IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 17 May 2023, accepted 30 May 2023, date of publication 12 June 2023, date of current version 25 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3285398

RESEARCH ARTICLE

Emotion Recognition for Affective Human Digital Twin by Means of Virtual Reality Enabling Technologies

KAHINA AMARA[1], OUSSAMA KERDJIDJ[1,2], AND NAEEM RAMZAN[3], (Senior Member, IEEE)
[1]Centre of Development of Advanced Techniques, Algiers 16081, Algeria
[2]College of Engineering and Information Technology, University of Dubai, Dubai, United Arab Emirates
[3]School of Engineering and Computing, University of the West of Scotland, PA1 2BE Paisley, Scotland, U.K.

Corresponding author: Naeem Ramzan (naeem.ramzan@uws.ac.uk)

This work was supported in part by the SAFE_RH Project under Grant ERASMUS+ CBHE-619483-EPP-1-2020-1-UK-EPPKA2-CBHE.

ABSTRACT Digital Twin is the seamless data integration between a physical and virtual machine in either direction. Emotion recognition in healthcare is becoming increasingly important due to recent developments in Machine Learning methods. However, it may face technical problems such as limited datasets, occlusion and lighting issues, identifying key features, incorrect emotion classification, high implementation costs, head posture, and a person's cultural background. This paper proposes a novel approach based on facial expression and body movement recognition for emotion recognition. It uses three devices (Kinect 1, Kinect 2, and RGB HD camera) to construct a new bi-modal database containing 17 participants' performances of six emotional states. Two mono-modal classifiers have been developed to obtain sufficient state information based on facial expression and body motion analysis. The system's performance is assessed using three algorithms: Bagged Trees, k-Nearest Neighbors (k-NN), and Support Vector Machine (Linear and Cubic). The acquired findings demonstrate the excellent performance of the suggested method and the effectiveness of the proposed features, particularly the combination of 3D distance and 3D angle, in characterising and identifying emotions. Results obtained using Kinect 2 marginally surpass those with Kinect 1. Comparing 2D RGB and RGB-D data reveals that the depth information significantly raises the recognition rate. RGB-D features can be used to represent emotions, but there are discrepancies between RGB and RG-D data.

INDEX TERMS Body movement, classification, emotion recognition, facial expression, geometrical feature, healthcare, human digital twin, RGB, RGB-D, virtual reality.

I. INTRODUCTION

The digital twin is a virtual representation of a real-world asset that uses data and simulators for real-time prediction, monitoring, and control. Recent advancements in various technologies such as computational pipelines, artificial intelligence, and big data cybernetics have made the potential of digital twins closer to reality and have increased their impact on society [1]. The ability of a digital twin of a person to recognise and comprehend emotional states or emotions is referred to as "emotion recognition in a digital twin". The authors of a comprehensive study, [2], on human-computer interaction and digital twins provide recommendations for further research. Digital twins can be used in healthcare to monitor patients' emotional states for better care and in customer service to provide tailored services. An end-to-end framework that integrates emotion recognition with a digital twin setup has been proposed in [3] to examine and test projected results before they develop into life-threatening illnesses. Machine learning and deep learning models can be trained using a person's emotional state data obtained from sensors. However, the accuracy of these models can

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

be influenced by various variables such as lighting, head pose, and cultural background. Emotion recognition in virtual reality (VR) enhances realism and immersion and can modify content and actions to match the user's emotional state. There are several techniques for recognising emotions, such as voice, body language, and physiological signs.

- Analysis of the user's facial expressions using computer vision algorithms allows for detecting emotions, including happiness, sadness, rage, surprise, fear, etc.
- Voice analysis: To identify emotions like happiness, sorrow, anger, and fear, speech recognition and natural language processing techniques are used to analyse the user's voice.
- Body language analysis: In order to identify emotions like happiness, sadness, anger, and fear, computer vision techniques are used to examine the user's posture, gestures, and movements.
- Assessments of the user's physiological signals, such as heart rate, skin conductance, and blood pressure, are used in physiological signal analysis to identify emotions like happiness, sorrow, rage, and fear.

Human Digital Twin collects data from wearable sensors to monitor weight, blood pressure, pulse, heart rate, respiration, blood glucose, exercise volume, and emotional changes. Literature explores methods to identify emotional states using machine learning, deep learning techniques, and physiological databases [4]. Affective human digital twin (AHDT) is a recently developed idea that uses biometrics and AI to create a digital representation of a person's emotions and behaviours. The ESG (Environmental, Social, and Governance) objectives are to reduce the carbon footprint, power data centres with renewable energy sources, and reduce trash produced during production and disposal. Privacy and data protection, social impact assessment, trust and transparency, and ethical technology use are among the potential social factors to take into consideration. These social dimensions highlight how important it is to think about the broader societal implications and duties related to the creation and use of human digital twins, making sure that they promote social fairness, empowerment, and well-being. Governance should ensure that AHDT technology is created and used in an ethical and responsible manner. To map the digital twin and the real world, a digital model must be made that replicates the physical characteristics, actions, and functionality of the real-world system or object. The digital twin is updated in real-time to reflect changes in physical systems, allowing engineers, designers, and operators to test scenarios and improve performance. Various applications and implementations of digital twin technology in different domains can be found. In [5], the authors explore the use of digital twins in healthcare and provide a paradigm of digitally twinned everything as a healthcare service. This is consistent with the digital twining as a service paradigm and the Internet of Things as a Service idea supporting Industry 4.0 [6]. [7] offers a systematic evaluation of the literature on DT technology and its implementation difficulties in critical engineering domains. Reference [8] describes a digital twin designed to replicate the human response to viral infections at different scales. Reference [9] provides a cooperative city DT idea for municipal crisis management, and [10] suggests using HDTs for elderly real-time monitoring, remote diagnosis, virtual surgery training, and health consulting. A system developed by [11] that uses web cameras to capture and interpret real-time images for emotion recognition in an emergency room. The system integrates with a digital twin setup to allow testing and examination of projected results, preventing life-threatening diseases and providing efficient care. CNNs have recently been used for facial emotion recognition [12], [13], [14].

Humans can understand, distinguish, and evaluate emotions, and our study aims to recognise emotions based on facial expressions and body movements [3], emotion identification has found fruitful ground for applications: humans-robots interaction [15], action tendency, and recently in Digital twin for health-care [10], [11].

Our work investigates people's emotions by recognising them accurately. Emotion recognition can improve the performance of VR systems, and machines are analysing people's expressions for emotion prediction. Emotions elicitation, detection, and modelling are important for understanding human emotional responses. One of the simplest models is one described in [16] and [17], it contains six emotions (fear, anger, disgust, sadness, happiness, and surprise). Other models are proposed in [18]. These models build a classification scheme and provide a prediction with regard to the user's emotional state [19]. The features used to build models can be of different natures. Many works investigated emotion recognition by using biological signals that include Electroencephalography (EEG) [19], an electrophysiological [20].

Facial and body expressions are effective ways of expressing emotion. They are the most expressive modalities for human emotion. In the literature, various studies addressed facial expression [3], [21], [22], [23], [24] and body movements and expressions [25], [26]. Multimodal emotion recognition has also attracted attention and has been widely examined [21], [27]. It tries to combine different modalities simultaneously to improve emotion recognition performance. Several works implemented such an approach and showed that the combination of the expression modalities expression improves significantly the emotion recognition system performance [3], [28]. In the same context, combinations of body, facial, speech, physiological, and text modalities have been surveyed and showed their better performance than the standard mono-modal approaches [29].

The current methods for identifying emotions through facial and bodily expressions have various drawbacks and restrictions. First, there is no openly accessible database. Additionally, using insignificant features to represent various emotions can lead to model failure, according to citation [30]. To solve a problem involving many classifications, the complexity of the suggested model and the complexity

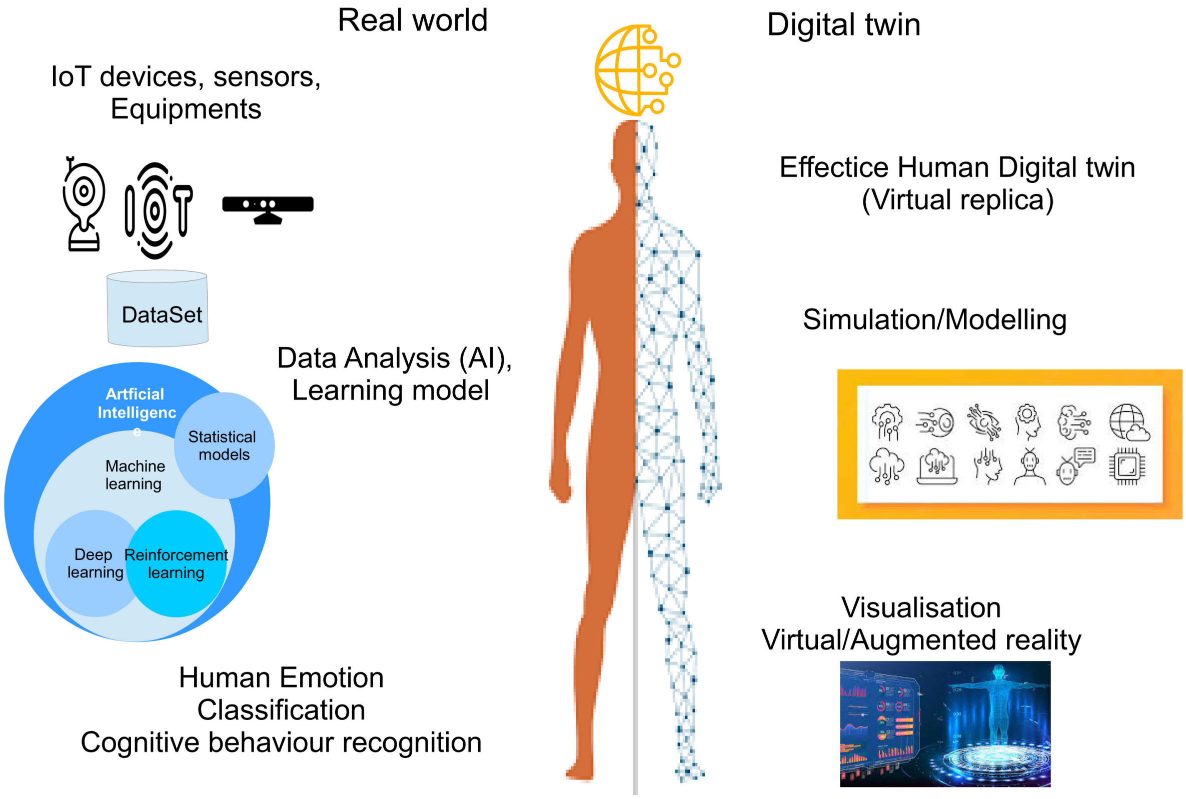
Real world
IoT devices, sensors, Equipments
DataSet
Artficial Intelligenc e
Statistical models
Machine learning
Deep learning
Reinforcement learning
Data Analysis (AI), Learning model
Human Emotion Classification Cognitive behaviour recognition
Digital twin
Effectice Human Digital twin (Virtual replica)
Simulation/Modelling
Visualisation Virtual/Augmented reality

FIGURE 1. The proposed framework for Affective human digital twin.

of the target emotions are crucial, especially as real-time applications are developed further. The fact that cultural influences on emotional expression result in varied personal styles of emotional performance is one of the significant challenges. Also worth mentioning are changes in the surroundings, such as variations in lighting and posture. In this study, we provide a system for emotion recognition based on body language and facial expressions to address these issues. The key contributions of the paper are:

- **Bi-modal RGB and RGB-D dataset:** We create a dataset, including the performance of 17 participants (9 males and 8 females). The participants are from more than ten different nationalities and have different skin tones. For the participant emotion elicitation, we used emotional images and videos. Unique features collected from Kinect sensors (version 1 and 2) and RGB HD camera were used to collect the data (table 1). The dataset was captured in controlled conditions of varying face appearance, body pose, and illumination. Our dataset is available at: https://zenodo.org/record/8015985.
- **Significant and consistent corporal and facial key points features selection for facial and corporal emotion recognition:** Important facial and skeletal features were chosen to represent the various emotional states. We reduced the number of features and used geometrical features, including a combination of 3D distance and 3D angle computed between each selected point for the data provided by Kinect 1 and Kinect 2 devices; further real-time implementation of the proposed system endorses this choice. RGB-D recordings and joint sequences, which are data similar to that from the Kinect, can be utilised to extract key features for gesture and face emotion recognition. The 3D body and facial features points are calculated using the 3D joint-oriented body skeleton, and the 3D facial points provided by Kinect SDK software. Feature selection is crucial for such a system; the multi-classification of emotions may fail because of the choice of non-suitable feature points.
- **Kinect 1 and Kinect 2 data comparison for emotion recognition:** In this study, we carried out a data performance comparison for emotion recognition using data provided by Kinect 1 and Kinect 2.
- **RGB and RGB-D data comparison:** The 2D RGB data provided by the RGB HD camera were tested against Kinect sensors RGB-D data. The use of depth data may improve emotion recognition performance.
- **Facial and corporal performance comparison for emotion recognition:** In this work and we provided a comparison between two modalities for emotion recognition: facial and corporal.
- **Comparison of the suggested method's performance with other cutting-edge techniques:** The method's effectiveness is evaluated in comparison to other cutting-edge techniques.

The remainder of this essay is structured as follows. The related work is discussed in section II. We describe the sug-

TABLE 1. Properties of emotional multimedia databases.

Database	Modalities	Sensory Data	Emotion Target	Application Target
[31]	Posture Gestures	Kinect 1 Kinect 2	Anger, Neutral, Happiness	Emotion recognition from gait
[25]	Posture Gestures	2 Kinect 2	Anger, Neutral, Happiness	Emotion recognition from gait
[3]	Bodily Expression Facial Expression	Kinect 2	Anger, Fear Happiness, Sadness, Surprise	Affective recognition in Serious Games Applications
[32]	Facial Expression Posture Gestures	2 HD cameras Kinect	Emotions, Mood	Emotion and mood recognition
[33]	Facial Expression Audio Signals, Physiological signals, Eye gaze	6 cameras were used for facial expression recording	Anxiety, Fear, Joy, Sadness, Disgust, Anger, Surprise, Amusement	Affect recognition
[28]	Bodily Expression Facial Expression	Kinect 1	Anger, Happiness Sadness,Disgust, and Surprise	Multimodal Affect recognition
[11]	Face Body Gestures	RGB camera	Disgust, Sadness, Happiness Surprise, Confusion, Aggressiveness	Emotion recognition for Digital twin application
Our dataset	**Bodily Expression** **Facial Expression**	**Kinect 1** **Kinect 2** **RGB HD camera**	**Anger, Happiness Sadness, Fear, Neutral, Surprise**	**Emotion recognition for AHDT**

gested strategy in more detail in section III. This encompasses every step of the data gathering process, experimental settings, feature extraction and selection, classification, and training. The results are shown, and the performance comparison is discussed in section IV. Finally, in Section V, we draw our findings and outline our next steps.

II. RELATED WORK

Across many contexts, digital twinning is currently a significant and developing trend. No wonder a digital twin, also known as a computational mega model, device shadow, mirrored system, avatar, or synchronised virtual prototype, plays a significant role in how we develop the modularity of multidisciplinary systems as well as how we design and operate cyber-physical intelligent systems [1]. Numerous studies on emotion identification utilising facial and bodily movements have been proposed, as we noted in section I. There were several feature extraction and selection techniques put forth. Using Kinect's facial and body data, [28] has looked into positional and temporal aspects. For the positional features, the authors developed a feature vector of the tracked points' coordinates, Euclidean distance, and angle. For the temporal features, they used a window of 10 frames for extracting temporal features across multiple frames. In [28], the authors concluded that fusing those features augmented the classification rates obtained from supervised learning. The facial expression-based emotion recognition line of the proposed work can be divided into three categories: image-based, video-based [34], and 3D surface-based [35]. Animation units (AUs) and point positions are extracted as features based on Kinect data for 3D surface-based emotion identification techniques. The hundreds of physiologically possible face expressions can be represented as combinations of 27 fundamental AUs, and a number of AU descriptors [36], which are used in an increasing number of studies on human facial behaviour [21]. The facial expression analysis is carried out using morphological parameters, such as the shapes of the facial regions (nose, eyes, mouth, face contour, etc.), as well as the locations of salient facial points (chin tip, corner of the eyebrows, lips, etc.). Mao et al. [36] proposed a real-time method for recognising facial expressions by combining AUs point features with Kinect feature position points. In [37], the facial expression recognition was based on motion features; they estimate the motion between two facial expressions. In [22], a facial feature extraction technique is described, Using the structured streaming skeleton (SSS) approach, the sequence of the normalised Euclidean distance between the monitored face points was examined. This later was first proposed for body emotion recognition [38]. For the evaluation, they constructed a database of fifteen actors for basic emotions (fear, happiness, surprise, sadness, disgust, anger, contempt further to neutral). Zhang et al. [23] addressed emotion recognition using 3D facial points provided by Kinect 2 to identify three emotions (sadness, happiness and neutral) using Kinect 2. To reduce the feature dimensionality PCA- Principal Component Analysis is used. A number of 100 key facial points were selected. Each point's temporal, frequency, and frequency-frequency domain properties are extracted. The authors [23] constructed the emotion recognition model using two datasets (male and female data), and they concluded that female data gives better results with a recognition rate of 80%.

People focus more on facial expressions than body gestures when trying to understand other people's feelings, but not all facial emotion recognition methods address which points are most relevant. This study is based on psychological studies on the perception of movements. [39], which reveal that specific facial and corporal key points are more relevant to characterising different emotions.

Body language constitutes a significant source of affective information. Recently, researchers focused on bodily emotion expression [25], [28], [33]. The study presented in [3] reported promising results with emotion recognition rates of 90%. The authors noted that a person's gait pattern and other aspects could alter depending on their emotional state. VR may be a beneficial tool for increasing physical activity and helping self-manage negative emotions. In [25], Li et al. discussed the gait recognition from information collected by the Kinect sensor. To categorise neutral, pleased, and furious affective states, the scientists collected gait data from six selected joints on the arms and legs in the frequency and temporal domains. They used PCA to lessen the redundant frequency features and found that the time-frequency features were more efficient. According to the authors, distinguishing neutral, anger, and happy states were not better due to the extracted features, which were not suitable. They claimed that using more convenient features would be more beneficial for depicting these two emotions (anger and happiness) in future work.

The proposed work uses 3D facial and body features from Kinect SDK software to improve the performance of the system, compared to traditional feature extraction methods. The first line of the work has focused on collecting a multimodal database using the sensors (Kinect 1, Kinect 2 and RGB HD camera). Emotion recognition classifiers have to be trained on databases comprising emotion manifestation, and related annotation [32]. Many databases were created for the study of affect recognition and used for different applications. Table 1 summarises some existing datasets; it shows the device used for data collection, the emotion target, and the application target. The experiment of data collection is discussed in the next section. The general framework is depicted in figure 1. We can describe three phases: data collection, feature extraction and selection, and classification, respectively.

III. PROPOSED APPROACH

To increase the performance of emotion recognition and accurately classify emotions, this study suggests a novel way based on facial expressions and body movements. The method is based on novel geometrical features and considers the most significant skeletal and facial landmarks.

A. RGB AND RGB-D BI-MODAL DATABASE

As depicted in figure 2, we used three devices, Kinect 1 Kinect 2 and RGB HD camera, to collect the used bi-modal dataset, including the corporal and facial recording of the emotional performance.

1) EXPERIMENT DESIGN

We ran an emotion priming experiment to gather accurate facial expression data using various emotional videos instead of performing on purpose [40] and images. The acquisition system recorded the facial expressions and corporal movements data of participants. For a fair comparison, the algorithms should perform under the same setting and conditions. For this reason, we adopted a pretty similar procedure as in [3] for a part of the experiment. This later was divided into two phases: emotion elicitation using emotional videos and emotion elicitation using emotional images. The acquisition system is shown in figure 2. We created a dataset with three devices, Kinect 1, Kinect 2 and RGB HD camera, including six emotions performances (fear, anger, sadness, happiness, surprise, and neutral). The neutral category was defined with no motion to distinguish the remaining emotion classes. The created dataset in this work contains 1581 RGB videos of 2s of length and 6000 Kinect RGB-D data files (skeleton joints and facial points from Kinect 1 and Kinect 2).

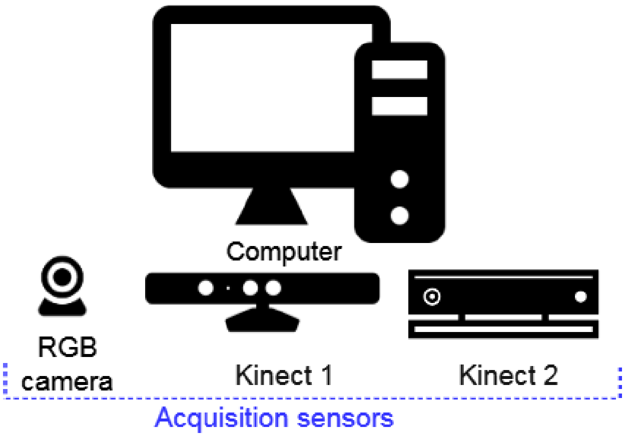

FIGURE 2. Acquisition system.

2) PARTICIPANTS

The participants were 17 students drawn from the School of Engineering and Computing at the UWS University of West of Scotland (9 men and 8 women). All of the volunteers ranged in age from 24 to 45. Participants come from ten different countries with a range of skin tones (figure 3).

3) STIMULI AND VIDEO/IMAGE SELECTION

The participants were first shown pictures illustrating the five basic emotions. They were asked to perform the same posture three times (each performance was 2 seconds in length) in front of the two Kinect sensors (Kinect 1 and Kinect 2) and RGB HD camera (figure 2). Then, the participants were shown emotional film clips.

We used the film clips studied in [41]; they were of different lengths ranging from 08 s to 55 s. We made the video segments as brief as possible to keep subjects from experiencing other emotions or becoming accustomed to the stimuli while keeping them long enough to capture the emotion. The order of the film clips was pseudo-randomly. Between each emotional film clip, a neutral video was projected.

4) MEASURES

After viewing each emotional video clip, each participant was asked to fill out a modified version of the post-film questionnaire [42]. Participants reported each emotion that they had experienced. Then, they were asked to perform the reported emotion three times. The form used in this work

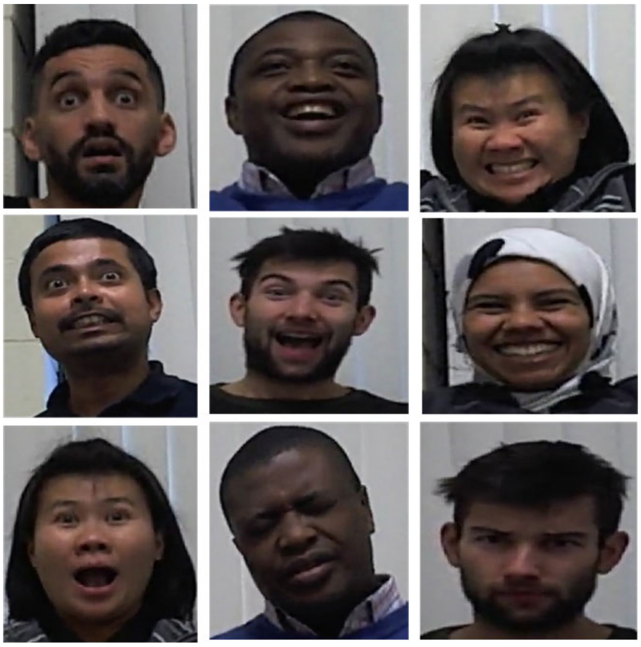

FIGURE 3. Some facial participants expression.

TABLE 2. Our dataset content summary.

Participants and modalities	
Participants	17 (9 males and 8 females).
Recorded Data	2D RGB Face videos, 3D facial points and 3D skeleton joints.
Emotional responses to videos	
Number of videos	20 videos
Self-report and Rating scale	Discrete scale of 0-10 for five emotional state (anger, fear, happiness, surprise, sadness).
Performance	The participants perform of the reported emotional state according to their personal style.
Emotional performance images	
Number of images	5 images (anger, fear, happiness, surprise, sadness).
Performance	The participants perform of the same projected emotional state on each image according to their personal style.

was expanded to include a rating scale from 0 (not at all) to 10 (extremely). Table 2 provides a summary of our bi-modal database's properties.

The body responses and recorded facial footage were manually segmented and entered into a database. We reduce our efforts in this work for the pre-processing since we are using the Microsoft Kinect toolkit. This device could provide a sequence of 3D Facial and Skeleton points.

Beyond the dataset collection description above, one more thing should be mentioned: we noticed that the expression style projected in videos influenced some performance of the emotional states. The participants were asked to express their different emotions according to their personal style. Despite this, some of them tended to imitate the film clip performance. Our emotion recognition has been performed on laboratory-controlled data.

TABLE 3. Extracted points.

Modality	Extracted Points
Face	Forhead, Eyes, Eyelids, Eyebrows, Nose, Jaw, Cheeks, Mouth, Chin
Body	Head, Spine base, Neck, Shoulders, Wrists, Elbows, Hips, Knees, and Ankles

TABLE 4. List of the facial expressions that our system can recognise, along with the corresponding modifications to the face.

Emotion	Changes that occur on the face
Anger	Brows lowered and drawn together Lips are either tightened/pressed firmly together with corners straight or down, Chin raised. Eyelids tightened, Upper lid raised.
Fear	Brows raised and drawn together. Lips are slightly tense or stretched and drawn back. Brows raised and drawn together. Inner brow raise, Outer brow raise. Upper lid raise. Upper eyelid is raised and lower eyelid is drawn up.
Happiness	Mouth may or may not be parted with teeth exposed or not cheeks are raised. Lower eyelid shows wrinkles below it, and may be raised but not tense. Corners of lips are drawn back and up. Lip corner pull, Lips parted.
Surprise	Brows raised, Eyelids opened. Jaw drops open or stretching of the mouth. Inner brow raised, Outer brow raised. Upper lid raised, Upper lid inner corner is raised. Chin raised, Corners of the lips are drawn downwards.
Sadness	Inner corners of eyebrows are drawn up. Chin raised, Corners of the lips are drawn downwards. Upper lid inner corner is raised.
Neutral	No change

B. FEATURE EXTRACTION/SELECTION

Six categories/emotional states were defined (sadness, fear, surprise, anger, happiness, and neutral). Some facial and corporal points were selected. Table 3 provides an overview of the points that have been tracked. The list of the bodily and facial emotion categories recognised by our system, as well as the changes that took place on the face and body, are shown in tables 4 and 4.

1) FACE FEATURE EXTRACTION

The face and skeleton tracking API found in the Microsoft Kinect Software Development SDK Toolkit was used to study facial and physical emotions. Kinect 1 and Kinect 2 sensors offer 121 and 1347 face key points, respectively. However, not all of these points are important to facial expressions, according to psychological studies [39]. The authors of [43] suggested that the areas, including the eyes, lips, and brows, are involved in facial emotion expression. To increase the recognition accuracy, we prioritised the essential facial areas surrounding the lips, brows, eyes, nose, cheeks, and chin above other important regions while choosing from the available points.

TABLE 5. List of the corporal emotions that our system can detect and the corresponding changes on the face.

Emotion	Changes that occur on the Body
Anger	Body extended Two hands up (close to head) Left hand moved up Right hand moved up Arms crossed
Fear	Left/Right hand touching the face Arms crossed Left/Right hand high up Body contracted Left/Right hand touching the neck Shoulders dropped Shoulder shrug
Happiness	Body extended Body extended, Two hands up (close to head) Two hands up Body extended, Two hands up
Surprise	Body extended Left/Right hand high up Two hands high up (towards the head) Shoulder shrug Palms up
Sadness	body contracted Left/Right hand up touching the cheek(s) Left/Right hand about to touch the head Left/Right hand moved down
Neutral	No change

For the 2D RGB sequences, we used the feature points offered by the open-source OpenFace [44], which offers 2D 68 facial landmarks. The 26 facial points we selected.

2) BODY FEATURE EXTRACTION

For the feature selection of the skeleton joints, Points describing limbs and trunk were sufficient to determine the correct emotion. For the skeleton joints, based on [40], and for simplification reasons, we choose corporal joints including ankles, shoulders, spine base, neck, elbows, wrists, knees, hips, and head. As a result, we could only monitor 15 of the 25 skeleton joints for Kinect 2 and 20 of the 20 skeleton joints for Kinect 1 that were available through the toolkit.

Finally, we select specific points that exhibit notable movement disparities to describe the slight variations in facial and physical expressions. We selected 26 vertices of facial points for the face modality and 15 skeleton joints for the body modality from among the points that could be tracked using the toolkit. Each point has 3D information (X, Y, Z).

The feature vectors, face feature vector and body feature vector (FV_{Face} and FV_{Body}), were determined by combining two geometrical characteristics: angle and the separation between each tracked point.

Given two facial points $P_n^{Face}(t)$ and $P_{n-1}^{Face}(t)$ with coordinates $(x_n(t), y_n(t), z_n(t))$ and $(x_{n-1}(t), y_{n-1}(t),$ $z_{n-1}(t))$ respectively at frame t, and two corporal points $P_n^{Body}(t)$ and $P_{n-1}^{Body}(t)$ with coordinates $(x_n(t), y_n(t),$ $z_n(t))$ and $(x_{n-1}(t), y_{n-1}(t), z_{n-1}(t))$ respectively at frame t. The position of the monitored points from one frame is the foundation for the feature vector. Feature vectors for face and body modalities are defined as follows (equation 1 and equation 2) respectively. It is a set of distance difference $D(t)$ and $\theta(t)$ which is the angle between each tracked facial or corporal tracked points [45], [46].

TABLE 6. Facial and corporal feature vectors length.

Modality	distance, angle	Vector length
Face	36,36	216
Body	11,11	66

$$\begin{aligned} FV_{Face} = {} & D_1^{Face}(P_0^{Face}(t), P_1^{Face}(t)), \dots, D_n^{Face}(P_{n-1}^{Face}(t), \\ & P_n^{Face}(t)), \theta_1^{Face}(P_0^{Face}(t), P_1^{Face}(t)), \dots, \\ & \theta_n^{Face}(P_{n-1}^{Face}(t), P_n^{Face}(t)) \end{aligned} \tag{1}$$

$$\begin{aligned} FV_{Body} = {} & D_1^{Body}(P_0^{Body}(t), P_1^{Body}(t)), \dots, D_n^{Body} \\ & (P_{n-1}^{Body}(t), P_n^{Body}(t)), \theta_1^{Body}(P_0^{Body}(t), \\ & P_1^{Body}(t)), \dots, \theta_n^{Body}(P_{n-1}^{Body}(t), P_n^{Body}(t)) \end{aligned} \tag{2}$$

The performance recording was 2 s in length. The RGB facial videos and skeleton joint sequences are synchronised and recorded at 30 frames per second. We chose key facial and corporal points, and we calculated the 3D distance and 3D angle. Table 6 summarises the feature vector length. For the face modality, we estimated 36 distances and 36 angles, and for the body modality, 11 distances and 11 angles.
For advanced real-time applications, we would include that optimising the parameters of the calculations has an enormous impact on the speed of meeting to a level, but it requires another step of cross-validation.

C. CLASSIFICATION, TRAINING AND EVALUATION

The distinct emotional states can be recognised using six computational models (neutral, fear, anger, sadness, happiness, and surprise). Different classifiers may offer varying classification accuracies for the same dataset. Bagged Trees, Linear SVM, Cubic SVM, and k-NN were some of the linear and non-linear classifiers we chose to assess in this study.
As one of the strategies utilised in statistics and machine learning, Support Vector Machine (SVM) displayed by Vapnik and Chervonenk could be a kind of capable factual learning strategy; it models the circumstance by making a feature space. We point to prepare a demonstration that categorises inconspicuous modern information into a specific category. Linear SVM tries to discover a linear combination of data that recognises or characterise distinctive classes. The facial and corporal affect information may not be directly distinct. For this reason, we inspected the non-linear SVM.

Leo Breiman formulated bagging, it is utilised as a strategy for moving forward the comes about of machine learning classification calculations. Its title was derived from the state "bootstrap aggregating". In arrange to decrease the change related to forecast and make strides the forecast handle, this

TABLE 7. The obtained emotion recognition results using Bagged trees.

Devices	Modalities	Accuracy%	Recall%	F1-score%	Precision%
Kinect 1	Face	97.44	92.94	92.77	92.74
	Body	97.51	93.27	93.11	93.12
Kinect 2	Face	97.58	93.43	93.18	93.10
	Body	98.46	95.39	95.52	95.68
RGB camera	Face	73.83	47.93	45.55	48.77

TABLE 8. The obtained emotion recognition results using *k*-NN.

Devices	Modalities	Accuracy%	Recall%	F1-score%	Precision%
Kinect 1	Face	97.09	91.94	91.74	91.67
	Body	97.40	92.87	92.79	92.81
Kinect 2	Face	97.40	92.86	92.65	92.49
	Body	98.06	94.40	94.44	94.52
RGB camera	Face	67.97	40.42	40.73	42.72

algorithm can be utilised. From the accessible information, numerous bagging samples are drawn. To each one of them, a few expectation strategies are connected and employing a voting process for classification; the results are combined. Typically to get the in general forecast, with the fluctuation being diminished due to the averaging. The bagging strategy gives extra information for preparing from the first dataset utilising combinations with redundancies to deliver multi-sets of the same cardinality/size as the initial information. By expanding the estimate of the preparing set, it cannot move forward the show prescient drive, but fair diminish the fluctuation, barely tuning the expectation to the expected result.

A non-parametric lazy learning algorithm is the *k*-NN algorithm. One of the most accessible categorisation algorithms, according to several experts. Despite its simplicity, *k*-NN can perform very well and give highly competitive results. The non-parametric algorithm means that it does not need to make any assumptions on the underlying data distribution. Most of the experimental data do not follow the typical theoretical assumptions (linearly separable, Gaussian mixtures, etc.), making these algorithms pretty relevant for the real world. The *k*-NN does not use the training data points to generalise, as it is a lazy algorithm.

We employed conventional leave-one-subject-out cross-validation to assess the performance of the models to develop a stable and trustworthy emotion model. We ran the trials for this study on a machine with an Intel$^{\delta}$ Xeon$^{\delta}$ CPU E3-1245 v3 3.40 GHz and 8 GB RAM. The Matlab classification techniques (Bagged Trees, *k*-NN, Linear SVM, and Cubic SVM) have been used in all tests.

IV. RESULTS AND DISCUSSION

Sixteen study participants' data were used for training purposes. We only used one participant's performance for testing purposes. The results of emotion recognition for face and body modalities are displayed in tables 7 and 8." When comparing the performance of several training methods, the Bagged Trees classifier outperforms them all. The outcomes produced by *k*-NN are comparable to those of the Bagged Trees method. The three algorithms' results, Bagged Trees, *k*-NN, and Support Vector Machine SVM, will be reported; (linear SVM and Cubic SVM).

The current results demonstrated the advantages of the suggested approach for emotion recognition with Kinect. The spatial characteristics help define and differentiate emotions.

One of the reasons for error in learning is noise. The bagged Trees with a single parameter its refinement is quite popular and can reduce the error. It gave the highest results for both modalities, face and body (table 7) with the data of Kinect 1 and Kinect 2; the results using OpenFace are also presented. The system achieved an accuracy rate of 98.46% for body modality and 97.58% for face modality using data collected by Kinect 2. The results using Kinect 1 are slightly close with accuracy rates of 97.44%, 97.51% for face and body modality, respectively, using Bagged Tress. Table 10 showcased the accuracy rates of each class separately.

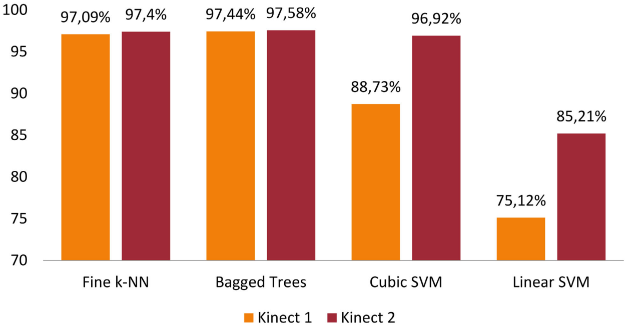

FIGURE 4. The outcomes for the face modality.

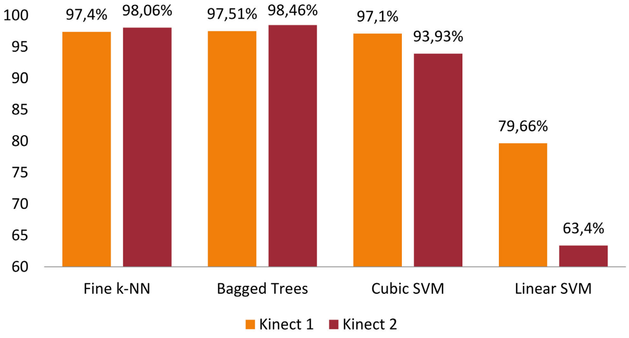

FIGURE 5. The outcomes for the body modality.

The *k*-NN algorithm with $k = 1$ achieved an accuracy rate of 97.09%, 97.40% for face modality (for Kinect 1 and Kinect 2, respectively). The emotion recognition for the body modality attained 97.40%, 98.06% of accuracy rate for Kinect 1 and Kinect 2, respectively. The *k*-NN, a simple as it is a lazy-learning algorithm, achieved high accuracy rates, close to the results obtained using Bagged trees. The benefits, of the Bagged trees, as a non-parametric learning algorithm over the other classification existing algorithms, are robust to the noisy training dataset, fast and effective for the large training dataset. Table 6 and table 7, we conclude that the RGB data showed the lowest performance compared to the RGB-D data collected by Kinect devices. Unlike k-NN, SVM requires training. It maps the features into a higher dimensional feature space. Consequently, SVM finds a linear hyperplane with a maximal margin to separate the different classes in this higher-dimensional space. The Linear-SVM performed poorly compared with the two previous classification algorithms; it gave the lowest accuracy rates for both modalities. The SVM poor performance is due to the feature vectors used for face and body modality, which are large, which degraded the performance of the SVM classifiers (Linear SVM, Cubic SVM).

In general, when the number of features is disproportionately significant, there is a possibility that the data are linearly separable in the original space. As a result, it is not necessary to map the data into a higher-dimensional space. However, non-linear SVM offers the opportunity to translate data that is linearly non-separable in a low dimensional space into a very high dimensional space for enhanced linear separability. The main distinction between non-linear and linear classifiers is that the non-linear classifiers might respond to high-level feature conjunctions differently from how they respond to individual features. Because our data structure is not linearly separable, the non-linear algorithms (Bagged Trees, k-NN, and Cubic SVM) perform better in the relative lower-dimensional space. In these circumstances, non-linear SVM performed better, and linear SVM had the lowest accuracy rate. This is so that a linear hyperplane may be identified easily in feature space, which is a higher-dimensional space where non-linear kernels change (map) the input data (input space). There is a comparison of various algorithms in figure 4 and figure 5.

The present results suggested that bodily emotion recognition slightly outperformed facial emotion recognition. Using Bagged Trees, it showed the highest classification rates compared to facial emotion recognition with 97.51% and 98.46% (for Kinect 1 and Kinect 2, respectively). That is can be explained by the body feature points, which tend to move more than the facial feature points in terms of distance and number of variations such as angles.
Unlike the results obtained in [31], the results of emotion recognition using Kinect 2 seem better than those using Kinect 1. The acquisition conditions of the database, such as the control of illumination intensity, may influence and explain this difference in the results. We recorded our data streams using an RGB HD camera and Kinect sensors under carefully regulated lighting and face appearance scenarios. The only requirement is that the participants need to stay in front of the acquisition system at appropriate distances and ensure their head yaw is less than 45°, in which condition their faces can be tracked successfully.

The comparison between 2D RGB data and RGB-D data is depicted in figure 6; compared to the RGB HD camera, the facial feature points offered by Kinect sensors produced more significant results. The RGB-D images provide essential geometrical features which yield accurate emotion distinction. The RGB 2D images were not robust enough for emotional face recognition, which are considered 3D objects. The use of depth information improved the recognition rate significantly.

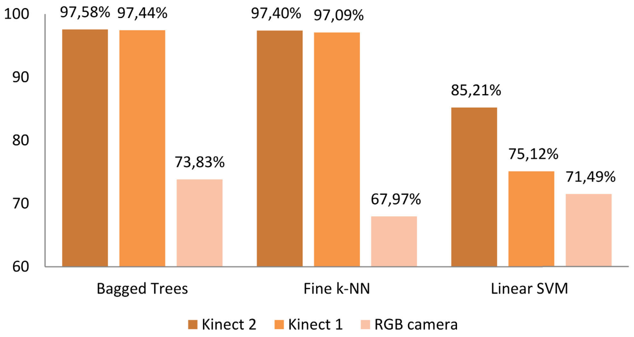

FIGURE 6. RGB data and RGB-D data performance comparison for face modality.

The existing facial and corporal emotion recognition systems could identify different emotional states through relatively facial and corporal expressions. The authors of [3] constructed their dataset containing the performance of 15 subjects of five primary emotional states. The best accuracy levels reached were 93% and 90% for face and body mono modality, respectively, using a deep learning network classifier. However, they got a recognition rate of 98.3% for bi-modal fusion. We are considering using the multimodal affective recognition method in future work. We expect improvement in recognition rates based on the study presented in [3], which investigated the multimodal fusion. In [28], the authors used temporal features across multiple frames; they reported an improvement in their results compared to the positional features. They said the results for three classes (happiness, anger, sadness). Even though we are using the positional features in our study, we got the highest recognition rates compared to [28]. Table 9 shows the performance comparison between the proposed approach for emotion recognition in this paper and state-of-the-art studies using the same devices and data.

Our results outperform the results reported in, which used the feature extraction method of structured streaming facial skeleton for facial emotion recognition. The results reported in [47] for eight classes with a dataset of five people are better than those reported in [22] with the recorded performance of 15 participants. In table 9, we expose the results of eight classes presented in [22] and [47]. Since we considered a problem of six classes in our work, we had to recalculate the accuracy of six classes again. The results of our study with a larger dataset (17 participants) outperform those reported in [22] and [47]. In [48], using the RGB images acquired by Kinect, the authors extracted geometric features of RGB images, and they achieved a recognition rate of 89.46%. To the best of our knowledge, we conducted fair comparisons between the proposed approach in this study and the state-of-the-art works by taking into account different dataset sizes and numbers of classes, as shown in table 9.

For monomodal emotion recognition, our results outperformed the existing works in emotion recognition with

TABLE 9. Performance evaluation in contrast to cutting-edge works. * Kinect 1, ** Kinect 2, ***RGB camera, ^{1} *k*-NN, ^{2} Bagged-Trees.

Works	Modality	Number of Emotional States	Accuracy %	Precision %	Recall%
[22] ^{**1}	Face	6 8	89.44 90.33	- -	- -
[36] *	Face	6 7	80.75 80.57	- -	- -
[3] **	Face Body	6	93 90	- -	- -
[28] *	Face Body	3	- -	77.33 78	55 52
[47] ^{**1}	Face	6 8	96.74 96.92	- -	- -
[48]*	Face	5	89.46	-	-
[12]***	Face	7	56.77	57	-
[13]***	Face	3	97.23	-	-
[14]	Face	-	95.64	-	-
[49]	Face	-	96.78	-	-
Proposed approach ^{**2}	Face Body	6	**97.58** **98.46**	**93.10** **95.68**	**93.43** **95.39**
Proposed approach ^{*2}	Face Body	6	**97.44** **97.51**	**92.74** **93.12**	**92.94** **93.27**

TABLE 10. Bagged trees accuracy performance.

Device	Modality	anger	fear	happiness	sadness	surprise	neutral
Kinect 1	Face	97.07	97.79	96.24	98.84	96.97	97.75
	Body	97.95	95.88	96.89	99.47	97.90	96.98
Kinect 2	Face	96.92	97.71	97.67	99.15	97.39	96.67
	Body	98.24	98.75	97.95	98.92	98.51	98.39
RGB camera	Face	67.62	77.31	64.73	86.92	75	71.39

recognition rates of 97.58% and 98.46% for face and body modality, respectively, using the Bagged Trees classifier. Besides, our study shows that the results using the Kinect 2 sensor are slightly better than those obtained using Kinect 1. Furthermore, we noticed that the body modality gives a higher recognition rate as compared to face modality results.

For comparison reasons and success measurement, we ran the experiment on the part of our face database. We took the recording of 09 participants, and using the leave-one-out subject cross-validation; we evaluated the system performance. The table below (table 11) addresses a comparison between existing datasets presented in [22] and [47] and our face dataset concerning the number of subjects and the computational complexity of feature extraction. The table also provides our method performance with two different dataset sizes. By increasing the size of the dataset, we obtained very close results, which means that our dataset is not noisy or does not contain unrepresentative training data.

As depicted in table 11, we presented the results using the Bagged Trees algorithm, and we compared our proposed method to two existent work [22], [47] which used the facial points provided by Kinect 2. The authors of these studies determined the Euclidean distance between paired facial skeleton model points. In this example, we also use the angle between each unique pair of tracked points to more precisely describe the facial and bodily motion and the distance between them. Based on the results, our strategy performs better than the approaches mentioned.

TABLE 11. Performance with regards to the complexity and the face dataset size.

Dataset	Subjects	complexity	Accuracy%
[47]	05	Euclidean distance	96.92
[22]	15	Euclidean distance	90.33
Proposed work	**17**	**3D Euclidean distance and 3D Angle**	**97.58**

V. CONCLUSION

By combining DT and human emotion recognition, a novel framework for affective human digital twins is proposed. This framework will make it easier to monitor, comprehend, and improve the capabilities of the physical entity. It will also provide continuous input to improve the quality of life and well-being for personalised healthcare to monitor a patient's health condition and early diagnosis of life-threatening diseases. The 3D positional features in this research, such as the 3D distance and the 3D angle between the monitored points, are offered as a new way for affective human digital twin emotion recognition via facial expressions and bodily movements. A new bi-modal dataset was introduced for the performance of facial and physical emotions. The six emotional states have been categorised using mono-modal classifiers (fear, anger, sadness, happiness, surprise, and neutral). Cross-validation using a leave-one-out subject was established for evaluating system performance. Due to the nature of the data, the non-linear algorithms produced consistent findings. All the classification methods consistently performed worse than the Bagged Trees and k-NN. Using Kinect 2, the system achieved accuracy rates of 97.58% for the face modality and 98.46% for the body modality. Using the Bagged Trees algorithm, Kinect 1 had the best accuracy at 97.44% for face modality and 97.51% for body modality. The findings show that Kinect 2 outperforms Kinect 1. Based on our findings, we conclude that the body modality produced consistent results. The results showed that the RGB-D data outperformed the RGB-2D data, which needed to be more reliable for emotional face recognition. Furthermore, the RGB-D data are more pertinent and provide crucial geometrical information. Our comparison of the dataset size, computational cost, and the number of classes revealed that the proposed methodology outperformed state-of-the-art classification accuracy. We hypothesise that in the future, emotion recognition and VR technology might incorporate capabilities for affective digital twin.

ACKNOWLEDGMENT

This work was supported in part by the SAFE_RH Project under Grant ERASMUS+ CBHE-619483-EPP-1-2020-1-UK-EPPKA2-CBHE.

REFERENCES

[1] A. Rasheed, O. San, and T. Kvamsdal, "Digital twin: Values, challenges and enablers from a modeling perspective," *IEEE Access*, vol. 8, pp. 21980–22012, 2020, doi: 10.1109/ACCESS.2020.2970143.

[2] B. R. Barricelli and D. Fogli, "Digital twins in human-computer interaction: A systematic review," *Int. J. Hum.-Comput. Interact.*, vol. 8, pp. 21980–22012, Sep. 2022, doi: 10.1080/10447318.2022.2118189.

[3] A. Psaltis, K. Kaza, K. Stefanidis, S. Thermos, K. C. Apostolakis, K. Dimitropoulos, and P. Daras, "Multimodal affective state recognition in serious games applications," in *Proc. IEEE Int. Conf. Imag. Syst. Techn. (IST)*, Oct. 2016, pp. 435–439.

[4] W. Shengli, "Is human digital twin possible?" *Comput. Methods Programs Biomed. Update*, vol. 1, 2021, Art. no. 100014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666990021000136

[5] H. Hassani, X. Huang, and S. MacFeely, "Impactful digital twin in the healthcare revolution," *Big Data Cogn. Comput.*, vol. 6, no. 3, p. 83, Aug. 2022, doi: 10.3390/bdcc6030083.

[6] C. Zhang, G. Zhou, J. Li, F. Chang, K. Ding, and D. Ma, "A multi-access edge computing enabled framework for the construction of a knowledge-sharing intelligent machine tool swarm in industry 4.0," *J. Manuf. Syst.*, vol. 66, pp. 56–70, Feb. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0278612522002060

[7] D. M. Botín-Sanabria, A.-S. Mihaita, R. E. Peimbert-García, M. A. Ramírez-Moreno, R. A. Ramírez-Mendoza, and J. D. J. Lozoya-Santos, "Digital twin technology challenges and applications: A comprehensive review," *Remote Sens.*, vol. 14, no. 6, p. 1335, Mar. 2022, doi: 10.3390/rs14061335.

[8] R. Laubenbacher, J. P. Sluka, and J. A. Glazier, "Using digital twins in viral infection," *Science*, vol. 371, no. 6534, pp. 1105–1106, Mar. 2021.

[9] J. Pang, Y. Huang, Z. Xie, J. Li, and Z. Cai, "Collaborative city digital twin for the COVID-19 pandemic: A federated learning solution," *Tsinghua Sci. Technol.*, vol. 26, no. 5, pp. 759–771, Oct. 2021.

[10] Y. Liu, L. Zhang, Y. Yang, L. Zhou, L. Ren, F. Wang, R. Liu, Z. Pang, and M. J. Deen, "A novel cloud-based framework for the elderly healthcare services using digital twin," *IEEE Access*, vol. 7, pp. 49088–49101, 2019.

[11] B. Subramanian, J. Kim, M. Maray, and A. Paul, "Digital twin model: A real-time emotion recognition system for personalized healthcare," *IEEE Access*, vol. 10, pp. 81155–81165, 2022.

[12] A. Santra, V. Rai, D. Das, and S. Kundu, "Facial expression recognition using convolutional neural network," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 10, no. 5, pp. 1081–1092, 2022.

[13] G. S. Monisha, G. S. Yogashree, R. Baghyalaksmi, and P. Haritha, "Enhanced automatic recognition of human emotions using machine learning techniques," *Proc. Comput. Sci.*, vol. 218, pp. 375–382, Jan. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050923000200

[14] X. Jin, Z. Lai, W. Sun, and Z. Jin, "Facial expression recognition based on depth fusion and discriminative association learning," *Neural Process. Lett.*, vol. 54, pp. 2025–2047, Jan. 2022.

[15] C. Zhang, G. Zhou, D. Ma, R. Wang, J. Xiao, and D. Zhao, "A deep learning-enabled human-cyber-physical fusion method towards human–robot collaborative assembly," *Robot. Comput.-Integr. Manuf.*, vol. 83, Oct. 2023, Art. no. 102571. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0736584523000479

[16] P. Ekman, "An argument for basic emotions," *Cognition Emotion*, vol. 6, nos. 3–4, pp. 169–200, May 1992. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/02699939208411068

[17] K. Amara, N. Ramzan, N. Zenati, O. Djekoune, C. Larbes, M. Guerroudji, and D. Aouam, "Towards emotion recognition in immersive virtual environments: A method for facial emotion recognition," in *ICCSA Conf. Comput. Sci. Complex Syst. Their Appl. 2021*, T. Marir, A. Bourouis, R. Benaboud, V. Gupta, and C. Gupta, Eds., vol. 2904. Oum El Bouagui, Algeria, May 2021, pp. 253–263.

[18] R. Plutchik, "The nature of emotions," *Amer. Scientist*, vol. 89, no. 4, pp. 344–350, Jul. /Aug. 2001.

[19] P. Arnau-González, M. Arevalillo-Herráez, and N. Ramzan, "Fusing highly dimensional energy and connectivity features to identify affective states from EEG signals," *Neurocomputing*, vol. 244, pp. 81–89, Jun. 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231217305222

[20] F. Mazza, M. P. Da Silva, and P. L. Callet, "Investigating electrophysiology for measuring emotions triggered by audio stimuli," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, Nov. 2013, pp. 1350–1364.

[21] H. F. García, M. A. Álvarez, and Á. A. Orozco, "Dynamic facial landmarking selection for emotion recognition using Gaussian processes," *J. Multimodal User Interfaces*, vol. 11, no. 4, pp. 327–340, Dec. 2017.

[22] N. Chanthaphan, K. Uchimura, T. Satonaka, and T. Makioka, "Novel facial feature extraction technique for facial emotion recognition system by using depth sensor," *Int. J. Innov. Comput. Inf. Control*, vol. 12, pp. 2067–2087, Dec. 2016.

[23] Z. Zhang, L. Cui, X. Liu, and T. Zhu, "Emotion detection using Kinect 3D facial points," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. (WI)*, Oct. 2016, pp. 407–410.

[24] N. Farajzadeh, G. Pan, and Z. Wu, "Facial expression recognition based on meta probability codes," *Pattern Anal. Appl.*, vol. 17, no. 4, pp. 763–781, Nov. 2014.

[25] B. Li, C. Zhu, S. Li, and T. Zhu, "Identifying emotions from non-contact gaits information based on Microsoft kinects," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 585–591, Oct. 2018.

[26] S. Piana, A. Staglianò, F. Odone, A. Verri, and A. Camurri, "Real-time automatic emotion recognition from body gestures," *CoRR*, vol. abs/1402.5047, pp. 1–7, Feb. 2014.

[27] L. Kessous, G. Castellano, and G. Caridakis, "Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis," *J. Multimodal User Interfaces*, vol. 3, nos. 1–2, pp. 33–48, Mar. 2010.

[28] A. S. Patwardhan and G. M. Knapp, "Multimodal affect recognition using Kinect," *CoRR*, vol. abs/1607.02652, pp. 1–9, Jul. 2016.

[29] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[30] H. Boubenna and D. Lee, "Feature selection for facial emotion recognition based on genetic algorithm," in *Proc. 12th Int. Conf. Natural Comput., Fuzzy Syst. Knowl. Discovery (ICNC-FSKD)*, Aug. 2016, pp. 511–517.

[31] S. Li, L. Cui, C. Zhu, B. Li, N. Zhao, and T. Zhu, "Emotion recognition using Kinect motion capture data of human gaits," *PeerJ*, vol. 4, p. e2364, Sep. 2016.

[32] C. Katsimerou, J. Albeda, A. Huldtgren, I. Heynderickx, and J. A. Redi, "Crowdsourcing empathetic intelligence: The case of the annotation of EMMA database for emotion and mood recognition," *ACM Trans. Intell. Syst. Technol.*, vol. 7, no. 4, pp. 51:1–51:27, May 2016. [Online]. Available: http://doi.acm.org/10.1145/2897369

[33] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Trans. Affect. Comput.*, vol. 3, no. 1, pp. 42–55, Jan. 2012, doi: 10.1109/T-AFFC.2011.25.

[34] S. E. Kahou, X. Bouthillier, P. Lamblin, C. C. Gülçehre, V. Michalski, K. R. Konda, S. Jean, P. Froumenty, Y. N. Dauphin, N. Boulanger-Lewandowski, R. C. Ferrari, M. Mirza, D. Warde-Farley, A. C. Courville, P. Vincent, R. Memisevic, C. J. Pal, and Y. Bengio, "EmoNets: Multimodal deep learning approaches for emotion recognition in video," *CoRR*, vol. abs/1503.01800, pp. 1–14, Mar. 2015.

[35] F. Barrett, E. Martin, M. Milonova, R. Gur, R. E. Gur, C. Kohler, and R. Verma, "Automated video-based facial expression analysis of neuropsychiatric disorders," *J. Neurosci. methods*, vol. 168, pp. 224–238, Feb. 2008.

[36] Q.-R. Mao, X.-Y. Pan, Y.-Z. Zhan, and X.-J. Shen, "Using Kinect for real-time emotion recognition via facial expressions," *Frontiers Inf. Technol. Electron. Eng.*, vol. 16, no. 4, pp. 272–282, Apr. 2015.

[37] H. da Cunha Santiago, I. R. Tsang, and G. D. C. Cavalcanti, "Facial expression recognition based on motion estimation," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 1617–1624.

[38] X. Zhao, X. Li, C. Pang, Q. Z. Sheng, S. Wang, and M. Ye, "Structured streaming skeleton—A new feature for online human gesture recognition," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 11, no. 1, pp. 22:1–22:18, Oct. 2014. [Online]. Available: http://doi.acm.org/10.1145/2648583

[39] M. A. Giese and T. Poggio, "Neural mechanisms for the recognition of biological movements," *Nature Rev. Neurosci.*, vol. 4, no. 3, pp. 179–192, Mar. 2003.

[40] N. F. Troje, "Decomposing biological motion: A framework for analysis and synthesis of human gait patterns," *J. Vis.*, vol. 2, no. 5, p. 2, Sep. 2002.

[41] C. A. Gabert-Quillen, E. E. Bartolini, B. T. Abravanel, and C. A. Sanislow, "Ratings for emotion film clips," *Behav. Res. Methods*, vol. 47, no. 3, pp. 773–787, Sep. 2015, doi: 10.3758/s13428-014-0500-0.

[42] J. Rottenberg, R. D. Ray, and J. J. Gross, *Emotion Elicitation Using Films*. London, U.K.: Oxford Univ. Press, 2007, ch. 2.

[43] B. Fasel and J. Luettin, "Automatic facial expression analysis: A survey," *Pattern Recognit.*, vol. 36, no. 1, pp. 259–275, Jan. 2016.

[44] T. Baltrušaitis, P. Robinson, and L. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–10.

[45] K. Amara, N. Ramzan, N. Achour, M. Belhocine, C. Larbes, and N. Zenati, "A new method for facial and corporal expression recognition," in *Proc. IEEE 16th Int. Conf Dependable, Autonomic Secure Comput., 16th Int. Conf Pervasive Intell. Comput., 4th Int. Conf Big Data Intell. Comput. Cyber Sci. Technol. Cong. (DASC/PiCom/DataCom/CyberSciTech)*, Aug. 2018, pp. 446–450.

[46] K. Amara, N. Ramzan, N. Achour, M. Belhocine, C. Larbas, and N. Zenati, "Emotion recognition via facial expressions," in *Proc. IEEE/ACS 15th Int. Conf. Comput. Syst. Appl. (AICCSA)*, Oct. 2018, pp. 1–6.

[47] N. Chanthaphan, K. Uchimura, T. Satonaka, and T. Makioka, "Facial emotion recognition based on facial motion stream generated by Kinect," in *Proc. 11th Int. Conf. Signal-Image Technol. Internet-Based Syst. (SITIS)*, Nov. 2015, pp. 117–124.

[48] L. Cai, H. Xu, Y. Yang, and J. Yu, "Robust facial expression recognition using RGB-D images and multichannel features," *Multimedia Tools Appl.*, vol. 78, pp. 28591–28607, May 2018.

[49] A. Y. Boumedine, S. Bentaieb, and A. Ouamri, "An improved KNN classifier for 3D face recognition based on SURF descriptors," *J. Appl. Secur. Res.*, pp. 1–19, Jul. 2022, doi: 10.1080/19361610.2022.2099688.

KAHINA AMARA received the master's degree in control and robotics and the Ph.D. degree from the University of Science and Technology Houari Boumediene (USTHB), Algiers, Algeria, in 2011 and 2018, respectively. She was a Consulting Engineer with the Centre of Development of Advanced Techniques (CDTA), IRVA Team. She is currently a full-time Researcher with CDTA. Her research interests include augmented and virtual reality, 3D interaction, affective computing, computer vision, emotion recognition, and healthcare.

OUSSAMA KERDJIDJ received the Ph.D. degree from the University of Laghoaut, Algeria, in 2019. His leading research interests include hardware and software implementation and artificial intelligence applied to healthcare applications.

NAEEM RAMZAN (Senior Member, IEEE) received the M.Sc. degree in telecommunication from the University of Brest, France, in 2004, and the Ph.D. degree in electronics engineering from the Queen Mary University of London, London, U.K., in 2008. He is currently a Full Professor in computing engineering and the Chair of the Affective and Human Computing for Smart Environment Research Centre and the Co-Lead of the Visual Communication Cluster, AVCN, University of the West of Scotland (UWS). He is currently focused on leading high quality interdisciplinary research and teaching in the areas of video processing, analysis and communication, multimedia search and retrieval, video quality evaluation, brain-inspired multi-modal cognitive technology, multimodal human–computer interfaces, DNA computing, fall detection, big data analytics, affective computing, the IoT/smart environments, natural multimodal human–computer interaction, and eHealth/connected health. Before joining UWS, he was a Senior Research Fellow and a Lecturer with the Queen Mary University of London, from 2008 to 2012. He has published more than 200 articles in journals, conferences, and book chapters, including standard contributions. He is a fellow of Royal Society of Edinburgh (FRSE), Senior Fellow of Higher Education Academy (SFHEA), the Co-Chair of the MPEG HEVC Verification (AHG5) Group, and a Voting Member of the British Standard Institution (BSI). In addition, he holds key roles with the Video Quality Expert Group (VQEG), such as the Co-Chair of the Ultra High Definition (UltraHD) Group, the Co-Chair of the Visually Lossless Quality Analysis (VLQA) Group, and the Co-Chair of the Psycho-Physiological Quality Assessment (PsyPhyQA). He was awarded the Staff Appreciation and Recognition Scheme (STARS) Award, in 2014 and 2016, for "Outstanding Research and Knowledge Exchange" (University of the West of Scotland) and awarded Contribution Reward Scheme, in 2011 and 2009, for "Outstanding Research and Teaching Activities" (Queen Mary University of London).