Socially assistive robotics (SARs) that employs a cognitive control approach to change their behaviour and actively plan the trajectory of interactions, thereby leading users towards learning or behavioural aims. The suggested system can adjust its behaviour by dynamically inferring the specific assistive services required to support a person in various circumstances based on his or her known health-related requirements [20].

A novel system named CATHI, which stands for cognitive assessment through human–robot interaction, was demonstrated in [88]. It combines a social robot and a cloud-based AI system to serve as a screening tool for mild cognitive impairment (MCI). MCI, a group of cognitive-related disorders that affect older adults, is often considered a prodromal stage of dementia [88]. Additionally, a new social robot, such as 'Pepper', was endowed with advanced cognitive abilities to provide a well-suited and effective impact in the healthcare assistance of the elderly within their home environments [20]. Moreover, an enhanced control architecture for assistive robots was synthesised using the constant flow of data provided by a sensor network [20]. Varrasi *et al.* contributed to the development of a robust automatic scoring system that can be implemented in a social robotic platform and used for cognitive impairment screening, early treatment and routine assessment for individualised care [21]. The robot assesses cognitive decline by providing orders to the patient, collecting their responses and objectively computing the final score to provide early therapy or routine evaluation for individualised care [21].

b: GUIDE-BASED SOCIAL ROBOT

Many researchers were focused on proposing robotic systems for guiding people in various scenarios. The research of Del Duchetto *et al.* presented an autonomous tour guide robot for use in a public museum to improve continually the robot's planning and activity based on human interactions. The proposed robot employs a unique regression model based on CNN and LSTM models to calculate a single scalar engagement value during HRIs by analysing conventional video streams from the perspective of an interacting robot [89]. Robust self-directed autonomous robots were designed to be evolving in public environments and

interacting with people. The developed framework is a 3D human recognition and tracking framework that possesses advanced capabilities in environment modelling and scene comprehension. It enabled the robots to make autonomous decisions, engage with humans in a short period and navigate safely in crowded public areas [90]. A robotic testbed incorporating a face detection and recognition hardware tracking system was proposed [8]. The robot has visual intelligence to recognise gestures and the capability of reading text, which can be applied in many real-world applications, such as video surveillance, biometric authentication and targeted advertising [8]. A shopping assistant robot was designed with a cognitive architecture that can learn and adjust its understanding of customers to aid individuals in real-life situations. Robots can influence customer purchase decisions and change shopping behaviour in smart technological stores by merging social robotics with machine learning technologies. [91].

c: SERVICE SECTOR-BASED SOCIAL ROBOT

Developing service robots is vital to conduct user-specified tasks (such as delivering, cleaning or monitoring) in indoor or outdoor conditions.

A cloud-based robotic system was proposed for Internet of Things (IoT) to deliver many reusable and shareable services for people's daily lives [18]. The robot utilises vision-based services, such as face and behaviour recognition, to locate a target individual in a home setting [18]. Moreover, Hu *et al.* implemented a monitoring system for NAO humanoid robot using the design of a fast background subtraction algorithm. The suggested approach tries to convert the background and foreground models into a problem of contiguous weighted linear regression [92]. The proposed monitoring system can monitor the living environment and serve as the foundation for various vision-based applications, such as fall detection. Nakamura *et al.* introduced the teahouse robot, which is based on the construction and applications of PRINTEPS, a collaborative environment for humans and robots. PRINTEPS is a user-centric platform that combines four types of modules, such as knowledge-based reasoning, vocal conversation, image sensing and motion management, to create integrated intelligent applications [93].

The studies in the robotics-based category have adopted the evaluation matrices to assess the performance of their AI classifiers. The performance effectiveness of the classification models was measured by measuring accuracy, confusion matrix, precision, recall (sensitivity), F-score and error rate (E-training and E-validation), which are illustrated in TABLE 2.

All the studies under the robotics-based class utilised the accuracy parameter as the main measure. The second-highest rate was obtained by time complexity, representing 57%. It is the classifier with the highest accuracy. The least time complexity is regarded as the best. The error rate has the third-highest percentage, particularly for E-training and E-validation, with 39.35% and 25%, respectively. The parameters of precision and recall have the same percentage,

TABLE 2. Measurement criteria used in the reviewed papers under the robotics-based category.

Ref	Reliability									Time complexity	Error Rate	
	Accuracy	Precision	Specificity	Recall	Confusion Matrix				Behaviour parameters of F-score		E-Training	E-validation
					TP%	TN%	FP%	FN%				
[65]	*											
[82]	*									*		
[5]	*											
[1]	*									*		
[86]	*									*		
[94]	*									*		
[89]	*	*		*						*	*	*
[67]	*											
[15]	*											
[72]	*											
[75]	*	*		*							*	
[90]	*										*	
[17]	*											
[79]	*									*		
[8]	*									*		
[83]	*									*	*	*
[20]	*									*		
[19]	*									*		
[80]	*									*	*	*
[18]	*									*	*	*
[84]	*	*		*	*	*	*	*	*	*	*	
[87]	*								*		*	*
[81]	*									*	*	*
[21]	*											
[92]	*	*		*	*	*	*	*	*	*	*	*
[74]	*									*		
[91]	*										*	
[85]	*											
Total	28	4	0	4	2	2	2	2	3	16	11	7
%	100%	14.3%	0%	14.3%	7%	7%	7%	7%	10.7%	57%	39.3%	25%

which is 14.3%. F-score and confusion matrix have the lowest interest, with 10.7% and 7%, respectively. Table 2 shows that no study includes all the evaluation matrices to evaluate the performance of their classification model.

D. AR-BASED CATEGORY

Access to a futuristic world built by a computer has become available because of the advancement of 3D technologies. The Smart Interconnected Interactive Classroom (SIIC) was introduced in [12]. It refers to the high-level design of a smart, modern, interactive laboratory class. It also refers to the ability of telecommunications, sensors and actuators to communicate with one another through a virtual world to improve the learning process and experience. A novel augmented and virtual service was proposed to help e-learning systems through VR and real-time interactions [12]. A review of several approaches, strategies and mixed reality device Microsoft HoloLens (MHL) for monitoring emotion recognition through facial expressions in AR was presented, followed

by a comparison of MHL and regular camera findings for emotion recognition [95].

This section highlights some aspects of the virtual environment and AR in the literature, i.e. (n = 6/97 papers). AR-based papers have includes the studies according to audience context for practice, training context with feedback and emotion embodiment.

Firstly, a VR audience environment was created and deployed at a university public speaking laboratory to allow undergraduate students to practise the speeches assigned as course assignments [96]. It covered system design, integration into a typical university course and the effect of practising in front of a virtual audience, particularly a student performance for a speech delivered in front of his or her class [96].

In the second study, frameworks for unsupervised physical exercise training in various applications have been developed to establish a training scenario and obtain feedback. For instance, a novel framework was built for the real-time detection, recognition and evaluation of the extended

TABLE 3. Measuring criteria adopted in the reviewed papers under the AR-based category.

Ref	Reliability								Time complexity	Error Rate		
	Accuracy	Precision	Specificity	Recall	Confusion Matrix					Behaviour of parameters	E-Training	E-validation
					TP%	TN%	FP%	FN%		F-score		
[99]	*									*		
[98]	*	*	*	*					*			
[96]	*								*	*		
[97]	*								*	*	*	*
Total	4	1	1	1	0	0	0	0	2	3	2	1
%	100%	25%	25%	25%					50%	75%	50%	25%

action-sequence motion capture data [97]. Furthermore, the proposed system was constructed to identify skeletal 3D joint positions received by depth sensors, such as Kinect [97].

Thirdly, an imitate mode of instruction using a ballet training application, ‘VR dance training with feedback protocols’, was demonstrated to enhance the traditional teacher-centred setup [98]. Automatic training activities for repetitive tasks were integrated with highly accurate analytical tools, such as automated gesture detection and assessment, to offer enough feedback and promote self-directed learning in an immersive environment [98].

Finally, engaging live storytelling was applied to the embodiment of emotion in the virtual environment. A new form of emotion embodiment creates an emotional experience and embodies the presenter’s emotional states by combining facial expression and auxiliary multimodalities and taking a two-dimensional (2D) approach for live storytelling in VR [99].

The studies in the AR-based category employed the evaluation criteria and sub-criteria to evaluate and benchmark their AI classification models, as presented in TABLE 3. The effectiveness of the performance for the classification models was mainly measured depending on the accuracy and time complexity, which have the highest percentage in the table. The second-highest rate was reported by F-score and e-training, which have 50%. The lowest rate was obtained by the confusion matrix and e-validation, which have 25%.

E. HYBRID-BASED CATEGORY

This category is the fusion of the capabilities of computer vision, robotics and AR approaches in implementing their systems. These approaches are combined to support the capabilities of each approach. From the literature, we found four ($n = 4/97$) papers for this category. The engagement and learning of students can be improved by experimenting with a digital theatre setting with a robot that can interact with them whilst practising drama-based activities in the digital theatre [100]. Besides the robot’s interaction

in the digital learning theatre, the interactive robot has interacted with the students as a classroom-learning peer through emotional expressions, hand gestures and speech characteristics [100]. Another study examined the problem of spontaneous dimensional emotion prediction to identify the affective states of children based on their spontaneous bodily expressions in child–robot communication scenarios [101]. The study explained the elicitation of emotions, data extracting, 3D skeleton modelling, feature designing and ML techniques [101]. A robotic system that can understand four gestures of human nonverbal communication was also developed [10]. The system recognises dynamic movements, such as waving and nodding, using a dynamic time warping (DTW) technique based on gesture-specific information derived from depth maps acquired with the Kinect v2 sensor. It can aid individuals with special needs, such as the elderly or people with disability [10]. In [102], an identity recognition method was developed to classify gesture operators in robotic sport instructor systems for rehabilitation and exercise training. The proposed sport instructor robot system adopted a hidden Markov model (HMM) with fine considerations of continuous-time gesture variations as the classification model for identifying 10 gesture operators. For effective identification, a new feature extraction scheme that considers the practical height of a person was created [102].

The performance of the classification systems utilised in the hybrid-based studies is measured by adopting the validation matrix, as illustrated in TABLE 4. All of the hybrid-based papers utilised accuracy and time complexity. The second-highest percentage is achieved by e-training at 66.7%. Finally, the F-score and e-validation have a percentage of 33.3%.

IV. DISCUSSION

This research aims to update a modern approach to automated assessment based on image-processing techniques. It also highlights the research efforts in this field. The four main aspects of the academic literature review in the current study area are discussed in the following subsections. These aspects include the datasets used in the articles in our survey,

TABLE 4. Measurement criteria used in the studies under the hybrid-based category.

Ref	Reliability								Time complexity	Error Rate		
	Accuracy	Precision	Specificity	Recall	Confusion Matrix					Behaviour of parameters	E-Training	E-validation
					TP%	TN%	FP%	FN%				
[101]	*									*	*	
[10]	*								*	*	*	*
[102]	*									*		
Total	3								1	3	2	1
%	100%								33.3%	100%	66.7%	33.3%

the motivations for utilising the automatic assessments systems or models based on vision, the existing challenges regarding our research area, the recommended solutions to solve or alleviate these difficulties.

A. DATASETS

In this section, we investigate the datasets used in our survey’s literature. In general, datasets are classified into two types: primary and secondary. The primary dataset refers to the data that researchers collect directly from the main sources during interviews, surveys or/and experiments [79]. The secondary dataset refers to the data that are collected by other researchers and readily accessible from their sources [79]. The descriptions for the datasets in our survey references are detailed in TABLE 5.

The primary type represents 26% (n = 31/121) of the datasets in the research in our survey. In particular, the primary-type datasets were distributed over the four categories of our research taxonomy: computer-vision-based class (n = 11/31), robotics-based class (n = 16/31), AR-based class (n = 1/31) and hybrid-based class (n = 1/31). In comparison, the secondary-type datasets represent 66% (n = 80 / 121), which are distributed over three of the four categories: computer-vision-based class (n = 72/80), robotics-based class (n = 6/80) and AR-Based class (n = 3/80). Furthermore, some papers include both types of datasets, and they account for 8% (n = 10/121) of the datasets in the four categories. The majority of the papers utilised secondary datasets, particularly in analysing the face-based behaviour of the computer-vision-based category. On the contrary, few researchers utilised the primary- and mixed-type datasets. The probable reason is that utilising the secondary dataset saves time, effort and cost in developing a model for a specific purpose. Some classifiers, e.g. deep CNN approaches, suffer from performance decrements when trained on a small image dataset [33]. Additionally, some secondary datasets involve images with variations in the form of changing illumination conditions, head pose variations, expression variations, sun-glass disguise and occlusions by hand [56].

B. MOTIVATION

This study illustrates the motivations of researchers to develop and implement automated vision-based assessment techniques from a varied point of view. Automatic assessment of human behaviour is essential for many concerns and challenging tasks. Technology is vital in supporting humans in many settings. This section presents the addressed motivations in the literature by organising them into six categories with related benefits for further discussion, as shown in FIGURE 3.

1) TA ROBOT

An intelligent robot based on AI and ML approaches is required to support students and teachers. Some learning materials are abstract and challenging for the teacher to interpret and illustrate. A robot that acts as a teacher assistant to co-learn with students and to provide them with a friendly environment helps students to overcome academic difficulties and improve their learning outcomes [5], [65], [68], [69], [75], [77]. Teachers obtain feedback on students’ progress whilst these students are learning [65], [69]. In Japan, the use of a humanoid robot developed to assist instructors, particularly in explaining and demonstrating games for students to learn the rules, is an innovative approach to game-based learning activity [11]. A robot was built in [17]. This robot assists students in correcting and evaluating the performance of their programs during lessons and detects students who raise their hands.

In general, technology is vital in supporting humans in many educational settings. One example of these settings is the automated assessment of student performance in presentation sessions [23].

2) COST AND USER-FRIENDLY ROBOTICS SYSTEM

The development of affordable Information and Communications Technology (ICT)-based educational tools for students and instructors in low- and middle- income schools is necessary to address issues in early childhood development [72]. A TA robot design tool was proposed depending on KC

TABLE 5. Datasets used in the reviewed research.

Usage in the previous study	Name of Dataset	Type	Description
[65], [1], [72], [14], [90]	Knowledge base	Primary	The knowledge base of input variables is stored as data structure and defined by the researchers to model considerable static/dynamic knowledge about the environment and to propose an intelligent system.
[67]	Knowledge chunk KC base	Primary	The database stores the KCs, which are components of the reusable workflow, the indexes necessary for searching and the video recording of robots moving in accordance with the workflow components.
[58]	Semantic rules	Primary	The researchers developed semantic principles for recognising human actions that are not dependent on visual analysis. Semantic rules define an action as a relationship between its constituents, including bodily parts, associated objects and/or the scene.
[66]	ROS Moveit database	Secondary	It utilised the ROS Moveit package for trajectory control and gripping tasks.
[66]	Unknown	Primary	The dataset for the teaching assistant test was created from 10 students in three different groups (preprimary, primary and secondary schools).
[82]	EMNIST digit dataset	Secondary	The EMNIST dataset contains balanced handwritten digit datasets with 280,000 samples of handwritten digits (from 28,000 samples per digit) and 145,000 samples of handwritten characters (both uppercase and lowercase alphabets).
[82]	Custom colour ball dataset	Primary	Colour ball datasets are composed of approximately 500 photos that have been gathered and labelled.
[27], [94], [75], [9], [84], [31], [45], [46], [33], [41], [42], [43], [50], [38], [39], [53]	Extended Cohn-Kanade (CK+)	Secondary	It is a database of facial expressions that contains 327 annotated video sequences from 123 participants in eight different expression states. The dataset's subjects are diverse in age, gender and ethnic origin, making it one of the most preferred datasets for FER research. The image has a resolution of 640×490 pixels. The researcher in [27] concentrated on five fundamental emotional states of CK+: happiness, sadness, surprise, anger and neutral.
[28], [99], [33], [42], [44], [35], [39], [52], [53], [54], [37]	Fer2013 Dataset	Secondary	The Fer2013 dataset features an extensive quantity of data, high-quality images and labelled data. It comprises 35,887 greyscale images of the human face, each with a size of 48×48 pixels and a facial area that is roughly centred. Each image has a label that ranges from 0 to 6 and represents anger, disgust, fear, happiness, sadness, surprise and neutrality. The training dataset has 28,709 images. The public test set and the resultant test set both have 3,589 images.
[5]	Google facial expression comparison dataset	Secondary	This dataset contains 500K triplets and 156K face photos. It is a large-scale facial expression dataset with a size of 200 MB.
[86]	Dlib face detection dataset	Secondary	It consists of over 3 million faces from various datasets, including the Face Scrub and Visual Geometry Group (VGG) datasets.
[22]	A new e-learning and classroom environment database	Primary	The dataset contains over 4000 manually labelled image frames with object localisation for facial expressions, hand gestures and body postures of Indian students. It includes posed (acting) and spontaneous (natural) expressions and single-person and multiperson images in a single image frame.
[89]	TOur GUide RObot (TOGURO) dataset	Primary	A long-term dataset acquired by an autonomous TOGURO was placed in a public museum with three independent coders continuously annotating a numeric engagement assessment of visitors.
[89]	User Engagement (UE-HRI)	Secondary	This publicly available HRI dataset contains 54 spontaneous interactions between a robot and humans. The interactions range from 4 to 15 minutes.
[99]	Closed Eyes in the Wild (CEW) dataset	Secondary	This dataset was adopted to detect and classify different eye blinking states.
[23]	Student behaviour dataset	Primary	This dataset is based on volunteer participants who presented a mock presentation lasting up to 90 s. The presentation is a video record of a group of participants. Their details were removed before the video files were analysed.
[28], [86], [103], [29], [36], [32], [96], [79], [8], [18], [84], [87], [101], [25], [102]	Real-time recognition dataset	Primary	The sensors, e.g. camera or Kinect V2 sensors, collect different human behavioural features or participants' facial expressions in real time during an experiment. Then, the researchers should annotate the collected video and label the data according to the research subjects to produce their own datasets.
[30]	CASIA-WebFace database and FaceScrub database	Secondary	These two large labelled face recognition datasets can be used for training a face recognition model on over 500K images by combining them.
[90]	NRIA Person dataset	Secondary	It is a dataset for the detection of a person in images and video.
[78]	Speech data	Primary	It comprises a collection of roughly 1500 Slovak and 1800 English read utterances. It is used to develop an identical synthetic voice that can speak both Slovak and English by utilising a bilingual male speaker who is both Slovak and English.
[78]	Fisher English database	Secondary	It is an English Training Speech database, which comprises time-aligned transcript data for 5850 complete conversations lasting up to 10 min and represents a collection of conversational telephone speech (CTS).
[9]	CK+ (7 expressions)	Secondary	It is the same as the normal CK+ that includes six facial expressions but with the addition of the natural expression.
[9], [41]	Radboud Faces Database (RaFD), (7 and 8 expressions)	Secondary	The dataset is appropriate for FER analysis. It contains photos of 67 people making eight various facial expressions (anger, disgust, fear, happy, contemptuous, sadness, surprise and neutral) with three different gaze directions and five different face orientations. Selecting frontal facial images with frontal gaze direction yields two subsets: the first

TABLE 5. (Continued.) Datasets used in the reviewed research.

			contains 469 images with seven expressions (anger, contemptuous, disgusted, fearful, happy, sad and surprised), whereas the second contains 536 images with 67 instances of the neutral expression added to the previous ones.
[83]	Multiset discriminant correlation analysis (MDCA)	Primary	It is a multimodal dataset of 35 children with autism. It was gathered using sensory inputs, such as cameras and microphones, to capture distinct modalities (facial, body and voice) of the children's behaviours.
[20]	Knowledge-based cOntinuous Loop (KOaLa)	Mixed	KOaLa is a realistic scenario-based architecture that incorporates sensor data representation, knowledge reasoning and decision-making capabilities.
[80]	Unknown	Primary	It is a dataset of images gathered through an interactive data collection based on the interaction between the robot Pepper and the participants using the 2D camera on the robot's forehead to record roughly 20 s of video at 5 fps. The image depicts the minutes shortly before the start of the interaction to teach the robot how to begin engaging with a person in both single-user and multiuser scenarios.
[18]	Face Detect_Pose Estimate Online dataset	Secondary	It has 90 image sets of various individuals. Each image set contains facial images of a particular subject taken from various angles, ranging from 90 to -90 degrees with a 5-degree step, with the human face oriented to the camera as 0.
[21]	Unknown	Primary	The data were acquired from 16 study participants using the microphones of the Pepper robot for audio recording, and the robot captured the images of the user drawings for the visuospatial and executive skills, which were then analysed.
[48], [45], [33], [24], [50], [38], [53], [54]	The Japanese Female Facial Expression (JAFFE) Dataset	Secondary	It features 213 facial photos of 10 Japanese women, each with seven different facial expressions: anger, disgust, fear, happiness, neutral, sadness and surprise. The image has a resolution of 256 × 256 pixels. Although it is a small dataset, it is considered to be quite challenging.
[26]	The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)	Secondary	The RAVDESS dataset consists of 24 professional actors (12 female and 12 male). Each delivers two lexically identical utterances in a neutral North American accent. The researcher chose the video songs without audio from the RAVDESS dataset [26] to focus on the distinctions of video frames with six different classifications of emotions.
[10]	Human Pose Recovery and Behavior Analysis (HuPBA) sequences dataset	Primary	A small dataset of dynamic gestures using specific features from the face and the body and static gestures was obtained from the depth images of the Kinect TM sensor by using the Kinect TM SDK v2.0. This dataset includes 30 sequences of six users (5 sequences per user). Each user randomly performs the four gestures (i.e. pointing at, waving, nodding and head negation).
[92]	I2R dataset	Secondary	The dataset consists of nine demanding videos with some frames, including ground truth. The videos feature the following elements: bootstrap (120 × 160 × 3057 frames; crowd scene), campus (128 × 160 × 1439 frames; waving trees), curtain (128 × 160 × 2964 frames; waving curtain), escalator (130 × 160 × 3417 frames; moving escalator), fountain (128 × 160 × 523 frames; fountain water), hall (144 × 176 × 3548 frames; crowd scene), lobby (128 × 160 × 1546 frames; switching light), shopping mall (256 × 320 × 1286 frames; crowd scene) and water surface (128 × 160 × 633 frames; water surface).
[93]	Mongodb_store - ROS	Primary	It is a MongoDB-based storage and analysis system for the data from the ROS system, which serves as the Information State database in PRINTEPS for data sharing amongst PRINTEPS's five subsystems.
[45]	Multimedia Understanding Group (MUG) Dataset	Secondary	It is an extensive database with 1462 sequences and 86 subjects. Each sequence includes 50–60 frames. The resolution of the image is 896 × 896 pixels.
[45]	Static Facial Expressions in the Wild (SFEW) Dataset	Secondary	It features photos acquired from movies with various illuminations and positions. The images are 720 × 576 pixels in size.
[45], [42]	Oulu-Chinese Academy of Science, Institute of Automation (Oulu-CASIA) datasets.	Secondary	It contains 480 image sequences taken from 80 individuals. Each series has five to six photos of peak expression. The first frame is assigned to the category of neutral emotions. Each category of emotion has 500 photos.
[46], [41], [38]	MMI dataset	Secondary	It is a database of facial expressions. It features 30 people of both sexes, whose ages range from 19 to 62 years. They represent various ethnic groups (European, Asian and South American). This dataset has 213 sequences annotated with six fundamental expressions, of which 205 are from the frontal view. Each sequence depicts a single facial expression, beginning with a neutral face (onset), progressing to a peak and concluding with an offset.
[46]	Acted Facial Expression in Wild (AFEW) dataset	Secondary	It is a compilation of short videos culled from videos shot in real-world settings. It is one of the most difficult datasets to analyse because it contains facial expressions from real-world situations. It has 1426 video sequences and 330 subjects of both sexes, whose ages range from 1 to 70 years. It was separated into three sets: training, validation and testing sets. Each set contains seven different facial expressions (angry, disgust, fear, happy, neutral, sad and surprise).
[55]	Dataset [I], Dataset [II], Dataset [III].	Primary	Three distinct datasets have been developed. Each dataset offers various face orientations and circumstances constrained to 181 × 181 pixels. No image blending was done to dataset [I]. However, a linear blending of 1.0 alpha (α) was applied to dataset [II], and a linear blending of 0.5 alpha (α) was applied to dataset [III].
[56]	CurtinFaces dataset	Secondary	It is the Microsoft Kinect face dataset, which contains over 5000 RGB-D photos of 52 participants obtained with the Kinect sensor. The photos of each subject were taken in various poses, lighting conditions and facial expressions, with and without sunglasses disguise. Each subject has at least 49 images at 7P × 7E and 35 images at 5I × 7E.
[56]	Biwi Kinect Head Pose dataset	Secondary	The dataset was created to estimate the head position. This dataset contains around 15,000 RGB-D photos of 20 individuals, including 6 females and 14 males. The head poses of the images vary, for example, along the yaw axis in the range from -75 to +75 and along the

TABLE 5. (Continued.) Datasets used in the reviewed research.

			pitch axis in the range from -60 to +60. Additionally, this dataset includes images that were partially obscured by a hand or a sunglass to increase the difficulty of the experiments.
[56]	University of Western Australia (UWA) Kinect Database	Secondary	It was obtained at the University of Western Australia in an indoor laboratory environment. It comprises almost 15,000 low-resolution RGB-D photos obtained with the Kinect sensor of 48 people. Each subject's image has a unique facial expression and a broad head posture rotation. The dataset contains between 289 and 500 images for each participant.
[56]	Honda/UCSD dataset	Secondary	It is a 2D image collection comprised of 59 video sequences (image sets) featuring 20 individuals. The image sets range in length from 12 to 645 images.
[56]	CMU Motion of Body (Mobo) dataset	Secondary	It was created to assist in determining the position of a human body. It has 96 sequences of 25 individuals walking on a treadmill in CMU 3D room. Numerous cameras captured four distinct walking styles: slow, quick, inclined and holding a ball. Each individual is represented by four video sequences, each depicting a distinct walking pattern.
[47]	Bank of Standardized stimuli (BOSS) dataset	Secondary	It is a new 2D dataset containing images of faces and non faces. The majority of face images were taken in uncontrolled environments and conditions, including variations in illumination, facial expressions (neutral, anger, scream, sad, sleepy, surprised, wink, smile, smile with teeth and open/closed eyes), head pose, contrast, sharpness and occlusion. The majority of individuals are between the ages of 18 and 20. Some older persons with a distinctive appearance and hairstyle and those with a scarf are also included. The images were captured with a 26-megapixel ZOOM CMOS digital camera with full HD capabilities. This database has 9619 images, containing 2431 training images (771 faces and 1660 nonfaces) and 7188 test images (178 faces and 7010 nonfaces).
[47], [42]	MIT Database	Secondary	It is a dataset of face and nonface images, which were used at MIT's Center for Biological and Computational Learning. This dataset has 6977 training photos (2429 face images and 4548 nonface images) and 24045 test images (472 face images and 23573 nonface images). The face images were saved in PGM format, with a resolution of 19*19 pixels and a greyscale of 256 levels.
[33], [39], [40]	Karolinska Directed Emotional Faces (KDEF) Database	Secondary	It has 4900 images of facial expressions. This dataset contains seven distinct facial expressions: angry, happy, neutral, surprised, sad, fearful and disgusted.
[33]	AffectNet Database	Secondary	It is a web-based dataset composed of around one million images of naturally occurring facial expressions gathered from the Internet via three search engines and 1250 emotive terms. It has 11 labels, including neutral, surprised, happy, sad, fearful, disgusted and angry for emotions and nonemotions, as well as the classes of contemptuous, none, uncertain and no-face.
[59]	Montalbano gesture dataset	Secondary	It features 13206 Italian gestures divided into 20 distinct categories (6862 for teaching, 2765 for confirmation and 3579 for testing). All movements were captured at 20 fps using a Kinect 360 sensor, which provides multimodal information, such as RGB, depth, skeleton and user mask. Each segmented video has approximately 30 frames.
[59]	Gesture Commands for Robot InTeraction (GRIT) gesture dataset	Secondary	It contains 543 gestures separated into nine categories. The frame sequences were captured with an RGB camera at 20 fps. Each sequence has approximately 30 frames. Each video features a single dynamic gesture. Additionally, the gestures are immediately discernible because it was built for usage with robots.
[59]	ChaLearn LAP Isolated Gesture Dataset (IsoGD)	Secondary	It is a vast database with 47,933 gestures classified into 249 distinct categories. This dataset was captured at 20 fps using a Kinect 360 camera. It comprises RGB and depth images. Each video shot has an average of 47 images and depicts only one gesture. This dataset can take several forms, including body language gestures, Indian mudras, spoken gesticulations, drawings, symbols, signs, signals and pantomimes.
[97]	Cardiovascular disease (CVD) exercise dataset	Primary	It is a new dataset that contains activities selected expressly to aid people with CVD in maintaining a minimum level of physical activity and improving their health. Each subset of this dataset has 15 exercises performed twice or thrice by 15 people, 4 females and 11 males, resulting in 656 sequences and 28979 frames. A single Kinect sensor positioned in front of the user was utilised to record data at 30 fps.
[97]	The Microsoft Research Cambridge-12 (MSRC-12) dataset	Secondary	It is a big gesture database for action recognition. It contains 594 unsegmented sequences for 30 subjects executing 12 gestures. The Kinect sensor gathered data at 30 fps. It has 6244 gesture instances and 719359 frames. Along with the 3D coordinates of 20 body joints, annotation files indicating the action points at which action detection should be triggered are provided. Examples of gestures include beat bow, change weapon, duck, goggle, had enough, kick, lift outstretched arms, push right, shoot, throw and wind it up.
[97]	Microsoft Research Cambridge-12 Kinect gesture data set (MSRC-12)	Secondary	It is a publicly available depth map and skeleton sequence dataset for action recognition consisting of 567 presegmented sequences of 20 actions performed twice or thrice by 10 people. The data were collected using a depth camera and infrared light at 15 fps. It contains 23,797 sample frames, each with 20 3D joint positions. The following are the actions that it covers: bend, draw circle, draw tick, draw x, forward kick, forward punch, golf swing, hammer, hand catch, hand clap, hand wave, high arm wave, high throw, horizontal arm wave, jogging, pick up and throw, side kick, tennis serve, tennis swing and two-side boxing.
[24]	The Hollywood2 dataset	Secondary	It has 1707 videos describing 12 different actions, including 823 videos used for the training and the remaining 884 videos used to measure performance. The video was collected from 69 different Hollywood films. Training and test videos were culled from various films.
[24]	The Olympic sports dataset	Secondary	It has 783 videos depicting athletes engaged in 16 different sports. It includes 649 videos used for training and the remaining 134 videos for measuring performance. Each video focuses on a single activity.

TABLE 5. (Continued.) Datasets used in the reviewed research.

[24]	The Hollywood 3D dataset	Secondary	It consists of 951 video pairs, i.e. left and right channels, describing 13 actions and a 'no-action' class from Hollywood films. These video pairs were divided into 643 videos for training and 308 videos for performance measurement. Training and testing videos were culled from various films.
[24]	ORL dataset	Secondary	It has 400 photographs of 40 persons' faces. The photographs were taken at different times under various conditions, including lighting, facial expressions (smiling / not smiling) and facial details (open/closed eyes, with/without glasses). Facial images were recorded in a frontal stance with a tolerance of up to 20 degrees for facial rotation and tilting.
[24]	The University of California Irvine (UCI) Machine Learning Repository Datasets	Secondary	This group of datasets include Glass, Heart, Iris, Parkinson, Relax, Seeds, Spectf and Tae datasets, which have been used to apply multiclass classification.
[60]	Data S1	Primary	It is a gesture dataset with a complicated background and apparent shifts in illumination. It has 3269 images divided into 16 gestures and 26 complex scenes.
[43]	The Denver Intensity of Spontaneous Facial Action (DISFA) database	Secondary	It includes 27 facial videos, each of which is dedicated to a distinct adult subject. Each subject was videotaped at 20 fps with a high-resolution camera (1024 × 768) whilst viewing a 4 min stimulus video clip meant to induce spontaneous facial expressions. Each video frame's intensity of 12 facial AUs was coded manually in six levels from 0 to 5. Additionally, this collection contains the X–Y coordinates of 66 points of facial landmarks for each video frame.
[34]	Audio-Visual Emotion Challenge 2013 and 2014 (AVEC2013 and AVEC2014) depression subchallenge datasets	Secondary	These datasets are two of the few that contain raw video and audio data. Both datasets are subsets of the audio-visual depressive language corpus, consisting of 150 films from 82 participants. A webcam and a microphone were used to capture an individual interacting with a computer during a human–computer interaction assignment. Each dataset was partitioned into three unique groups: training, development and test. Each set contains 50 movies labelled with the subject's depression degree. Each video lasts between 20 and 50 min. On average, the depression levels are 15.1 and 14.8 for training and development sets, respectively.
[35]	IMDB WIKI dataset	Secondary	It is the largest dataset (over 500K), including high-quality images of male and female faces with age and gender labels from various nations.
[39]	FEEDTUM dataset	Secondary	It comprises short videos of 18 participants reacting to stimuli to elicit an emotional response. In this dataset, every sequence corpus begins and finalises with a neutral face.
[39]	Multi-PIE dataset	Secondary	It features 750,000 images captured from 337 individuals. The facial expression images were recorded in various postures, with each image containing 19 degrees of illumination. The images of pose variants were obtained from a front-facing camera at angles ranging from 0 to ±90 degrees. The multi-PIE corpus includes unlabelled images of emotions.
[49]	Caltech Occluded Faces in the Wild (COFW) dataset	Secondary	It is one of the most challenging datasets to produce because it was designed to show faces in real-life situations with partial occlusions. It comprises 1852 face images, 507 of which are for the testing set. The occlusion of the face images in this dataset varies greatly because of many accessories on the face, such as sunglasses and hats. An average occlusion of more than 23% was obtained.
[49]	The Wild (300W) dataset	Secondary	It is a compilation of many facial recognition datasets, including the LFPW, HELEN, AFW and IBUG. It includes 300 indoor and outdoor face images in the wild. Each image was labelled with 68 face landmarks.

reuse to reduce development costs by using a user-friendly scenario editor for end users known as PRACTICAL INTELLIGENT APPLICATIONS (PRINTEPS) [67]. PRINTEPS platform assists end users to participate in the design of AI applications and develop applications easily [93]. Moreover, a low-cost and user-friendly robotic system was designed; this system can perform many activities in a developing country community with limited technology literacy [66]. Compared with traditional techniques, automated behaviour detection, diagnosis and assessment systems are inexpensive [86]. For example, a low-cost and user-friendly mental health care system was developed with efficient face emotion analysis systems [33]. The other example is the automatic depression diagnosing system [34].

3) ONLINE EDUCATION

The quality of teachers' online education content, the role of online education and the influencing factors of online teaching are fundamental in distance education. Automated

detection of learners' learning emotions in real time offers many benefits for online education, such as enhanced interaction between the lecturer and learner and personalised education [28], [29], [32]. The difference in space and time between students and instructors generates a loss of emotion in the learning process, which has consequences on students' learning and teachers' teaching quality [37]. Students' affective state analysis can be used to personalise a smart tutoring system and an intelligent class setting. It is also based on DL or ML methods. Therefore, the accuracy of the adopted database is crucial for the effective recognition of affective states [22]. In the real world, automated classroom attendance evaluation was demonstrated utilising face recognition and ML techniques [30], [55].

4) SUPPORT SPECIAL NEEDS

Numerous studies were conducted to develop SARs capable of adapting users' long-term behaviour and to assist in behavioural, therapeutic or educational purposes.

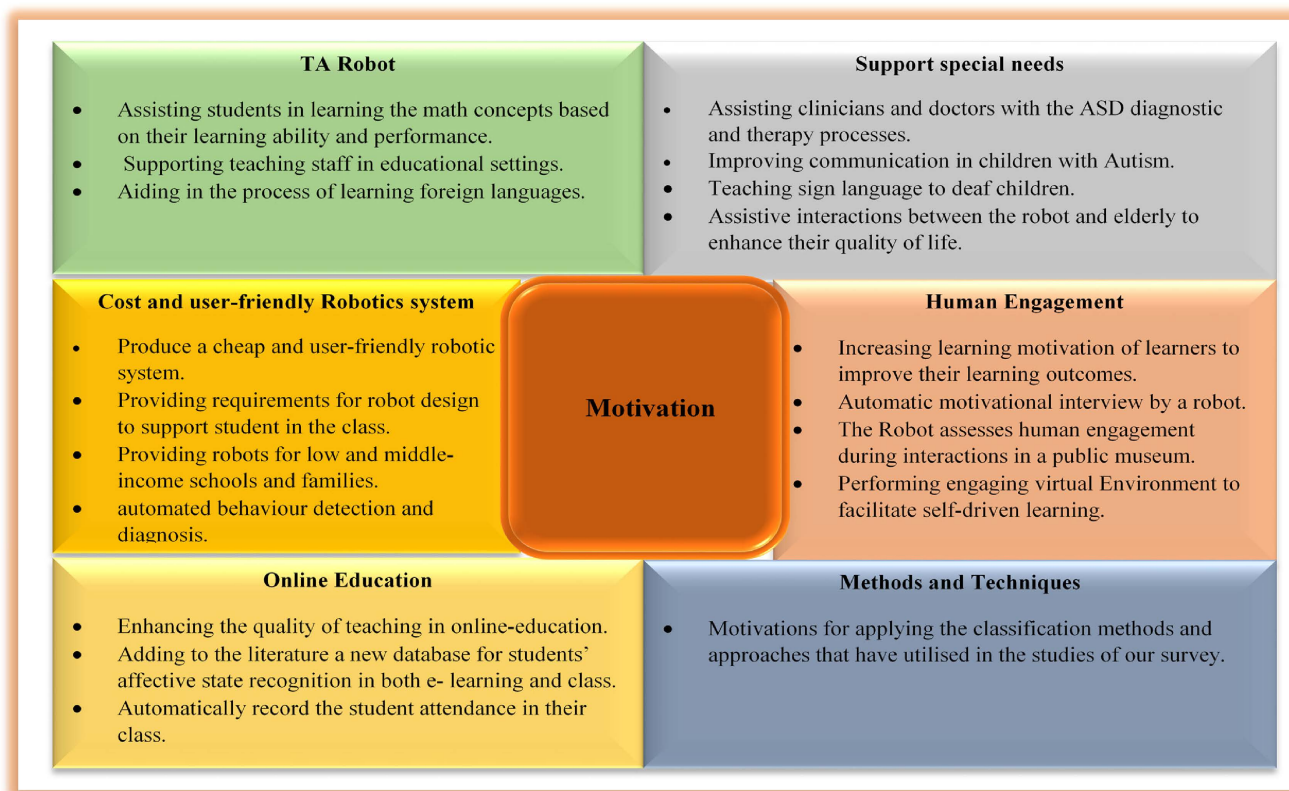


FIGURE 3. Categories of motivations in automated vision-based assessment.

For example, a robot interacts with an elderly to monitor, assist and guide him or her towards learning or behaviour goals inside the domestic home environment [2], [10], [20], [21]. In Brazil, a robot-assisted framework with an artificial reasoning module was developed to aid physicians in diagnosing ASD and reduce diagnostic delay [86]. At Sheffield Hallam University, an intelligent robot was programmed to provide preliminary remote diagnostic information about patients, thereby reducing the workload of doctors and expanding the population eligible for screening of neurological disorders amongst the elderly (e.g. dementia) and younger population (schizophrenia and depression) [88]. Additionally, a robot was developed to teach hearing impaired children Persian sign language by recognising human emotional states from frontal facial photos, i.e. looking into the user's face and imitating the user's emotional states [27]. A robot's head was designed to aid teachers in developing communication skills in children with ASD [85] and in second language learning and native language communication with adults who have various communication disabilities [78].

5) HUMAN ENGAGEMENT

The paradigm of learning engagement has transformed from manual engagement to personal response systems or mixed reality on portable devices platforms. As attractive agents

for humans, robots were used to fulfil different needs and were brought into the realm of second language acquisition (SLA). A robot was implemented to gain students' attention during learning [15], [72], [75], [94] and motivate children with ASD to participate in the therapy [14], [83], [84], [87]. Robots make learning interactive by facilitating engaging learning experiences and offering rapid and intuitive feedback [19], [73]. In China, a robotic Chinese language tutor was developed to demonstrate the Chinese words in an engaging manner, increase the understanding of learners and enhance their memory of the learned words [1]. Social robotic motivational interviewers were also developed to motivate participants during interaction and increase their physical activity [70]. A robot was trained to analyse people's degree of engagement when they initialise to communicate with it [80], [89]. However, AR and VR assist in encouraging self-directed learning by implementing an engaging virtual environment [96], [98], [100]. In Taiwan, a robot in a digital theatre environment was proposed to investigate the impact of the robot engaging with students whilst conducting drama-based activities in the digital theatre and enhance student engagement and learning [100]. The speech practice in front of a virtual audience positively affects students' performance when they deliver a speech in front of their classmates [96].

TABLE 6. Motivations of the reviewed studies for utilising the methodologies and approaches.

No.	Methods	Motivation
1	SVM	It is the most effective classifier utilised in ML because of its high classification accuracy [32] and convenience in classifying behaviour recognition problems [18] and human action recognition in unconstrained environments [24].
2	HOG	It is a powerful technique that can be effectively exploited for the FER problem [9] [18].
3	HOG- SVM algorithm	FER engine combines HOG descriptor with SVM to enhance the classification performance of traditional HOG in real-time recognition [84].
4	K-means	This method is adopted to classify the head posture vectors from each HRI session into a hard histogram [87]. It assigns data points by computing the centroids and iterating until it finds the optimal centroid.
5	TSVM	It is a popular algorithm that applies low-density separation to improve classification performance [84], takes advantage of both labelled and unlabelled data points and tests the effectiveness of the features generated [84].
6	Online Recursive Gaussian Process (GP)	It is a nonparametric kernel-based technique that performs successfully in time series prediction issues; the recursive kernel is utilised to form the time dynamics [101].
7	Uniform LBP-KNN	It combines the uniform LBP and KNN algorithm and uploads the binary code to the FPGA chip, utilised for saving computational time and memory [25].
8	IBM AI Cloud services ('Watson')	It provides inclusive user-friendly tools for speech recognition (speech-to-text) and production (text-to-speech) and object recognition from images [21].
9	Discriminated ELM classifier	It selects the active neurons to create a simple network topology and uses few neurons whilst preserving high recognition accuracy as in the traditional ELM [48].
10	Semantics-based approach	It utilises semantic rules instead of ML methods to detect human behaviours without training systems; the recognition task is more reliable, particularly
11	CNN	It is a potent classification model extracts features using convolution layers with an accuracy rate higher than that registered by humans [26] [51]. CNN can discover deep feature representations of facial expressions to achieve automatic recognition [50]. It is useful in face recognition based on a feature-based approach [35] and powerful for facial emotion recognition [52][54]. It can also utilise unconstrained environments concerning facial pose and illumination invariance [39]. Moreover, it can learn a nonlinear relationship between feature vectors [53].
12	DTW	The DTW technique detects certain geometric features computed from depth maps and dynamic gestures, matches two temporal sequences and determines the low-cost alignment between them [10].
13	HMM	An HMM is adaptable in nature; it can achieve excellent identity recognition accuracy with enhanced feature extraction-based feature parameters [102].
14	Bayesian rating method	It utilises a limited quantity of data to establish item ratings; it can also measure the human players' ability and performance in a game with a robot [74].
15	MDP + PVK-GDA + PVK-ELM	It combines MDP with PVK to control the time spent on selecting the optimal kernel and dimensional reduction [45]. The PVK-based ELM performs high recognition accuracy because the function of a single kernel PVK is sufficient for all datasets in different categories. It is computationally robust and generates stable results compared with typical ELMs [45].
16	HOG-CNN	Two classifiers were combined to overcome the possible FNs and to complete each another because the HOG detector was first applied to the testing image; if the approach fails, then the CNN detector would be applied to the image [30].
17	HOG-BPNN	These two methods were combined to increase the robustness and effectiveness of the present face classification research findings [47].
18	You Only Look Once (YOLO)	It is one of the CNN networks; it is fast and accurate in the detection and recognition of objects and hand gestures in real time [60].

TABLE 6. (Continued.) Motivations of the reviewed studies for utilising the methodologies and approaches.

19	Deep Belief Network (DBN)	It is a semisupervised classifier effective for predicting the most important facial features with few computations; compared with human learning, it can use labelled and unlabelled data [41].
20	CNN-SVM	Two machine learning algorithms were integrated to improve performance in facial image classification [44].
21	Multi-scale Spatiotemporal Network (MSN)	It utilises 3D convolutional layers with various kernel sizes to extract detailed spatiotemporal information about the changes in facial expressions; the tool can record facial expressions and dynamics over a wide range of values. [34].

6) METHODS AND TECHNIQUES

The motivations for utilising the methods and approaches applied in the studies from our survey are discussed in this section. Some of the researchers in our survey were motivated to adopt specific classification systems and methods for many reasons. One of these reasons is that proposing a new method by combining image-processing techniques improves the accuracy of the classifier, thereby enhancing the system performance [55]. Moreover, researchers were motivated to propose a method for a robust recognition system for images with various conditions [56]. TABLE 6 clarifies some of the motivations for utilising different classification systems and methods.

The majority of the papers in our survey expect to reach the highest level of recognition accuracy and the most excellent result for the classification models. The researchers developed novel classification methods or enhanced the current systems to improve the accuracy and efficiency of classification models. The classification models with high accuracy and performance were demonstrated in the findings of the majority of reviewed papers using the proposed methodologies. Thus, this research field offers various effective approaches for the automatic vision-based detection and assessment of human behaviour for many purposes and interests. On the contrary, specific challenges from the reviewed studies are discussed in the next section.

C. CHALLENGES

Many fields, such as AI, robotics and AR, concerned with the automated vision-based assessment of human behaviour involve many challenges. The literature outlines several challenges in many important respects. This study is limited to summarising and discussing only the crucial challenges. Figure 4 illustrates the key points of the main issues, which are illustrated in detail in the following subsections.

1) CHALLENGES RELATED TO CO-LEARNING OF ROBOT AND STUDENT/TEACHER

Many researchers focus on the effectiveness of robotics in education. In this era, co-learning between humans and machines is a critical problem for the global communities. Learning about the use of educational robots, the development of appropriate teaching materials and curriculum and

the confirmation of quality dimensions in robotic education from the perspectives of instructors is necessary [6], [16], [62], [64]. The application of robots in education is still investigated, and related technologies must be validated and enhanced [63]. It can reduce teachers' workloads whilst increasing students' enthusiasm and learning [63]. Machine learning is one way of offering intelligent learning capabilities to a robot. The ML techniques investigate the algorithms for acquiring knowledge from data. For example, the structure of an intelligent agent with fuzzy mark-up language, knowledge base, rule base and optimisation model was proposed [65].

Another issue is the robot–student co-learning for the second language teaching [1]. RALL (robot-assisted language learning) involves several limitations and possibilities for teaching communication skills in both native and second languages [78]. The shortage of Chinese language teachers results in the need to design an educational robot that utilises its automated emotion recognition and body language detection to demonstrate the Chinese words [1]. The majority of studies focus on teaching restricted vocabulary in a second language (L2) and the usage of NAO robots [78]. However, Benus *et al.* investigated whether the social robotic head (Furhat) could aid human teachers in developing communication skills in SLA and native language with people with diverse communication disabilities.

For robot–teacher collaborative teaching, including a robot as a teaching assistant in a class is a challenging task, particularly in teaching numeracy to children [68], [69]. The effectiveness of the robot in education was assumed to increase student participation, strengthen the comprehension of mathematical concepts and analytical and critical reasoning and develop children's cognitive skills [68], [69]. In the physical classroom, no efficient technique exists to improve class motivation and learning effectiveness. Most students may believe that the class is tedious or fail to understand the goal of learning if the teaching methods are not engaging, resulting in low classroom participation and poor learning effectiveness [5]. Moreover, computer programming is a challenging aspect of learning. Student learning can be improved by regular practice in application programming. However, teachers often find that first-year bachelor's students do not engage in such activities [73]. Therefore, a system that can increase student performance and teacher teaching skills is necessary.

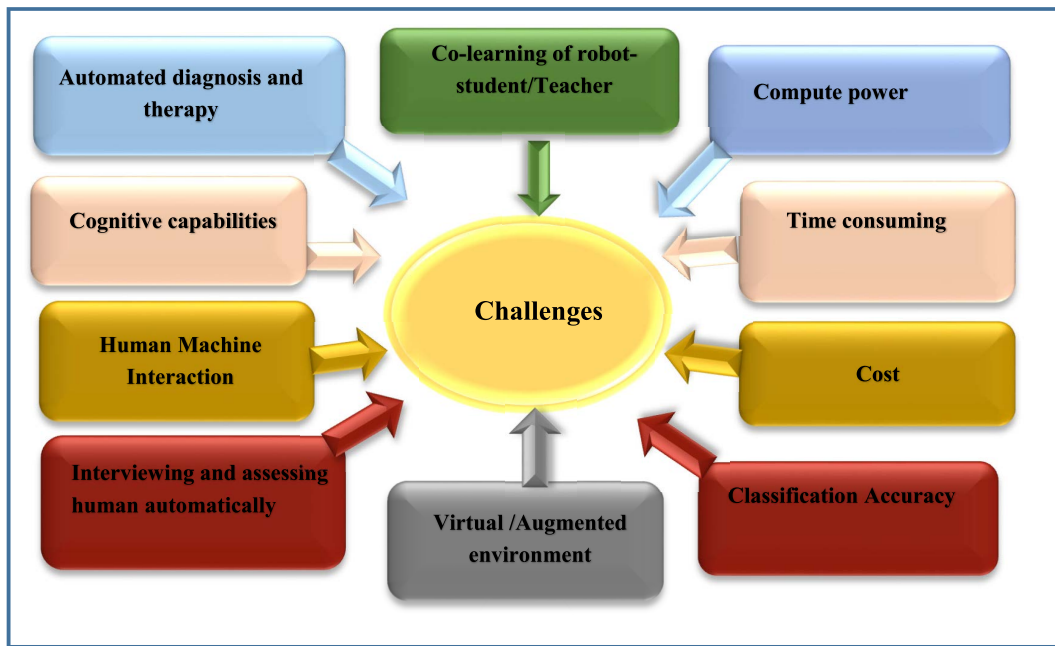


FIGURE 4. Categories of the automated vision-based assessment model.

A robot can provide instant feedback to students depending on their emotions, which is considered a challenging task [5], [76]. Humans differ in their way of expressing emotions [5]. Educational robots must possess features that enable them to maintain learners' attention for extended periods whilst they learn STEM concepts and abilities [15]. However, most educational robot platforms are powered by two motors and equipped with rudimentary sensing capabilities, such as infrared or ultrasonic collision detectors or possibly light sensors. These characteristics limit the capabilities of the interaction between robots and students. Robots lack the ability to communicate emotionally or recognise children based on their facial expressions. Teaching must be viewed as a social activity to provide a comprehensive learning experience; hence, educational robots must possess additional skills [15]. The works in the aspect of robots as a teaching assistant in the classroom are rare. The TA robots are expected to assist human teachers with a wide range of duties, particularly those that are difficult for teachers to complete individually [11].

In general, classic classroom lectures can have problems in eliciting student interaction. For effective education, social robots should motivate students to learn for an extended period by playing games. Maintaining children's interest in interacting with an assistant robot may be challenging once the initial enthusiasm wears off [74].

2) CHALLENGES RELATED TO AUTOMATED DIAGNOSIS AND THERAPY

The smart technologies, particularly robots, may provide potential methods for advancing ASD research and therapy. The next generation of therapy robots should be adopted for

individual needs [85]. The use of ICT as an intervention technique has benefited persons with multiple disabilities considerably because it can be tailored to each individual's requirements, capabilities and unique qualities. The majority of individuals with autism are visual thinkers. They have issues recalling successions and engine control in their grasp [82]. Performing as an application that understands and logically analyses the learning of a child with autism is a challenging task for a robot [82]. Therapy and ASD diagnosis based on robotic devices cannot provide autonomous feedback to increase children's interaction [86]. The challenge in ASD diagnosis requires the inclusion of a visual system in a robot assistant based on an artificial reasoning module to aid doctors with the diagnostic process. The vision system is capable of face detection, recognition and tracking of facial landmarks, head posture and gaze and estimating the visual focus of attention [86]. A study [84] focused on the tasks of evaluating children's behaviours and on the possibility of introducing a metric for assessing the efficacy of therapy based on machine learning strategies during robot-child with ASD interactions, particularly in terms of facial expression imitation. The study also considered automatically detecting and monitoring the child's face, followed by emotion recognition using a machine learning pipeline based on the HOG descriptor and SVM [84].

Predicting children's performance in HHI based on their HRI process is a critical challenge. Nevertheless, research in this area is extremely limited. The Early Social Communication Scales (ESCS) is a test that assesses nonverbal social abilities in early childhood, including RJA. Designing and applying robotic systems are necessary to assist children with

ASD improve their response to RJA skills [87]. Additionally, predicting participants' RJA performance in HHI using their HRI head position patterns is a challenge in a semisupervised machine learning context [87].

Each year, hundreds of children attend schools with various challenges, the most common of which is dyslalia or communication problems [14]. For this issue, a robotic assistant must provide auditory, tactile and visual stimulation to children receiving SLT [14]. Current robots have limitations with respect to automatic detection and response to human effects, which are required for establishing and maintaining engaging communications [83]. The ability to detect children's affective states and participation automatically during robot-assisted autism therapy is critical for the development of future autism therapies [83].

3) CHALLENGES RELATED TO COGNITIVE CAPABILITIES

Integrating robotics with AI is one of the recent research challenges. The integration of several cognitive capabilities, such as learning, context reasoning and planning, is an open problem in robotics [2]. Cognitive robotics aims to develop a new generation of control systems and provide robots with social, empathic and affective skills [79], [93]. Developing the mind of a social and emotional robot is a challenging task that requires imitating human characteristics and behaviours on a robotic artefact. Although biological systems give a wealth of inspiration, a technological design approach is still required [79].

The world's population is increasingly ageing, posing critical social and healthcare system issues. The mutual contact of a robotic platform and human users can fulfil the assistance purposes. In certain instances, robots cannot rely on human operators. Thus, robots must obtain social intelligence in a self-sufficient manner [80]. The development of robots skilled in assisting older people with various tasks involves several challenges; the most crucial of these challenges include the requirement for intelligent and continuous behaviour, robust and flexible services and the ability to respond to various conditions and demands [2], [10], [20]. Assistive robots should be provided with intelligent capabilities that enable them to think at multiple levels of analysis, recognise specific health-related requirements and realise how to act in executing specialised assistive activities [2], [20], [80]. Thus, robots must gain social intelligence to interact socially with human users by using modern machine learning algorithms [80]. Additionally, designing a multirobot system capable of real-time interactions with people based on gestures is critical for assisting users with special needs, such as the elderly or people with disability [10].

A robot is viewed as a non-invasive screening tool for assessing cognitive deterioration in the health care sector. The robot administers evaluation by giving instructions to the patient, recording his or her responses and objectively computing the final score. From a technical perspective, the challenge is to achieve a reliable voice and visual recognition

to generate a legitimate score of cognitive ability that may be used to support professional assessment [21].

Another challenge facing the researchers is specifying the necessary competencies for AI-based robot development. In particular, the relative significance of these competencies must be evaluated as perceived by the AI-based robot design professionals [3].

Frequently, vision-based systems are required to achieve continuous automatic video content analysis [18]. Two essential topics were addressed in the analysis: face recognition and human behaviour recognition. Vision-based tasks are configured to run in a cloud computing environment to improve the performance of these tasks [18]. Additionally, creating compelling cognitive applications and services has been mostly limited to powerful graphics processing units (GPUs) or the cloud because of their high processing requirements. However, clear bandwidth, energy consumption and privacy concerns exist with cloud processing [94]. In the face of an unprecedented number of modern mobile and ubiquitous applications, no adaptable open platform can be employed at present for the mobile embedded vision [94]. One of the critical difficulties is developing a robotics system that can perform advanced processing locally without the need for further computation in the cloud to reduce latency and eliminate the majority of ethical concerns [94].

4) CHALLENGES RELATED TO HUMAN–MACHINE INTERACTION

Human characteristics are the fundamental aspects to achieve effective communication. Instructors cannot observe the emotion of learners in face-to-face online learning. This issue leads to a lack of communication and interaction between the instructors and learners. Other issues include determining students' learning emotions in real time to increase lesson effectiveness and improving the communication and interaction with students [28], [31], [37]. Given the variety of expressive patterns and the difficulty of obtaining credible data, modelling emotive human expressions, particularly in a naturalistic situation, is extremely challenging [95], [101]. Developing a system with face detection and emotion recognition capabilities has been a critical field of research [95], particularly in the future ATS system [31]. The automatic recognition of the students' affective states must be performed using machine or deep learning techniques [22], [37].

Deep learning demonstrates its superior capacity in complex multiclass classification problems. One of the most popular classification areas is video facial expression detection, which is expected to become increasingly important in the robotics field [26]. Human emotions are dynamic and complicated. They are not easy to learn even when a deep CNN that performs best in numerous image databases is employed.

Few effective databases utilise students' facial expressions, hand gestures and body postures in online and classroom situations [22]. Other issues include monitoring the fatigue levels of students throughout the learning process and providing them feedback [75]. Studies in the field of detecting

learners' emotional status are few. Academic emotion recognition based on facial expressions research mostly focuses on student participation [29].

Few studies are limited to the interaction of learning procession and learning state, particularly for fatigue and confusion states [29], [75]. In addition, the automatic detection of learner's uncertainty through face based on 2D databases is another challenge [32]. Finding a global set of descriptors capable of robustly characterising human expression features is incredibly challenging [9]. A comprehensive investigation of how the HOG descriptor can be used to recognise facial expressions effectively was conducted [9]. The study of human internal state elements, such as prior knowledge, which influence the behaviour of users, is still in its early stages. Prior knowledge influences an individual's gaze behaviours, such as mutual gaze convergence between two people during a lecture task, with two conditions: interaction with and without prior knowledge [36].

Moreover, understanding human behaviour is a difficult task. Action detection has become an essential element in identifying emotional activity because of many of its essential applications. Machine learning applications have many issues, such as the need to train the system to detect human actions in real time [58]. According to research, automatic facial expression recognition is a vital field of research in mental health, as facial expressions can convey information about a person's mental condition [33]. Computer vision researchers must develop methods for accurately estimating a patient's state of depression based on nonverbal behaviour patterns to overcome the issues associated with diagnosing depression [34]. Current deep learning techniques based on 2D CNNs typically analyse spatial and temporal information separately to examine appearance information. Then, this analysis is followed by the mapping of the differences in facial features or averaging of the depression level across video frames. This technique has limitations in terms of its capacity to capture dynamic information that can aid in distinguishing different levels of depression [34].

One of the most desirable capabilities of assistive robots is emotion recognition and interpretation, which greatly improve the quality and spontaneity of HRI. [101]. However, recognising the emotional states of children based on their spontaneous bodily expressions during child-robot interactions is considered a challenging task [101]. A socially interactive robot must be constructed to investigate its capabilities for face-to-face connection with students and enable the robot to execute various tasks, such as facial expressions, gestures and speech [77]. A robot capable of learning and experimenting with the various available reinforcement techniques can gain a comprehensive understanding of human behaviour. The performance, interest level and motivation of students can all be influenced by positive reinforcement from humans or robots [19].

Robotics studies that involve long-term interactions in a real-life environment are very few. The challenge of placing a robot in a museum environment, where no one expects a

robot to wander alone and learn artwork appreciation (artificial aesthetics taste) through social interactions, was chosen [81]. Additionally, Bertacchini *et al.* (2017) integrated robotic services in the retail sector and focused on how HRI affects the shopping behaviour in smart technological stores. The robot must learn to recognise humans' emotional states either through their emotional experiences or emotional speech or via social media containing information on the customer's tastes and cultural background. Then, the robot expresses its emotional state to influence the customer's purchasing choice [91]. Developing integrated intelligent applications, such as service robot apps, including Teahouse Robot, remains challenging. These applications require various cognitive actions, such as listening, conversing, thinking, observing and moving; thus, integrating them is difficult [93]. Numerous current studies on tools for researchers, developers and end users remain limited in scope [93].

5) CHALLENGES RELATED TO INTERVIEWING AND ASSESSING HUMANS AUTOMATICALLY

Developing social robots by including a motivational interview protocol has difficulties within the constraints of current AI because participants find such kinds of robots engaging and helpful [70]. Providing robots with the capabilities to maintain user engagement is a challenging task. These capabilities include making a robot aware of the level of engagement humans demonstrate as part of interactive behaviour [89]. Moreover, studies comparing robotic assessment with human assessment or robotic assessment with computerised assessment are often lacking. Thus, additional experimental validation and investigation are required [88]. In-group presentations, a collection of specific behaviours expressed by students, are selected. Identifying and analysing student behaviours are challenging tasks because of the vast volume of observation points in a normal presentation video and the amount of time required to analyse recordings. This process is not possible without enabling analysis in settings [23], [71]. Additionally, the prospects for future HRI research were examined, including the usage of social robots as educational evaluators [76].

Another emerging challenge for the near future is designing cooperative robots skilled for aiding in public places; these robots can share tasks and communicate with humans [90]. The fundamental issues that must be addressed include the robust system design of self-directed autonomous robots with advanced environment modelling and scene understanding capabilities, distributed autonomous decision-making, short-term interaction with humans and robust and safe navigation in crowded spaces [90].

6) CHALLENGES RELATED TO VIRTUAL/AUGMENTED ENVIRONMENT

VR environment attempts to increase the sensation of being present in a virtual world. However, some approaches of VR, e.g. VTube, are technically challenging to be implemented by an ordinary person [99]. VTubers require a

high-quality and full-body motion capture system to track their motions, which is costly for average users broadcasting from a home computer. Additionally, people with little or no experience in the performing arts cannot often portray strong emotions naturally and dynamically [99]. Thus far, considerable research has concentrated on the visualisation phase [98]. The importance of rote learning and continuously imitating the dancing teacher was emphasised. Thus, quantifiable measurements and feedback are either rudimentary or non-existent. However, simply repeating and imitating the dancing instructors without receiving feedback does not always lead to increased performance [98]. Furthermore, recognising the identity of the gesture operator is a critical challenge in robotic sports teacher systems for rehabilitation and exercise instruction [102]. The accuracy of identity recognition utilising the Kinect sensor must be improved to facilitate user adaptability and assess trainee learning [102].

In drama-based learning classrooms with digital theatre, teachers devoted considerable time to teaching and classroom administration but little time to student audience engagements. However, students' engagement and learning must be enhanced when they participate in drama-based activities in the digital theatre [100]. Consequently, the attention and enthusiasm of students to study in digital learning theatres must be improved [100].

In [12], a SIIC environment that regards a smart, modern, interactive laboratory class was reviewed. The study demonstrated how telecommunication technologies, sensors and actuators can be used in a virtual environment to improve the learning process and experience [12].

VR environments have mostly been established in controlled environments, such as a clinical or research lab setting. A VR audience environment was built and implemented within a university public speaking laboratory for use in practising speeches, which were course assignments for undergraduate students [96]. The unique problems of such a system include making it accessible and useable for non-technical students with minimal supervision and determining whether utilising the system affects the speech grades of students [96].

7) CHALLENGES RELATED TO CLASSIFICATION ACCURACY

Efforts to develop interfaces that classify human interaction intuitively, such as through facial expressions and human motion recognition, have been increasing. The classifier methods should be evaluated for accuracy and efficiency [44], [51].

Facial expressions are critical in expressing and recognising emotions. Recognising facial expressions regards a crucial and challenging research area [31], [41], [43], [51], [53]. Although remarkable progress has been achieved in the field of face detection and recognition for security, identity and attendance, the issues preventing the advancement to human-level accuracy remain. Criminal and terrorist identification is a primary responsibility of police, military and security personnel. Combating the increasing rates of criminal

and terrorist activities is considered a challenging task for all security agencies. However, current technologies are not as efficient or accurate as promised in terms of identifying criminals and terrorists through facial recognition [35].

Variations in the appearance of human facial expression are caused by many factors, such as varying lighting conditions, noise in face images, scale, head poses, camera angles and occlusion [31], [41], [44], [55], [56]. These conditions influencing the most obvious feature points of the face are regarded as an RGB-D image set classification problem [31], [54]. Vision-based classification methods are highly sensitive to such external conditions. The process of FER must be executed in real time with minimum latency considering continuous changes in facial expressions [44].

In addition, the primary challenge is finding a classifier with high generalisation performance [26], [48], [52]. One of the reasons may be that the standard SoftMax loss function has limited discriminative ability [42]. Another concern is overfitting, which occurs as a result of the enormous number of parameters. The generalisation performance can be poor when the differences in geometry and texture amongst various people and databases are high [42]. Furthermore, the number of neurons is hard to determine in a given problem by utilising the traditional ELM classifier. Improving recognition accuracy whilst using few neurons and obtaining more sparse solutions than those provided by the typical ELM method is necessary [48].

The difficulties associated with face classification stem from the effort of extracting features and classifying them based on their visual appearance [47]. The recognition accuracies of most traditional approaches, such as Haar cascade, HOG descriptor, neural networks and deep learning, largely depend on feature extraction [35], [50]. An issue in a feature-extracted process is reducing the time spent on selecting the best kernel [45], [43]. The majority of current automated FER models are based on manually engineered features and pooled information from selected video frames. Although these features perform adequate results, they are not totally automated because of their need for human experience to create preferable descriptors [46], [52]. Localising face landmarks is a critical step in facial image processing. However, the issue of substantial fluctuations caused by pose disparity, illumination, expression, and occlusion remains a problematic undertaking [49]. Thus, a model that can effectively learn the relationships between image pairs associated with different facial expressions is necessary [38], [43]. Furthermore, it should have the capability to distinguish between pose variations and facial emotions by encoding them into two distinct hidden sets, namely, facial expression morphlets and non-facial expression morphlets, such as pose and occlusion [39], [46], [49], [53], [54].

Creating robust databases is a difficult task because various image variants should be regarded [22]. Several concerns exist about the datasets used to test the FER systems. Some facial expression recognition systems may perform well in some image datasets; however, they may perform poorly

in others [33]. An example of these systems is the deep learning approach, which often requires determining a large number of weights in the training phase [35], [39], [40]. These techniques have low performance when trained on a small image dataset. Another challenge with datasets is the environment's validity because most methods at present neglect real-world factors, such as illumination, image quality and background complexity. Thus, addressing such challenges for real-time applications is critical. Finally, both traditional and deep learning-based techniques have limitations. Traditional techniques, such as LBP, scale-invariant feature transform (SIFT), and SVM, rely on manually constructed features, thereby possibly resulting in unsatisfactory performance when dealing with unseen images [33]. Additionally, analysing the classification accuracy of a database utilising a few present machine and deep learning approaches is another challenge [22].

On the contrary, social robots must be able to interpret human emotions to enhance human-like performance. The issue lies in employing efficient CNNs on robotic platforms for accurate recognition tasks based on a small test dataset to establish effective HRI [67]. Additionally, the recognition accuracy of the proposed FER engine on public datasets based on HOG descriptor and SVM was compared with that of current ML strategies [84].

The issue of intended or unintended image occlusions can affect the accuracy of object tracking achieved by an active tracking robot with visual intelligence capabilities [8]. In classification-based target tracking, a huge number of labelled samples are needed to deal with the changes in the target and the complex background, and the expense of manually labelling the samples is enormous and sometimes unrealistic [57].

Human motion evaluation, which involves motion capture, data processing and motion detection and recognition, has been the focus of recent studies [97]. Current approaches for solving motion and gesture recognition problems have many issues, including a low recognition rate, slow speed and poor performance in complex environments when recognising multiple or long-distance targets. [60]. The majority of present deep learning algorithms focus on extending the neural network layer to increase the recognition performance, thereby slowing down the network's recognition speed [60]. The common trend for dynamic gesture recognition is to use multimodal data, such as depth and skeleton information. However, using only colour information would be convenient because RGB cameras are commonly available in public places and no extra equipment is needed. The difficulty of conveying spatiotemporal information by using colours only is the main issue of such a technique [59].

8) CHALLENGES RELATED TO COST

Numerous experiments, including localisation, teaching assistant and grasping, must be conducted to develop the robots' hardware. The cost of designing a robot system is considerably less than the cost of purchasing standard

robots [66]. Furthermore, a cognitive assessment via HRI with the autonomous scoring system is needed. It is inexpensive and can quickly diagnose people with psychiatric disorders, such as schizophrenia and depression [88].

The cost of developing intelligent assistive robot applications is still high for some developers. Further issues associated with the creation and reuse of workflows for robots' actions and HRI are released [67]. End users are typically unaware of the functions and properties of AI and robot applications. Discovering and altering reusable components to meet the reuse requirements when reusing complete workflows are difficult for end users [67].

In most cities, such as Ecuador's Cuenca City, only a few schools incorporate ICT and intelligent system-based technologies to enhance the intervention process amongst children from low- and middle-income families. Utilising the new technologies in the early childhood education of children is critical [72].

9) CHALLENGES RELATED TO TIME

Some tasks are time-consuming for teachers. These tasks can be achieved by automated systems or TA robots. For instance, assessing student attendance is critical within the classroom setting, although this task is a tiresome and time-consuming process [30]. In the context of restricted resources, the student attendance evaluation task must be completed automatically using facial recognition technologies [30].

Another example is that the time that teachers spend in repairing faults in the student's programs is more than the time they spend in teaching programming courses. These mistakes differ by students; consequently, a method for assisting these students without disrupting the class is necessary [17].

10) CHALLENGES RELATED TO COMPUTING POWER

According to studies on human facial emotion recognition, computational models based on regression modelling can achieve viable performance. However, many systems necessitate a large amount of computing power to function; this scenario has implications for many applications, including robotics and smart devices [25]. Furthermore, a low computing power setup is another problem, particularly when the robot is utilised to recognise gestures and read text [8]. In a robot-based monitoring system, the background subtraction algorithm must run in an online scheme at a speed that meets the requirements of certain applications. Minimising the algorithm's computational and memory costs whilst retaining high accuracy results is challenging [92].

D. RECOMMENDATIONS

This section presents the valuable recommendations of the literature in our survey, as shown in FIGURE 5.

1) RECOMMENDATIONS FOR USERS

The deployment of intelligent robots can aid instructors and minimise their burden whilst improving student excitement and learning effect in various learning

circumstances [63], [65]. Many instructors seek assistance for the regular or administrative aspects of their jobs [16]. The development of a humanoid robot by integrating the capabilities of emotion recognition and facial expression mimicking is necessary for a teaching assistant dealing with children [68] with autism [27], [82]. Sophisticated planning techniques on NAO, such as methods to assist the robot in deciding how to progress with instructional tasks based on an assessment of the students' academic achievements, emotional state and motivation, are necessary [69]. Furthermore, integrating ZENBO robot with auditory and visual signal recognition is recommended to design it with cognition capabilities. Thus, the robot may engage with learners in various independent and intelligent ways through the help of a virtual environment [100]. Numerous educational tools and experiments incorporating AR or VR are required to optimise learning outcomes [12].

Considerable research on the characteristics of robots, including role models, communication, cooperation, nonverbal feedback, attention building and empathy, is required [69], [79], [86]. Building a new generation of emotional machines, such as robots with interaction capabilities, is expected to draw the attention of students and encourage their active involvement [11]. People who may not feel ready to talk to a human counsellor would engage in a conversation by interacting with a robot [70]. Adding a level of emotional and behavioural learning engagement and improving self-learning based on various scales with a large number of research participants are recommended for robotics systems [1]. Developing a TA robot by combining technologies and conducting experiments in a real classroom is recommended to enhance the robot's performance based on student and instructor feedback [17]. Further research is needed to determine the relevant and long-term implications of social robots interacting and engaging with students in real educational environments [62], [73]. Exploring other strategies instead of a neutral feedback comment by the robot for each user and adding studies to literature for human-machine communication instead of classic human-to-human communication science are necessary [76].

2) RECOMMENDATIONS FOR DEVELOPERS

This section contains relevant advice for researchers and developers in this field of study. The majority of recommendations concern performance, accuracy and system model.

a: RECOMMENDATIONS FOR PERFORMANCE

This subsection discusses the recommendations of the studies in our survey according to the solutions to improve the performance by focusing on the hardware tools, long-term studies and datasets. Utilising feedback sensors and the camera can reduce the error in different tests performed and localise the system correctly [66]. Placing low-power cameras minimises the device's power consumption, extends its battery life and improves face detection and emotion recognition ability [94]. Additionally, adopting time-of-flight

cameras, such as the Kinect2, or other sophisticated data-collection equipment is recommended for precisely detecting body joints [98]. Furthermore, an extensive analysis of the influence of non-frontal face in future work is necessary [9].

Utilising the emotion recognition technology is one of the interactive methods for emotion detection in any scenario, such as academic emotions in online learning [29] or civilians in an emergency scenario [41]. Various kinds of academic emotions, such as engagement, boredom and frustration, can have a substantial impact on learning; thus, several levels of academic emotion recognition methods can be considered by utilising efficient algorithms [29]. Exploring machine learning methods is recommended to take advantage of the large dataset for training to increase the reliability and robustness of the system [30]. The performance of the TSVM framework can be enhanced further by training this classifier with a large amount of data [87]. Additionally, the framework for mental state detection and diagnosis must be trained using huge samples of natural facial expressions taken from older persons to optimise its performance [33]. Moreover, further research on cultural and gender differences is needed to obtain a deep understanding of human behaviour variation and its potential impacts [36], [64]. These studies are crucial for improving the effectiveness and adaptability of user-centred adaptive systems [36].

Additionally, considering multiscale features and multiple network fusion can boost generalisation performance between databases. For instance, expression synthesis can enhance the MMI dataset and fine-tune the FER2013 database [42]. Considering the different sparseness techniques and regularisation coefficients may be acceptable for different databases; multiple sparseness losses can be dynamically modified to adapt the sparseness strategy and regularisation parameter setting for a particular database [42].

b: RECOMMENDATIONS FOR ACCURACY

The articles in our survey have many recommendations that aim to solve system problems or develop system models. One of the problems is that the poor samples in the dataset may lead to reduced accuracy. Thus, integrating classifications methods or models further distinguishes these hard samples [54]. Other aspects that assist intelligent systems to be increasingly accurate in the future must be investigated [37]. Considerable research is needed to solve accuracy issues, such as lighting and occlusion, so that the solution may be used by developers for low-spec processing machines [31]. Other accurate models of emotion recognition and effective ways of exploring behaviours should be also studied [28], [45].

Emotion recognition is a multidisciplinary field that is challenging to recognise precisely. Accurate results are produced from utilising good face detection and face alignment techniques [26] [40]. In addition, various techniques can improve the models of emotion recognition. Numerous facial feature landmarks would be used to increase the recall and precision rates for recognising facial emotions, such as

fatigue [75]. Many absolute criteria for facial AU packaging, such as simultaneous detection of AUs that are completely unrelated or even exclusively behaved, must be considered, and AU packaging should be disregarded to determine whether training the model concurrently would affect the algorithm's performance [43]. Furthermore, facial tilt, i.e. the angle of the user's face when he or she faces upwards, must be considered; some of the user's face parts are not seen because of facial tilt, resulting in information loss [39], [40].

The performance of the learnt feature representations must be evaluated when applied to the objective of RGB-D based image set classification using the deep CNN because a considerable quantity of training data is necessary to learn the parameters of a deep CNN model [56]. A refined version of the suggested CNN would be developed to boost the proportion of correctly detected emotions [38], [94]. The deep CNN methods can be extended to determine the changes in facial emotion using a video sequence in many real-time feedback applications [51], [53]. As a pre-training strategy for deep CNNs, using SCAE with big datasets increases performance whilst reducing training time [40]. The proposed CNN-SVM algorithm in [44] is efficient and accurate for emotion recognition and can be embedded for human-computer or human-robot interaction applications. In addition, the method (HOG, BPNN) would be included in conjunction with feature selection algorithms on face recognition to increase the recognition rate [47]. Furthermore, employing various LPQ, EOH feature extraction, or regression modelling approaches, such as SVM, can improve the accuracy of the emotional state recognition system [25]. Approximate solutions should be developed to enable the new approach of SVM-based multiclass classification that takes advantage of geometric data connections applied to various classification issues [24].

Expression change should be detected when a person's expression changes over time. In this scenario, functioning with face detection and emotion recognition in real time is challenging for Microsoft HoloLens (MHL). The algorithm of MHL needs to work in real time rather than just in video mode, and its accuracies can be increased by utilising a large dataset for emotion recognition in real time [95]. Moreover, the motion analysis framework proposed by [97] can be utilised to train unsupervised physical exercise with any application, as it can perform accurate and comprehensive motion analysis using any available device. The MSN model developed in [34] functions accurately in other health care applications that utilise face information for automatic depression assessment.

3) RECOMMENDATIONS FOR SYSTEM MODEL

Combining many modalities, such as bodily, facial and verbal cues, has a considerable impact on the robot's perception of human behaviour [101]. Additionally, a complex sample selection technique can improve prediction accuracy by retaining the most informative and contributing data frames in the model [101]. Combining the concept of gesture

recognition with intelligent robots for use in specific scenarios has been planned [60]. A new classifier system that considers the hand information of each gesture, such as a real-time spotting gesture algorithm, should be developed for the communication with the robot [59]. Increasing the algorithm's speed and achieving high accuracy with a real-time background subtraction system are also essential for robotic systems when they are employed in the real world [92]. The speed of the discriminating ELM algorithm, which is a new facial recognition approach based on ELM, must be increased by improving the formula (L2) because its training time is longer than the training time of the standard ELM [48].

The comparison of the robot's performance as a deep learning agent with its performance as a deep reinforcement-learning agent in the same environment gains attention [80]. The CNN used in [80] would be integrated into a deep RL agent to conduct an experiment in which the robot makes decisions in an uncontrolled environment using the trained CNN. On the contrary, integrating RL approaches into the engaging prediction model is a component of the robot's reward function [89].

A long-term study and a large sample of participants would precisely measure the educational effectiveness of the robot as a teacher [15], [69]. A long-term study should be performed to validate the interaction approach and enhance the user experience [75]. The user interface of the current robotic system must be improved by providing effective functionalities [77]. Additional research is required to provide important knowledge about how robots train themselves over time to perform tasks and make appropriate judgments in collaboration with their human partners [19]. In [9], a humanoid robot would be supplied with the proposed FER system to learn about the emotional state of the interacting user and respond accordingly with a specific action. Optimising the FER engine's algorithms is essential to utilise fully the processing resources on board the R25 robot [84].

Additional training and post-processing are required for the IBM Watson Broadband model, which was implemented in the pepper robot for cognitive assessments, to improve the quality of speech transcriptions and visual recognition of people who have pronunciation problems, such as stroke patients or those with a low education level [21].

4) OTHERS

This section includes recommendations that are not relevant to the groups described above.

The recommendation of [35] can be expanded to the automated criminal identification system (CIS, allowing live broadcasting and sending alert messages to users' mobile phones with some extra capabilities. The existing system can be handled only by the administrator.

The next task of [81] is to enhance the readability behaviour of the robot museum visitor by incorporating a mechanism that would cause the robot to pause in front of artworks during the period of visual scene habituation. Another recommendation is to create evaluation criteria for

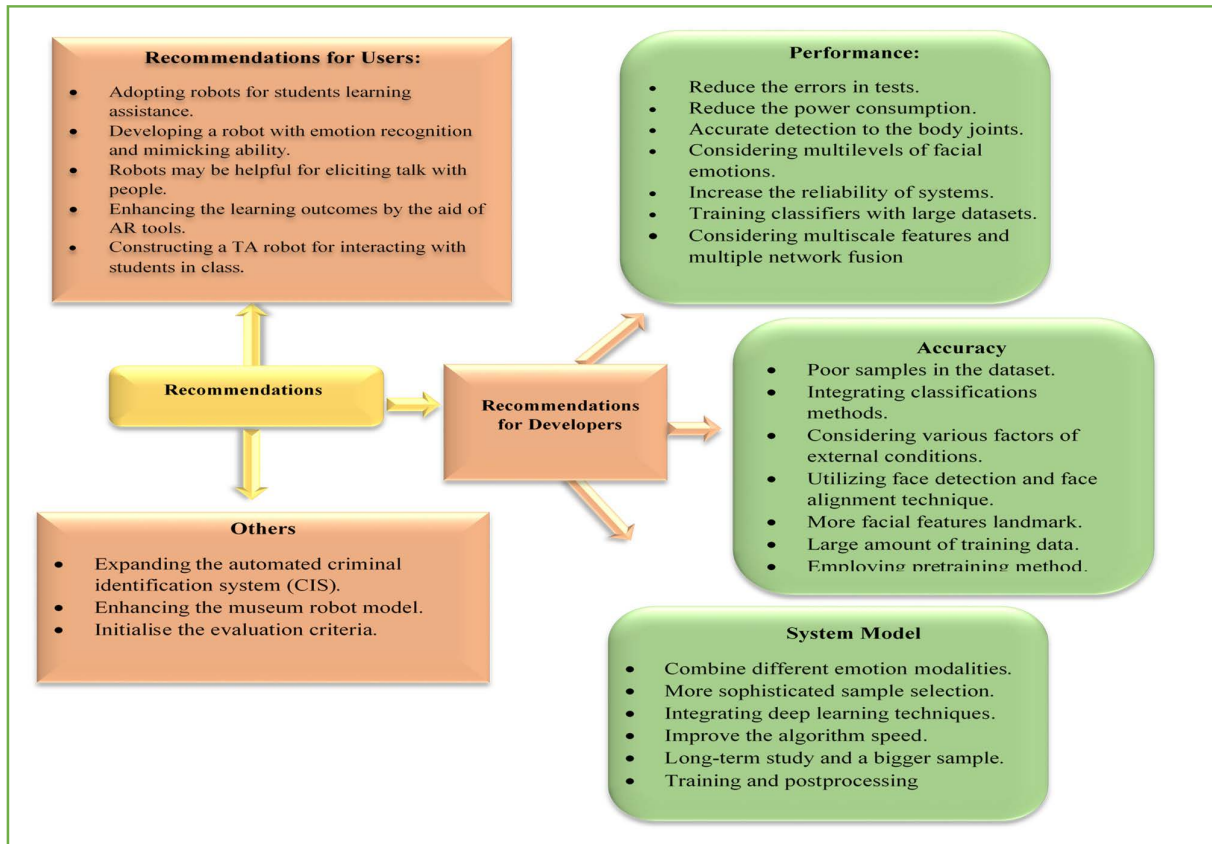


FIGURE 5. Categories of recommendations for the automated vision-based assessment models.

advanced intelligent applications and utilise these metrics for intelligent platform evaluation [93].

V. BIBLIOGRAPHY ANALYSIS

As illustrated in the following subsections, the key demographic statistical findings from the articles in survey are presented based on three aspects: years of publication, publisher databases and countries of publication.

A. YEARS OF PUBLICATION

FIGURE 5 shows the increasing interest in developing the automated vision-based assessment model.

FIGURE 6 also indicates the articles in our survey according to the publication year. The highest number of publications is recorded in 2019, including 12 papers in computer-vision-based category, 14 papers in robotics-based category and 1 paper in AR-based category. The year 2018 has the second-highest number of published articles, including 5, 13 and 3 papers in computer-vision-based, robotics-based and AR-based categories, respectively. The years 2017 and 2020 have 18 and 17 published papers, respectively. The year 2015 has the minimum number of publications, with only six papers from all the four categories.

B. PUBLISHER DATABASES

The papers in our survey were collected from the three databases, namely, IEEE, SD and WOS. Each of these databases is involved with various journals. FIGURE 7 shows the number of papers according to the main category per database. Computer-vision-based, robotics-based, AR-based and hybrid-based categories have 41, 47, 6 and 4 articles, respectively. The robotic-based category has the higher number of articles, followed by the computer-vision-based category, with 23 and 18 papers from IEEE, respectively. Moreover, the robotics category has the highest number of published articles, followed by computer-vision-based category, involving 18 and 14 papers, respectively. By contrast, the computer-vision-based category has the highest number of published articles in the SD database, followed by robotics-based category, with 14 and 5 papers, respectively. The AR-based category has the following number of articles per database: three papers in WOS, two papers in SD and one paper in the IEEE database.

C. COUNTRIES OF PUBLICATION

The countries of the corresponding authors are shown in FIGURE 8. The majority of the studies related to the

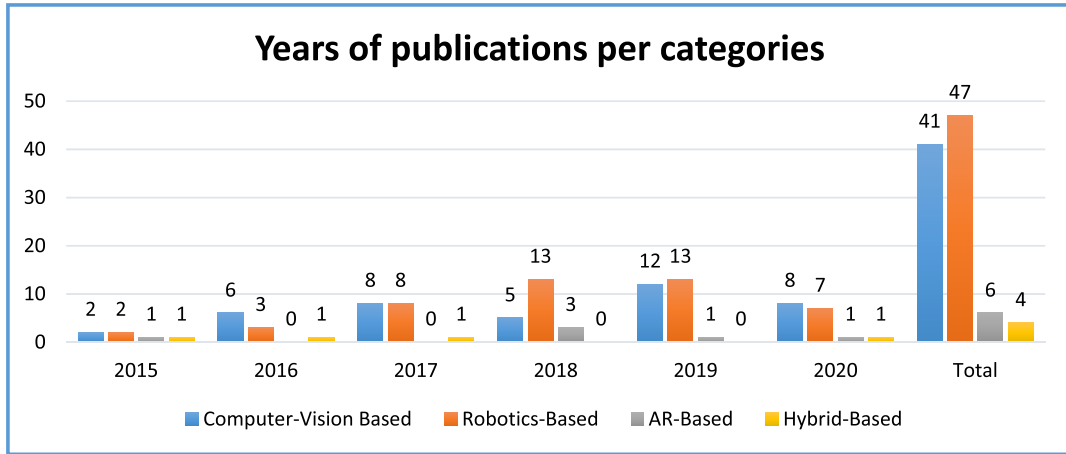


FIGURE 6. Statistics of the studies according to category in our taxonomy per year of publication.

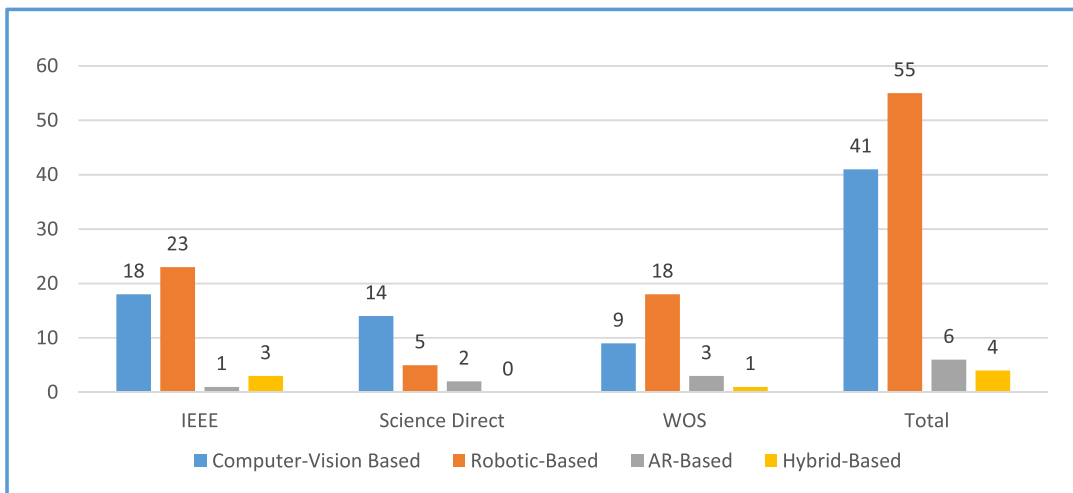


FIGURE 7. Statistics of the studies according to the main categories per database.

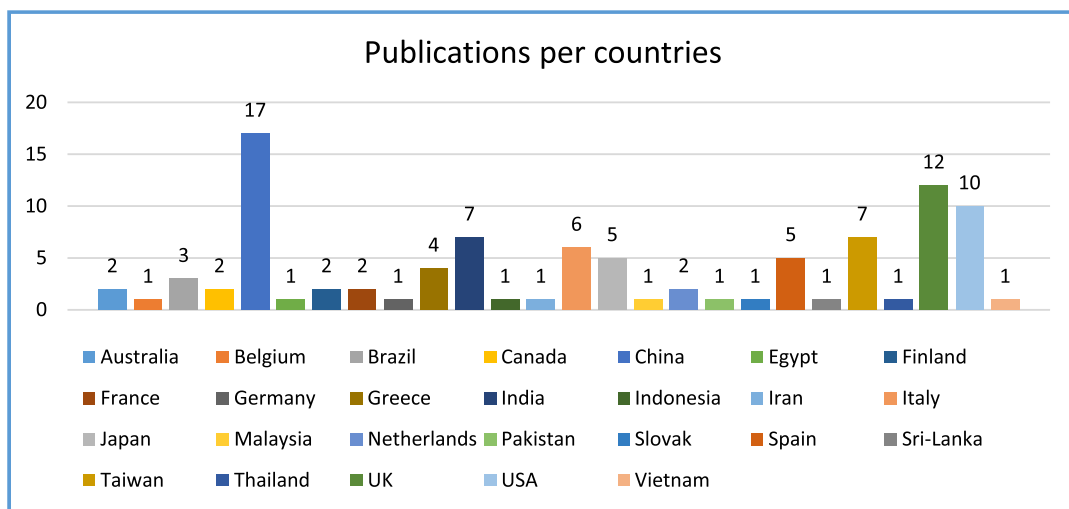


FIGURE 8. Publications per country.

automated vision-based assessment model are from China, with 17 papers. Moreover, 12 and 10 studies come from

UK and USA. India and Taiwan published seven studies each. Six studies come from Italy. Japan and Spain have

five studies each. Greece and Brazil published four and three studies, respectively. Australia, Canada, Finland, France and Netherlands have two studies each. Furthermore, Belgium, Egypt, Germany, Indonesia, Iran, Malaysia, Pakistan, Slovak, Sri-Lanka, Sri-Lanka and Vietnam have one study each.

VI. CONCLUSION

As far as we know, the research effort exploring the area of automated vision-based assessment systems is still increasing. The current study presents a systematic literature review structure, which describes the environment of the investigated field. The study initially presented the systematic review protocol in the area of automated vision-based assessment systems to obtain understanding and insights in this field. A large amount of important information was obtained through extensive reading and analysis of various reviewed articles to highlight the challenges and issues, motivations and benefits, recommendations related to automated vision-based assessment models and their various datasets. The studies in this review were categorised into four, computer-vision-based, robotics-based, AR-based and hybrid-based, through survey and classification methods to obtain deep insights from the investigated field. The goal is to present the automatic evaluation framework for assistive technologies and improve the perception on these technologies. This study provides researchers with essential guides and valuable information for future research. Furthermore, this study addresses the relevant descriptions and boundaries that remain ambiguous in the fields of AI, Robotic, and AR.

ACKNOWLEDGMENT

The authors would like to thank Universiti Kebangsaan Malaysia (UKM) for the financial support.

REFERENCES

- [1] E. Chew and X. N. Chua, "Robotic Chinese language tutor: Personalising progress assessment and feedback or taking over your job?" *Horizon*, vol. 28, no. 3, pp. 113–124, Jul. 2020, doi: [10.1108/OTH-04-2020-0015](https://doi.org/10.1108/OTH-04-2020-0015).
- [2] A. Umbrico, A. Cesta, G. Cortellessa, and A. Orlandini, "A holistic approach to behavior adaptation for socially assistive robots," *Int. J. Social Robot.*, vol. 12, no. 3, pp. 617–637, Jul. 2020, doi: [10.1007/s12369-019-00617-9](https://doi.org/10.1007/s12369-019-00617-9).
- [3] W. Shyr, F. Yang, P. Liu, Y. Hsieh, C. You, and D. Chen, "Development of assessment indicators for measuring the Student learning effects of artificial intelligence-based robot design," *Comput. Appl. Eng. Educ.*, vol. 27, no. 4, pp. 863–868, Jul. 2019, doi: [10.1002/cae.22118](https://doi.org/10.1002/cae.22118).
- [4] T. J. Sheng, M. S. Islam, N. Misran, M. H. Baharuddin, H. Arshad, M. R. Islam, M. E. H. Chowdhury, H. Rmili, and M. T. Islam, "An Internet of Things based smart waste management system using Lora and tensorflow deep learning model," *IEEE Access*, vol. 8, pp. 148793–148811, 2020, doi: [10.1109/ACCESS.2020.3016255](https://doi.org/10.1109/ACCESS.2020.3016255).
- [5] Y.-Z. Hsieh, S.-S. Lin, Y.-C. Luo, Y.-L. Jeng, S.-W. Tan, C.-R. Chen, and P.-Y. Chiang, "ARCS-assisted teaching robots based on anticipatory computing and emotional big data for improving sustainable learning efficiency and motivation," *Sustainability*, vol. 12, no. 14, p. 5605, Jul. 2020, doi: [10.3390/su12145605](https://doi.org/10.3390/su12145605).
- [6] Z. Pei and Y. Nie, "Educational robots: Classification, characteristics, application areas and problems," in *Proc. 7th Int. Conf. Educ. Innov. Through Technol. (EITT)*, Dec. 2018, pp. 57–62, doi: [10.1109/EITT.2018.00020](https://doi.org/10.1109/EITT.2018.00020).
- [7] N. S. Sani, I. I. S. Shamsuddin, S. Sahran, A. H. A. Rahman, and E. N. Muzaffar, "Redefining selection of features and classification algorithms for room occupancy detection," *Int. J. Adv. Sci., Eng. Inf. Technol.*, vol. 8, nos. 4–2, p. 1486, Oct. 2018, doi: [10.18517/ijaseit.8.4-2.6826](https://doi.org/10.18517/ijaseit.8.4-2.6826).
- [8] P. B. Nithin, R. A. Francis, A. J. Chemmanam, B. A. Jose, and J. Mathew, "Interactive robotic testbed for performance assessment of machine learning based computer vision techniques," *J. Inf. Sci. Eng.*, vol. 36, no. 5, pp. 1055–1067, Sep. 2020, doi: [10.6688/IJSE.202009_36\(5\).0008](https://doi.org/10.6688/IJSE.202009_36(5).0008).
- [9] P. Carcagni, M. D. Coco, M. Leo, and C. Distanto, "Facial expression recognition and histograms of oriented gradients: A comprehensive study," *SpringerPlus*, vol. 4, no. 1, p. 645, Dec. 2015, doi: [10.1186/s40064-015-1427-3](https://doi.org/10.1186/s40064-015-1427-3).
- [10] G. Canal, S. Escalera, and C. Angulo, "A real-time human-robot interaction system based on gestures for assistive scenarios," *Comput. Vis. Image Understand.*, vol. 149, pp. 65–77, Aug. 2016, doi: [10.1016/j.cviu.2016.03.004](https://doi.org/10.1016/j.cviu.2016.03.004).
- [11] Z. Sun, Z. Li, and T. Nishimori, "Development and assessment of robot teaching assistant in facilitating learning," in *Proc. Int. Conf. Educ. Innov. through Technol. (EITT)*, Dec. 2017, pp. 165–169, doi: [10.1109/EITT.2017.47](https://doi.org/10.1109/EITT.2017.47).
- [12] C. Stergiou, A. P. Plageras, K. E. Psannis, T. Xifilidis, G. Kokkonis, S. Kontogiannis, K. Tsarava, and A. Sapountzi, "Proposed high level architecture of a smart interconnected interactive classroom," in *Proc. South-Eastern Eur. Design Autom., Comput. Eng., Comput. Netw. Soc. Media Conf. (SEEDA_CECNSM)*, Sep. 2018, pp. 1–6, doi: [10.23919/SEEDA-CECNSM.2018.8544922](https://doi.org/10.23919/SEEDA-CECNSM.2018.8544922).
- [13] N. F. Zainal, R. Din, M. F. Nasrudin, S. Abdullah, A. H. A. Rahman, S. N. H. S. Abdullah, and N. A. A. Majid, "Robotic prototype and module specification for increasing the interest of Malaysian students in STEM education," *Int. J. Eng. Technol.*, vol. 7, pp. 286–290, Aug. 2018, doi: [10.14419/ijet.v7i3.25.17583](https://doi.org/10.14419/ijet.v7i3.25.17583).
- [14] V. Velasquez-Angamarca, K. Mosquera-Cordero, V. Robles-Bykbaev, A. Leon-Pesantez, D. Krupke, J. Knox, V. Torres-Segarra, and P. Chicaiza-Juela, "An educational robotic assistant for supporting therapy sessions of children with communication disorders," in *Proc. 7th Int. Eng., Sci. Technol. Conf. (IESTEC)*, Oct. 2019, pp. 586–591, doi: [10.1109/IESTEC46403.2019.00110](https://doi.org/10.1109/IESTEC46403.2019.00110).
- [15] M. Naya, G. Varela, L. Llamas, M. Bautista, J. C. Becerra, F. Bellas, A. Prieto, A. Deibe, and R. J. Duro, "A versatile robotic platform for educational interaction," in *Proc. 9th IEEE Int. Conf. Intell. Data Acquisition Adv. Comput. Syst., Technol. Appl. (IDAACS)*, Sep. 2017, pp. 138–144, doi: [10.1109/IDAACS.2017.8095065](https://doi.org/10.1109/IDAACS.2017.8095065).
- [16] Y.-W. Cheng, P.-C. Sun, and N.-S. Chen, "An investigation of the needs on educational robots," in *Proc. IEEE 17th Int. Conf. Adv. Learn. Technol. (ICALT)*, Jul. 2017, pp. 536–538, doi: [10.1109/ICALT.2017.115](https://doi.org/10.1109/ICALT.2017.115).
- [17] K. Yoshino and S. Zhang, "Construction of teaching assistant robot in programming class," in *Proc. 7th Int. Congr. Adv. Appl. Informat. (IIAI-AAI)*, Jul. 2018, pp. 215–220, doi: [10.1109/IIAI-AAI.2018.00049](https://doi.org/10.1109/IIAI-AAI.2018.00049).
- [18] J.-Y. Huang and W.-P. Lee, "Enabling vision-based services with a cloud robotic system," in *Proc. Asia-Pacific Conf. Intell. Robot Syst. (ACIRS)*, Jul. 2016, pp. 84–88, doi: [10.1109/ACIRS.2016.7556193](https://doi.org/10.1109/ACIRS.2016.7556193).
- [19] S. Roy, C. Crick, E. Kieson, and C. Abramson, "A reinforcement learning model for robots as teachers," in *Proc. 27th IEEE Int. Symp. Robot Hum. Interact. Commun. (RO-MAN)*, Aug. 2018, pp. 294–299, doi: [10.1109/ROMAN.2018.8525563](https://doi.org/10.1109/ROMAN.2018.8525563).
- [20] A. Umbrico, G. Cortellessa, A. Orlandini, and A. Cesta, "Toward intelligent continuous assistance," *J. Ambient Intell. Humanized Comput.*, vol. 12, no. 4, pp. 4513–4527, Apr. 2021, doi: [10.1007/s12652-020-01766-w](https://doi.org/10.1007/s12652-020-01766-w).
- [21] S. Varrasi, A. Lucas, A. Soranzo, J. McNamara, and A. D. Nuovo, "IBM cloud services enhance automatic cognitive assessment via human-robot interaction," in *Mech. Mach. Sci.*, vol. 65, pp. 169–176, Sep. 2019, doi: [10.1007/978-3-030-00329-6_20](https://doi.org/10.1007/978-3-030-00329-6_20).
- [22] A. T. S. and R. M. R. Guddeti, "Affective database for e-learning and classroom environments using Indian students' faces, hand gestures and body postures," *Future Gener. Comput. Syst.*, vol. 108, pp. 334–348, Jul. 2020, doi: [10.1016/j.future.2020.02.075](https://doi.org/10.1016/j.future.2020.02.075).
- [23] A. Fekry, G. Dafoulas, and M. Ismail, "Automatic detection for students behaviors in a group presentation," in *Proc. 14th Int. Conf. Comput. Eng. Syst. (ICCES)*, Dec. 2019, Art. no. 201914, doi: [10.1109/ICCES48960.2019.9068128](https://doi.org/10.1109/ICCES48960.2019.9068128).
- [24] A. Iosifidis and M. Gabbouj, "Multi-class support vector machine classifiers using intrinsic and penalty graphs," *Pattern Recognit.*, vol. 55, pp. 231–246, Jul. 2016, doi: [10.1016/j.patcog.2016.02.002](https://doi.org/10.1016/j.patcog.2016.02.002).

- [25] S. Turabzadeh, H. Meng, R. M. Swash, M. Pleva, and J. Juhar, "Real-time emotional state detection from facial expression on embedded devices," in *Proc. 7th Int. Conf. Innov. Comput. Technol. (INTECH)*, Aug. 2017, pp. 46–51, doi: [10.1109/INTECH.2017.8102423](https://doi.org/10.1109/INTECH.2017.8102423).
- [26] Z. He, T. Jin, A. Basu, J. Soraghan, G. Di Caterina, and L. Petropoulakis, "Human emotion recognition in video using subtraction pre-processing," in *ACM Int. Conf. Proc. Ser.*, vol. 2019, vol. Part F1481, pp. 374–379, doi: [10.1145/3318299.3318321](https://doi.org/10.1145/3318299.3318321).
- [27] A. Esfandbod, Z. Rokhi, A. Taheri, M. Alemi, and A. Meghdari, "Human-robot interaction based on facial expression imitation," in *Proc. 7th Int. Conf. Robot. Mechatronics (ICRoM)*, Nov. 2019, pp. 69–73, doi: [10.1109/ICRoM48714.2019.9071837](https://doi.org/10.1109/ICRoM48714.2019.9071837).
- [28] C. Ma, C. Sun, D. Song, X. Li, and H. Xu, "A deep learning approach for online learning emotion recognition," in *Proc. 13th Int. Conf. Comput. Sci. Educ. (ICCSE)*, Aug. 2018, pp. 567–571, doi: [10.1109/ICCSE.2018.8468741](https://doi.org/10.1109/ICCSE.2018.8468741).
- [29] Z. Shi, Y. Zhang, C. Bian, and W. Lu, "Automatic academic confusion recognition in online learning based on facial expressions," in *Proc. 14th Int. Conf. Comput. Sci. Educ. (ICCSE)*, Aug. 2019, pp. 528–532, doi: [10.1109/ICCSE.2019.8845348](https://doi.org/10.1109/ICCSE.2019.8845348).
- [30] A. G. Menezes, J. M. D. D. C. Sa, E. Llapa, and C. A. Estombelo-Montesco, "Automatic attendance management system based on deep one-shot learning," in *Proc. Int. Conf. Syst., Signals, Image Process.*, Jul. 2020, pp. 137–142, doi: [10.1109/IWSSIP48289.2020.9145230](https://doi.org/10.1109/IWSSIP48289.2020.9145230).
- [31] P. Utami, R. Hartanto, and I. Soesanti, "A study on facial expression recognition in assessing teaching skills: Datasets and methods," *Proc. Comput. Sci.*, vol. 161, pp. 544–552, Jan. 2019, doi: [10.1016/j.procs.2019.11.154](https://doi.org/10.1016/j.procs.2019.11.154).
- [32] S. A. A. Daud and S. L. Lutfi, "Towards the detection of learner's uncertainty through face," in *Proc. 4th Int. Conf. User Sci. Eng. (i-USER)*, Aug. 2016, pp. 227–231, doi: [10.1109/IUSER.2016.7857965](https://doi.org/10.1109/IUSER.2016.7857965).
- [33] Z. Fei, E. Yang, D. D.-U. Li, S. Butler, W. Ijomah, X. Li, and H. Zhou, "Deep convolution network based emotion analysis towards mental health care," *Neurocomputing*, vol. 388, pp. 212–227, May 2020, doi: [10.1016/j.neucom.2020.01.034](https://doi.org/10.1016/j.neucom.2020.01.034).
- [34] W. C. de Melo, E. Granger, and A. Hadid, "A deep multiscale spatiotemporal network for assessing depression from facial dynamics," *IEEE Trans. Affect. Comput.*, early access, Sep. 4, 2020, doi: [10.1109/taffc.2020.3021755](https://doi.org/10.1109/taffc.2020.3021755).
- [35] K. Rasanayagam, S. D. D. C. Kumarasiri, W. A. D. D. Tharuka, N. T. Samaranyake, P. Samarasinghe, and S. E. R. Siriwardana, "CIS: An automated criminal identification system," in *Proc. IEEE Int. Conf. Inf. Autom. Sustainability (ICIAfS)*, Dec. 2018, pp. 1–6, doi: [10.1109/ICIAfS.2018.8913367](https://doi.org/10.1109/ICIAfS.2018.8913367).
- [36] C. Thepsonthorn, T. Yokozuka, J. Kwon, R. M. S. Yap, S. Miura, K.-I. Ogawa, and Y. Miyake, "Does user's prior knowledge worth consideration?: The influence of prior knowledge toward mutual gaze convergence," in *Proc. IEEE 40th Annu. Comput. Softw. Appl. Conf. (COMPSAC)*, Jun. 2016, pp. 515–520, doi: [10.1109/COMPSAC.2016.99](https://doi.org/10.1109/COMPSAC.2016.99).
- [37] G. Li and Y. Wang, "Research on learner's emotion recognition for intelligent education system," in *Proc. IEEE 3rd Adv. Inf. Technol., Electron. Autom. Control Conf. (IAEAC)*, Oct. 2018, pp. 754–758, doi: [10.1109/IAEAC.2018.8577590](https://doi.org/10.1109/IAEAC.2018.8577590).
- [38] S. Sharma, K. Kumar, and N. Singh, "D-FES: Deep facial expression recognition system," in *Proc. Conf. Inf. Commun. Technol. (CICT)*, Nov. 2017, pp. 1–6, doi: [10.1109/INFOCOMTECH.2017.8340635](https://doi.org/10.1109/INFOCOMTECH.2017.8340635).
- [39] N. Webb, A. Ruiz-Garcia, M. Elshaw, and V. Palade, "Emotion recognition from face images in an unconstrained environment for usage on social robots," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2020, pp. 1–8, doi: [10.1109/IJCNN48605.2020.9207494](https://doi.org/10.1109/IJCNN48605.2020.9207494).
- [40] A. Ruiz-Garcia, M. Elshaw, A. Althahhan, and V. Palade, "Stacked deep convolutional auto-encoders for emotion recognition from facial expressions," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 1586–1593, doi: [10.1109/IJCNN.2017.7966040](https://doi.org/10.1109/IJCNN.2017.7966040).
- [41] A. R. Kurup, M. Ajith, and M. M. Ramón, "Semi-supervised facial expression recognition using reduced spatial features and deep belief networks," *Neurocomputing*, vol. 367, pp. 188–197, Nov. 2019, doi: [10.1016/j.neucom.2019.08.029](https://doi.org/10.1016/j.neucom.2019.08.029).
- [42] W. Xie, X. Jia, L. Shen, and M. Yang, "Sparse deep feature learning for facial expression recognition," *Pattern Recognit.*, vol. 96, Dec. 2019, Art. no. 106966, doi: [10.1016/j.patcog.2019.106966](https://doi.org/10.1016/j.patcog.2019.106966).
- [43] X. Zhang and M. H. Mahoor, "Task-dependent multi-task multiple kernel learning for facial action unit detection," *Pattern Recognit.*, vol. 51, pp. 187–196, Mar. 2016, doi: [10.1016/j.patcog.2015.08.026](https://doi.org/10.1016/j.patcog.2015.08.026).
- [44] X. Liu and K. Lee, "Optimized facial emotion recognition technique for assessing user experience," in *Proc. IEEE Games, Entertainment, Media Conf. (GEM)*, Aug. 2018, pp. 1–9, doi: [10.1109/GEM.2018.8516518](https://doi.org/10.1109/GEM.2018.8516518).
- [45] A. S. Alphonse and D. Dharma, "A novel monogenic directional pattern (MDP) and pseudo-Voigt kernel for facilitating the identification of facial emotions," *J. Vis. Commun. Image Represent.*, vol. 49, pp. 459–470, Nov. 2017, doi: [10.1016/j.jvcir.2017.10.008](https://doi.org/10.1016/j.jvcir.2017.10.008).
- [46] S. Elaiwat, M. Bennamoun, and F. Boussaid, "A spatio-temporal RBM-based model for facial expression recognition," *Pattern Recognit.*, vol. 49, pp. 152–161, Jan. 2016, doi: [10.1016/j.patcog.2015.07.006](https://doi.org/10.1016/j.patcog.2015.07.006).
- [47] B. Nassih, A. Amine, M. Ngadi, and N. Hmina, "DCT and HOG feature sets combined with BPNN for efficient face classification," *Proc. Comput. Sci.*, vol. 148, pp. 116–125, Jan. 2019, doi: [10.1016/j.procs.2019.01.015](https://doi.org/10.1016/j.procs.2019.01.015).
- [48] L.-C. Wang, M. Liu, X. Lu, and H.-J. Lin, "Facial expression recognition via neurons partially activated discriminated ELM," in *Proc. 13th World Congr. Intell. Control Autom. (WCICA)*, Jul. 2018, pp. 710–715.
- [49] J. Zhang and H. Hu, "Stacked hourglass network joint with salient region attention refinement for face alignment," in *Proc. 14th IEEE Int. Conf. Autom. Face Gesture Recognit., (FG)*, May 2019, pp. 1–7, doi: [10.1109/FG.2019.8756590](https://doi.org/10.1109/FG.2019.8756590).
- [50] K. Shan, J. Guo, W. You, D. Lu, and R. Bie, "Automatic facial expression recognition based on a deep convolutional-neural-network structure," in *Proc. IEEE 15th Int. Conf. Softw. Eng. Res., Manage. Appl. (SERA)*, Jun. 2017, pp. 123–128, doi: [10.1109/SERA.2017.7965717](https://doi.org/10.1109/SERA.2017.7965717).
- [51] E. Pranav, S. Kamal, C. Sathesh Chandran, and M. H. Supriya, "Facial emotion recognition using deep convolutional neural network," in *Proc. 6th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, Mar. 2020, pp. 317–320, doi: [10.1109/ICACCS48705.2020.9074302](https://doi.org/10.1109/ICACCS48705.2020.9074302).
- [52] D. V. Sang, N. Van Dat, and D. P. Thuan, "Facial expression recognition using deep convolutional neural networks," in *Proc. 9th Int. Conf. Knowl. Syst. Eng., (KSE)*, Oct. 2017, pp. 130–135, doi: [10.1109/KSE.2017.8119447](https://doi.org/10.1109/KSE.2017.8119447).
- [53] K.-Y. Tsai, J.-J. Ding, and Y.-C. Lee, "Frontalization with adaptive exponentially-weighted average ensemble rule for deep learning based facial expression recognition," in *Proc. IEEE Asia Pacific Conf. Circuits Syst. (APCCAS)*, Oct. 2018, pp. 447–450, doi: [10.1109/APCCAS.2018.8605689](https://doi.org/10.1109/APCCAS.2018.8605689).
- [54] Y. Zeng, N. Xiao, K. Wang, and H. Yuan, "Real-time facial expression recognition using deep convolutional neural network," in *Proc. IEEE Int. Conf. Mechatronics Autom. (ICMA)*, Aug. 2019, pp. 1536–1541, doi: [10.1109/ICMA.2019.8816322](https://doi.org/10.1109/ICMA.2019.8816322).
- [55] S. M. Bah and F. Ming, "An improved face recognition algorithm and its application in attendance management system," *Array*, vol. 5, Mar. 2020, Art. no. 100014, doi: [10.1016/j.array.2019.100014](https://doi.org/10.1016/j.array.2019.100014).
- [56] M. Hayat, M. Bennamoun, and A. A. El-Sallam, "An RGB-D based image set classification for robust face recognition from Kinect data," *Neurocomputing*, vol. 171, pp. 889–900, Jan. 2016, doi: [10.1016/j.neucom.2015.07.027](https://doi.org/10.1016/j.neucom.2015.07.027).
- [57] M. Yuan, "An analysis model of sports human body based on computer vision tracking technology," *DEStech Trans. Social Sci., Educ. Human Sci.*, vol. 2017, pp. 220–224, Sep. 2017, doi: [10.12783/dtssehs/ssme2017/12960](https://doi.org/10.12783/dtssehs/ssme2017/12960).
- [58] N. Shirbhate and K. Talele, "Human body language understanding for action detection using geometric features," in *Proc. 2nd Int. Conf. Contemp. Comput. Informat. (IC3I)*, Dec. 2016, pp. 603–607, doi: [10.1109/IC3I.2016.7918034](https://doi.org/10.1109/IC3I.2016.7918034).
- [59] C. C. dos Santos, J. L. A. Samatelo, and R. F. Vassallo, "Dynamic gesture recognition by using CNNs and star RGB: A temporal information condensation," *Neurocomputing*, vol. 400, pp. 238–254, Aug. 2020, doi: [10.1016/j.neucom.2020.03.038](https://doi.org/10.1016/j.neucom.2020.03.038).
- [60] Z. Cao, X. Xu, B. Hu, M. Zhou, and Q. Li, "Real-time gesture recognition based on feature recalibration network with multi-scale information," *Neurocomputing*, vol. 347, pp. 119–130, Jun. 2019, doi: [10.1016/j.neucom.2019.03.019](https://doi.org/10.1016/j.neucom.2019.03.019).
- [61] M. A. Alsalem, A. A. Zaidan, B. B. Zaidan, M. Hashim, O. S. Albahri, A. S. Albahri, A. Hadi, and K. I. Mohammed, "Systematic review of an automated multiclass detection and classification system for acute leukaemia in terms of evaluation and benchmarking, open challenges, issues and methodological aspects," *J. Med. Syst.*, vol. 42, no. 11, p. 204, Sep. 2018, doi: [10.1007/s10916-018-1064-9](https://doi.org/10.1007/s10916-018-1064-9).

- [62] H. Ahmed and H. M. La, "Education-robotics symbiosis: An evaluation of challenges and proposed recommendations," in *Proc. IEEE Integr. STEM Educ. Conf. (ISEC)*, Mar. 2019, pp. 222–229, doi: [10.1109/ISECCon.2019.8881995](https://doi.org/10.1109/ISECCon.2019.8881995).
- [63] Z. Li, J. Li, H. Zhang, and Z. Li, "Study on the application of robot-assisted education in industrial design," in *Proc. Int. Joint Conf. Inf., Media Eng. (IJCIME)*, Dec. 2019, pp. 51–54, doi: [10.1109/IJCIME49369.2019.00-90](https://doi.org/10.1109/IJCIME49369.2019.00-90).
- [64] S. Chootongchai, N. Songkram, and K. Piromsopa, "Dimensions of robotic education quality: Teachers' perspectives as teaching assistants in Thai elementary schools," *Educ. Inf. Technol.*, vol. 26, no. 2, pp. 1387–1407, Mar. 2021, doi: [10.1007/s10639-019-10041-1](https://doi.org/10.1007/s10639-019-10041-1).
- [65] C.-S. Lee, M.-H. Wang, T.-X. Huang, L.-C. Chen, Y.-C. Huang, S.-C. Yang, C.-H. Tseng, P.-H. Hung, and N. Kubota, "Ontology-based fuzzy markup language agent for Student and robot co-learning," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, Jul. 2018, pp. 1–8.
- [66] T. Salim, S. R. U. N. Jafri, M. Ahmed, A. Jamal, S. Malik, and R. Uddin, "Humanoid robot for educational and assistive applications," in *Proc. 4th Int. Conf. Emerg. Trends Eng., Sci. Technol. (ICEEST)*, Dec. 2019, pp. 1–6, doi: [10.1109/ICEEST48626.2019.8981684](https://doi.org/10.1109/ICEEST48626.2019.8981684).
- [67] T. Morita, N. Takahashi, M. Kosuda, and T. Yamaguchi, "A teaching assistant robot design tool based on knowledge chunks reuse," in *Proc. IEEE 43rd Annu. Comput. Softw. Appl. Conf. (COMPSAC)*, Jul. 2019, pp. 68–73, doi: [10.1109/COMPSAC.2019.10185](https://doi.org/10.1109/COMPSAC.2019.10185).
- [68] A. Di Nuovo and T. Jay, "Development of numerical cognition in children and artificial systems: A review of the current knowledge and proposals for multi-disciplinary research," *Cognit. Comput. Syst.*, vol. 1, no. 1, pp. 2–11, Mar. 2019, doi: [10.1049/ccs.2018.0004](https://doi.org/10.1049/ccs.2018.0004).
- [69] E. Vrochidou, A. Najoua, C. Lytridis, M. Salomidis, V. Ferelis, and G. A. Papakostas, "Social robot NAO as a self-regulating didactic mediator: A case study of teaching/Learning numeracy," in *Proc. 26th Int. Conf. Softw., Telecommun. Comput. Netw. (SoftCOM)*, Sep. 2018, pp. 93–98, doi: [10.23919/SOFTCOM.2018.8555764](https://doi.org/10.23919/SOFTCOM.2018.8555764).
- [70] J. G. G. da Silva, D. J. Kavanagh, T. Belpaeme, L. Taylor, K. Beeson, and J. Andrade, "Experiences of a motivational interview delivered by a robot: Qualitative study," *J. Med. Internet Res.*, vol. 20, no. 5, p. e116, May 2018, doi: [10.2196/jmir.7737](https://doi.org/10.2196/jmir.7737).
- [71] M. Hussin, M. S. Said, N. M. Norowi, N. A. Husin, and M. R. Mustaffa, "Authentic assessment for affective domain through student participant in community services," *Asia-Pacific J. Inf. Technol. Multimed.*, vol. 10, no. 1, pp. 52–62, 2021.
- [72] V. Robles-Bykbaev, C. Contreras-Alvarado, L. Matute-Sanchez, E. Lema-Condo, Y. Robles-Bykbaev, and P. Suquilanda-Cuesta, "An educational environment based on stuffed toy robots, mobile apps, and expert systems to provide support in the early development of children," in *Proc. IEEE Colombian Conf. Commun. Comput. (COLCOM)*, Jun. 2019, pp. 1–6, doi: [10.1109/ColComCon.2019.8809113](https://doi.org/10.1109/ColComCon.2019.8809113).
- [73] M. J. Scott, S. Counsell, S. Lauria, S. Swift, A. Tucker, M. Shepperd, and G. Ghinea, "Enhancing practice and achievement in introductory programming with a robot Olympics," *IEEE Trans. Educ.*, vol. 58, no. 4, pp. 249–254, Nov. 2015, doi: [10.1109/TE.2014.2382567](https://doi.org/10.1109/TE.2014.2382567).
- [74] B. R. Schadenberg, M. A. Neerincx, F. Cnossen, and R. Looije, "Personalising game difficulty to keep children motivated to play with a social robot: A Bayesian approach," *Cognit. Syst. Res.*, vol. 43, pp. 222–231, Jun. 2017, doi: [10.1016/j.cogsys.2016.08.003](https://doi.org/10.1016/j.cogsys.2016.08.003).
- [75] D. Liao, T. Wu, and Y. Chen, "An interactive robot for fatigue detection in the learning process of children," in *Proc. 2nd Int. Conf. Adv. Robot. Mechatronics (ICARM)*, Aug. 2017, pp. 218–222, doi: [10.1109/ICARM.2017.8273163](https://doi.org/10.1109/ICARM.2017.8273163).
- [76] C. Edwards, A. Edwards, F. Albrehi, and P. Spence, "Interpersonal impressions of a social robot versus human in the context of performance evaluations," *Commun. Educ.*, vol. 70, no. 2, pp. 165–182, Apr. 2021, doi: [10.1080/03634523.2020.1802495](https://doi.org/10.1080/03634523.2020.1802495).
- [77] V. Bantia, Y. Maddahi, M. May, D. Blakley, Z. Chang, A. Gbur, C. Tu, and N. Sepehri, "Development of a graphical user interface for a socially interactive robot: A case study evaluation," in *Proc. IEEE 7th Annu. Inf. Technol., Electron. Mobile Commun. Conf. (IEMCON)*, Oct. 2016, pp. 1–8, doi: [10.1109/IEMCON.2016.7746294](https://doi.org/10.1109/IEMCON.2016.7746294).
- [78] S. Benus, R. Sabo, and M. Trnka, "Teaching L₁ and L₂ communication skills with a robotic head," in *Proc. 17th Int. Conf. Emerg. Elearn. Technol. Appl. (ICETA)*, Nov. 2019, pp. 69–75, doi: [10.1109/ICETA48886.2019.9040019](https://doi.org/10.1109/ICETA48886.2019.9040019).
- [79] N. Lazzeri, D. Mazzei, L. Cominelli, A. Cisternino, and D. D. Rossi, "Designing the mind of a social robot," *Appl. Sci.*, vol. 8, no. 2, p. 302, Feb. 2018, doi: [10.3390/app8020302](https://doi.org/10.3390/app8020302).
- [80] M. Romeo, A. Cangelosi, and R. Jones, "Developing a deep learning agent for HRI: Dataset collection and training," in *Proc. 27th IEEE Int. Symp. Robot Human Interact. Commun. (RO-MAN)*, Aug. 2018, pp. 1150–1155, doi: [10.1109/ROMAN.2018.8525512](https://doi.org/10.1109/ROMAN.2018.8525512).
- [81] A. Moualla, A. Karaouzene, S. Boucenna, D. Vidal, and P. Gaussier, "Readability of the gaze and expressions of a robot museum visitor: Impact of the low level sensory-motor control," in *Proc. 26th IEEE Int. Symp. Robot Hum. Interact. Commun. (RO-MAN)*, Aug. 2017, pp. 712–719, doi: [10.1109/ROMAN.2017.8172381](https://doi.org/10.1109/ROMAN.2017.8172381).
- [82] M. Aniketh and J. Majumdar, "Humanoid robotic head teaching a child with autism," in *Proc. 3rd Int. Conf. Circuits, Control, Commun. Comput. (IAC)*, Oct. 2018, pp. 1–7, doi: [10.1109/CIMCA.2018.8739603](https://doi.org/10.1109/CIMCA.2018.8739603).
- [83] O. Rudovic, J. Lee, M. Dai, B. Schuller, and R. W. Picard, "Personalized machine learning for robot perception of affect and engagement in autism therapy," *Sci. Robot.*, vol. 3, no. 19, Jun. 2018, Art. no. eaao6760, doi: [10.1126/scirobotics.aao6760](https://doi.org/10.1126/scirobotics.aao6760).
- [84] M. Leo, M. D. Coco, P. Carcagni, C. Distanto, M. Bernava, G. Pioggia, and G. Palestra, "Automatic emotion recognition in robot-children interaction for ASD treatment," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 537–545, doi: [10.1109/ICCVW.2015.76](https://doi.org/10.1109/ICCVW.2015.76).
- [85] A. Amanatiadis, V. G. Kaburlasos, C. Dardani, and S. A. Chatzichristofis, "Interactive social robots in special education," in *Proc. IEEE 7th Int. Conf. Consum. Electron. Berlin (ICCE-Berlin)*, Sep. 2017, pp. 126–129, doi: [10.1109/ICCE-Berlin.2017.8210609](https://doi.org/10.1109/ICCE-Berlin.2017.8210609).
- [86] A. A. Ramírez-Duque, A. Frizera-Neto, and T. F. Bastos, "Robot-assisted autism spectrum disorder diagnostic based on artificial reasoning," *J. Intell. Robot. Syst.*, vol. 96, no. 2, pp. 267–281, Nov. 2019, doi: [10.1007/s10846-018-00975-y](https://doi.org/10.1007/s10846-018-00975-y).
- [87] G. Nie, Z. Zheng, J. Johnson, A. R. Swanson, A. S. Weitlauf, Z. E. Warren, and N. Sarkar, "Predicting response to joint attention performance in human-human interaction based on human-robot interaction for young children with autism spectrum disorder," in *Proc. 27th IEEE Int. Symp. Robot Hum. Interact. Commun. (RO-MAN)*, Aug. 2018, pp. 1069–1074, doi: [10.1109/ROMAN.2018.8525634](https://doi.org/10.1109/ROMAN.2018.8525634).
- [88] A. D. Nuovo, S. Varrasi, A. Lucas, D. Conti, J. McNamara, and A. Soranzo, "Assessment of cognitive skills via human-robot interaction and cloud computing," *J. Bionic Eng.*, vol. 16, no. 3, pp. 526–539, May 2019, doi: [10.1007/s42235-019-0043-2](https://doi.org/10.1007/s42235-019-0043-2).
- [89] F. D. Duchetto, P. Baxter, and M. Hanheide, "Are you still with me? Continuous engagement assessment from a Robot's point of view," *Frontiers Robot. AI*, vol. 7, p. 116, Sep. 2020, doi: [10.3389/frobt.2020.00116](https://doi.org/10.3389/frobt.2020.00116).
- [90] L. Jeanpierre, A.-I. Mouaddib, L. Locchi, M. T. Lazaro, A. Pennisi, H. Sahli, E. Erdem, E. Demirel, and V. Patoglu, "COACHES: An assistance multi-robot system in public areas," in *Proc. Eur. Conf. Mobile Robots (ECMR)*, Sep. 2017, pp. 1–6, doi: [10.1109/ECMR.2017.8098710](https://doi.org/10.1109/ECMR.2017.8098710).
- [91] F. Bertacchini, E. Bilotta, and P. Pantano, "Shopping with a robotic companion," *Comput. Hum. Behav.*, vol. 77, pp. 382–395, Dec. 2017, doi: [10.1016/j.chb.2017.02.064](https://doi.org/10.1016/j.chb.2017.02.064).
- [92] Y. Hu, K. Sirlantzis, G. Howells, N. Ragot, and P. Rodríguez, "An online background subtraction algorithm deployed on a NAO humanoid robot based monitoring system," *Robot. Auto. Syst.*, vol. 85, pp. 37–47, Nov. 2016, doi: [10.1016/j.robot.2016.08.013](https://doi.org/10.1016/j.robot.2016.08.013).
- [93] K. Nakamura, T. Morita, and T. Yamaguchi, "A user-centric platform PRINTEPS to develop integrated intelligent applications and application to robot teahouse," *Proc. Comput. Sci.*, vol. 112, pp. 2309–2318, Jan. 2017, doi: [10.1016/j.procs.2017.08.266](https://doi.org/10.1016/j.procs.2017.08.266).
- [94] J. Espinosa-Aranda, N. Vallez, J. Rico-Saavedra, J. Parra-Patino, G. Bueno, M. Sorci, D. Moloney, D. Pena, and O. Deniz, "Smart doll: Emotion recognition using embedded deep learning," *Symmetry*, vol. 10, no. 9, p. 387, Sep. 2018, doi: [10.3390/sym10090387](https://doi.org/10.3390/sym10090387).
- [95] D. Mehta, M. Siddiqui, and A. Javaid, "Facial emotion recognition: A survey and real-world user experiences in mixed reality," *Sensors*, vol. 18, no. 2, p. 416, Feb. 2018, doi: [10.3390/s18020416](https://doi.org/10.3390/s18020416).
- [96] A. D. Lee, A. F. D. Costa, A. Davis, D. L. Linvill, and L. F. Hodges, "Virtualized speech practice for the college classroom," in *Proc. IEEE Conf. Virtual Reality 3D User Interface Abstr. Workshops (VRW)*, Mar. 2020, pp. 133–137, doi: [10.1109/VRW50115.2020.00028](https://doi.org/10.1109/VRW50115.2020.00028).

- [97] F. Patrona, A. Chatzitofis, D. Zarpalas, and P. Daras, "Motion analysis: Action detection, recognition and evaluation based on motion capture data," *Pattern Recognit.*, vol. 76, pp. 612–622, Apr. 2018, doi: 10.1016/j.patcog.2017.12.007.
- [98] P. Muneesawang, N. M. Khan, M. Kyan, R. B. Elder, N. Dong, G. Sun, H. Li, L. Zhong, and L. Guan, "A machine intelligence approach to virtual ballet training," *IEEE MultimediaMag.*, vol. 22, no. 4, pp. 80–92, Oct. 2015, doi: 10.1109/MMUL.2015.73.
- [99] Z. Zhao, F. Han, and X. Ma, "A live storytelling virtual reality system with programmable cartoon-style emotion embodiment," in *Proc. IEEE Int. Conf. Artif. Intell. Virtual Reality (AIVR)*, Dec. 2019, pp. 102–109, doi: 10.1109/AIVR46125.2019.00024.
- [100] V. G. A. Hakim, S.-H. Yang, T.-H. Tsai, W.-H. Lo, J.-H. Wang, T.-C. Hsu, and G.-D. Chen, "Interactive robot as classroom learning host to enhance audience participation in digital learning theater," in *Proc. IEEE 20th Int. Conf. Adv. Learn. Technol. (ICALT)*, Jul. 2020, pp. 95–97, doi: 10.1109/ICALT49669.2020.00036.
- [101] W. Wang, G. Athanasopoulos, G. Patsis, V. Enescu, and H. Sahli, "Real-time emotion recognition from natural bodily expressions in child-robot interaction," in *Proc. Eur. Conf. Comput. Vis.*, vol. 8927, 2015, pp. 424–435, doi: 10.1007/978-3-319-16199-0_30.
- [102] I.-J. Ding and Y.-J. Chang, "HMM with improved feature extraction-based feature parameters for identity recognition of gesture command operators by using a sensed Kinect-data stream," *Neurocomputing*, vol. 262, pp. 108–119, Nov. 2017, doi: 10.1016/j.neucom.2016.11.089.
- [103] M. Irsan, R. Hassan, and M. C. Lam, "The process of using face detection through convolutional neural network," in *Proc. Int. Conf. Bus. Anal. Technol. Secur. (ICBATS)*, 2022, pp. 1–5, doi: 10.1109/ICBATS54253.2022.9759092.



RAFEEF FAUZI NAJIM ALSHAMMARI received the B.Sc. degree in computer science from the University of Kerbala, Iraq, in 2006, and the M.Sc. degree in information technology from the University of Nottingham, U.K., in 2015. She is currently pursuing the Ph.D. degree with Universiti Kebangsaan Malaysia, Malaysia. Her research interests include artificial intelligent (AI), robotics, computer vision, multiple-criteria decision-making, and augmented reality (AR).



HASLINA ARSHAD received the B.Sc. degree in computer science from the University of Bridgeport, USA, the M.Sc. degree in IT for manufacture from Coventry University, Coventry, U.K., and the Ph.D. degree in manufacturing systems (virtual systems) from University Putra Malaysia. She had been working as an Analyst Programmer and a Systems Analyst at IBM, before she joined Universiti Kebangsaan Malaysia as a Lecturer. She is currently a Professor and the Director of the Institute of IR4.0, Universiti Kebangsaan Malaysia (UKM). Her research interests include augmented reality and virtual reality.



ABDUL HADI ABD RAHMAN received the Bachelor of Electrical Engineering degree in (computer) from UTM, in 2006, the Master of Computer Science degree from UPM, in 2009, and the Doctoral degree from UTM, in 2016. He had an internship at Tokai University during his Ph.D. study. He is currently with the Center of Artificial Intelligence Technology, Universiti Kebangsaan Malaysia (UKM). His research interests include robotic, artificial intelligence, computer vision, and mobile apps.



O. S. ALBAHRI received the B.Sc. degree in computer science from Al-Turath University, Baghdad, Iraq, in 2011, the M.Sc. degree in computer science and communication from the Arts, Sciences and Technology University in Lebanon, Lebanon, in 2014, and the Doctor of Philosophy (Ph.D.) degree from Universiti Pendidikan Sultan Idris (UPSI), Tanjong Malim, Malaysia, in 2019. His research interests include artificial intelligence, multi-criteria decision-making, multi-attribute decision-making, information, and networks security.

...