





Article

# Emotional Intelligence for the Decision-Making Process of Trajectories in Collaborative Robotics

Michele Gabrio Antonelli <sup>1,\*</sup>, Pierluigi Beomonte Zobel <sup>1</sup>, Costanzo Manes <sup>2</sup>, Enrico Mattei <sup>2</sup>  
and Nicola Stampone <sup>1</sup>

<sup>1</sup> Dipartimento di Ingegneria Industriale e dell'Informazione e di Economia, Università degli Studi dell'Aquila, P.le Pontieri Monteluco di Roio, 67100 L'Aquila, Italy; pierluigi.zobel@univaq.it (P.B.Z.); nicola.stampone@graduate.univaq.it (N.S.)

<sup>2</sup> Dipartimento di Ingegneria e Scienze dell'Informazione e Matematica (DISIM), Università degli Studi dell'Aquila, Via Vetoio, 67100 L'Aquila, Italy; costanzo.manes@univaq.it (C.M.); enrico.mattei@graduate.univaq.it (E.M.)

\* Correspondence: michelegabrioernesto.antonelli@univaq.it

**Abstract:** In collaborative robotics, to improve human–robot interaction (HRI), it is necessary to avoid accidental impacts. In this direction, several works reported how to modify the trajectories of collaborative robots (cobots), monitoring the operator's position in the cobot workspace by industrial safety devices, cameras, or wearable tracking devices. The detection of the emotional state of the operator could further prevent possible dangerous situations. This work aimed to increase the predictability of anomalous behavior on the part of human operators by the implementation of emotional intelligence (EI) that allows a cobot to detect the operator's Level of Attention (LoA), implicitly associated with the emotional state, and to decide the safest trajectory to complete a task. Consequently, the operator is induced to pay due attention, the safety rate of the HRI is improved, and the cobot downtime is reduced. The approach was based on a vision transformer (ViT) architecture trained and validated by the Level of Attention Dataset (LoAD), the ad hoc dataset created and developed on facial expressions and hand gestures. ViT was integrated into a digital twin of the Omron TM5-700 cobot, suitably developed within this project, and the effectiveness of the EI was tested on a pick-and-place task. Then, the proposed approach was experimentally validated with the physical cobot. The results of the simulation and experimentation showed that the goal of the work was achieved and the decision-making process can be successfully integrated into existing robot control strategies.

**Keywords:** emotional intelligence; vision transformer; collaborative robotics; digital twin; human–robot interaction



**Citation:** Antonelli, M.G.; Beomonte Zobel, P.; Manes, C.; Mattei, E.; Stampone, N. Emotional Intelligence for the Decision-Making Process of Trajectories in Collaborative Robotics. *Machines* **2024**, *12*, 113. <https://doi.org/10.3390/machines12020113>

Academic Editor: Raul D. S.

G. Campilho

Received: 15 November 2023

Revised: 31 January 2024

Accepted: 4 February 2024

Published: 7 February 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Collaborative robotics belongs to advanced robotics, one of the enabling technologies of Industry 4.0 [1], which aims to improve smart production. Collaboration expects the cobot and man to share the same workspace and exchange forces [2]. This paradigm could extend the application of cobots to domestic, healthcare [3], and medical applications [4]. The hybrid man–cobot system can exploit the high dexterity and versatility of the man and the robot's precision, rapidity, and capability to perform repetitive and often alienating tasks [5]. Therefore, this system allows greater productivity, better ergonomics, and significantly reduced process errors [6,7].

However, the resulting human–robot interaction (HRI) introduces a new class of risks due to accidental man–cobot contact. The Standard ISO/TS 15066 [8] defines the safety requirements of collaborative working modes regarding cobots and workstation design and control software development. Several researchers are involved in this field: a cobot was covered by a sensitized skin [9] or air chambers [10] to promptly detect accidental collisions;

other solutions exploit the intrinsic compliance of the adopted materials [11,12], combined with the pneumatics [13,14]. Collision avoidance is a most interesting topic [15–17]. Several works are aimed at visual-based or wearable-device-based commands of a cobot's trajectories or its speed: vision systems were proposed in [18,19]; a hybrid visual-and-wearable-based solution was proposed for identifying object positions and the operator's body parts [20]; wearable devices were adopted in [21,22]. The common aspect of these approaches is the ability to take advantage of signals based on human motions in a closed loop as feedback to command the robot. However, during a work shift, an operator does not always manage to maintain the due attention. The changes in the human operator's Level of Attention (LoA) could reduce efficiency and safety, in total disagreement with Industry 5.0, the new industry paradigm aimed at the human-centric environment to increase human operators' efficiency [23]. Emotions were demonstrated to influence a person's judgment, decision-making process, and attitude [24] or induce the avoidance or performance of some actions [25], creating a sort of ranking of the stimulus and influencing the behavior towards doing or not doing [26]. Some works demonstrate that emotional states induced by images or video [27,28] cause changes in reaction times and cardiac responses. In the industrial workplace, the alteration of the emotional state can lead to a change in reaction times, with an increase in the probability of sudden movements or the exit from the perimeter of the working safety area; in the HRI, these behaviors could stop the cobot, reducing the efficiency, and be dangerous, reducing the safety. For this reason, the operator's emotional state should be monitored. To improve efficiency and safety, like a man, a cobot should be equipped with Emotional Intelligence (EI): men with a high level of EI pay attention to emotions and use, understand, and manage them, bringing benefit to themselves and others [29]; a cobot should capture the operators' emotions, evaluating them and finally performing tasks in which the actions are regulated by the computed emotional states [30].

The implementation of EI typically requires the implementation of algorithms to detect emotional states from facial expressions, gaze, body language, voice, and physiological cues [31]. In the scientific literature, the approaches refer to Artificial Intelligence (AI) methodologies, ranging from classic machine learning to deep learning techniques, using single or multiple input types. In different areas, such as autism therapy [32,33], education [34], and automotive [35,36], EI based on vision systems has been applied. Several AI architectures have been developed for emotion recognition by facial expressions [37–40]. Heredia et al. [41] adopted multimodal approaches to support the recognition of facial expressions and voice by convolutional neural networks (CNNs) and text by the transformer architecture. A method based on the latter was adopted to elaborate the electrocardiogram (ECG) [42]. Chaudhari et al. [43] explained how to implement EI, in terms of recognition of facial expressions, by the ViT architecture. Transformer-based multiple-input systems are widespread and take input from video signals of facial expressions, audio signals of voice, or text signals [44]. Systems combining CNNs and transformers are also used to classify emotions using facial and body features [45].

In [46], the electroencephalography (EEG) signal of a worker busy in an industrial environment was acquired by a wearable device, processed, and used to modify the position of a cobot to help the operator in an assembly operation. Other studies adopted bio-signal monitoring, as reported in [47]. However, monitoring an operator with non-wearable sensors in such an environment would be more appropriate. Aside from the fact that the operator's movements can change the position of the sensors, this solution can create apprehension in the operator and convey to him the unsettling feeling of being controlled. To overcome this issue, some researchers proposed assessing an operator's stress state and mental effort with a stereo camera [47] in a HRI with a cobot. In [48], the operator emotions were identified by a CNN in addition to a recurrent neural network (RNN) architecture. Based on the detected emotion, a light color is adjusted to create a comfortable working environment.

The present work is focused on the implementation of EI in a cobot to be used mainly in the industrial field but potentially also in the domestic and healthcare fields. The work

reports the development and experimental validation of an EI system for the decision-making process of the trajectories of a cobot based on the real-time estimation of the LoA of an operator by a vision transformer (ViT). Such an architecture, developed in the Pytorch environment, is trained and validated by the ad hoc created dataset, the Level of Attention Dataset (LoAD), based on images of facial expressions or hand gestures to be referred to as three LoAs: “normal”, “medium”, and “low”. During an in-execution pick-and-place task of a cobot, the decision-making process computes a specific trajectory and velocity profile for each of the three estimated LoAs. The higher the deviation from the “normal” LoA, the farther the trajectories will be from the operator and the lower the execution speeds. Since the adopted cobot (Omron TM5-700) does not have a simulator, a novel Digital Twin (DT) of the cobot was designed and developed in the MATLAB environment to validate the proposed system. Based on the ViT recognition, the DT calculates, generates, and numerically simulates the specific trajectory and speed profile. Hence, the numerical validation of the ViT-DT system allowed the experimental implementation of the proposed system: in the MATLAB environment, ViT outputs are sent to a National Instrument (NI) USB6001 board, which generates corresponding analog input signals for the control unit of the cobot, whose parametric programming software calculates the associated trajectories. Experimental validation with ten subjects demonstrates that EI can improve HRIs and operator safety and induces the operator to pay due attention. Indeed, the modified trajectory and speed induced the subjects to pay the proper attention, avoiding accidental collisions, in compliance with a high safety rate. The last aspect should reasonably improve the acceptance of collaboration and the success of the robotic teammate as it occurs between human colleagues. Conversely, the stop of the task or acoustic or light alert signals could create discomfort for the operator and the acceptance of the cobot.

This work does not intend to develop a new mathematical theory or new algorithms for AI. It aims to create and experimentally validate a more collaborative, safe, efficient, and comfortable HRI by increasing the predictability of anomalous behavior of human operators and guiding the latter to have the proper attitude during a work shift.

The following items represent the originality of the work:

- A novel dataset, coming from existing ones but ad hoc manipulated, for a typical HRI in the industrial environment was created;
- The ViT architecture was trained and validated for the specific application in which hands and facial expressions are monitored. It was demonstrated that this kind of architecture, requiring millions of images, can be adopted in scenarios like the one reported in the manuscript;
- Since the outcome of the proposed system was a trajectory generated by emotional intelligence, whose value is predictable in the described application but for more complex situations could generate more complex trajectories, it was necessary that the trajectory was not harmful to the cobot. In the case of the adopted cobot, a simulator and a software integrator of the simulator with an external ViT architecture did not exist. For this reason, a DT of the adopted cobot for communication with the ViT architecture, simulation of the cobot behavior, and communication with the control unit of the real cobot was created;
- The aim of the work was not to create a comfortable working environment or monitor the operator’s emotions due to the HRI; on the contrary, it aimed to create a more collaborative, safe, efficient, and comfortable HRI when the cobot guides the operator to pay the due attention, as occurs among human colleagues.

The paper is organized as follows: the new dataset and the adopted ViT architectures are described in Section 2; in Section 3, the HRI application in which the EI is implemented for the decision-making process of trajectories is described; the DT and its validation are detailed in Section 4; Section 5 reports the experimental activity and the discussion of the achieved results.

## 2. The LoA Estimation

### 2.1. The Dataset

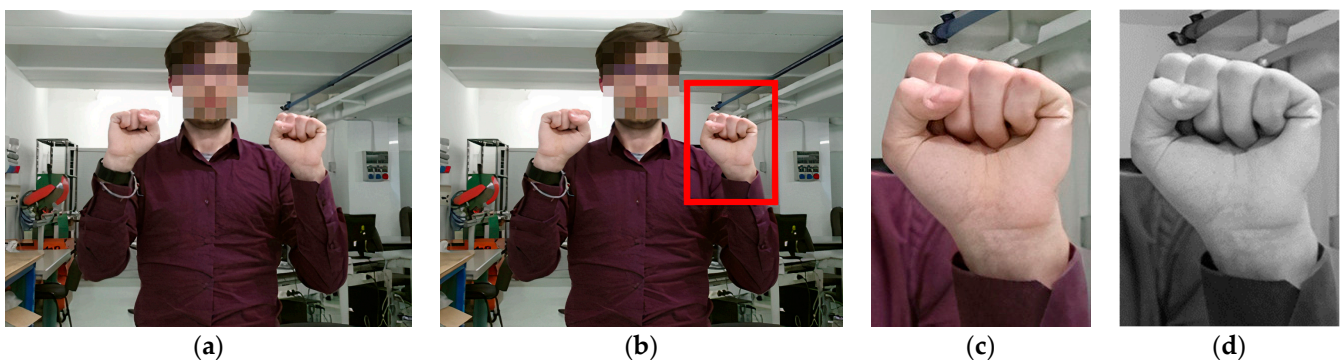
The LoAD was built with images of facial expressions and hand gestures, organized into two different sections. Each of them is made of three classes: Normal Level of Attention (NLoA), Medium Level of Attention (MLoA), and Low Level of Attention (LLoA). The LoAD arises from the union and manipulation of the Facial Expression Recognition 2013 (FER-2013) dataset [49] for facial expressions, HANDS [50] and the HAnd Gesture Recognition Image Dataset (HaGRID) [51] for hand gestures.

Of the original FER-2013 dataset (made of 35,888 grayscale images,  $48 \times 48$  pixel sized, and organized into seven different classes), only the *neutral*, *happy*, and *angry* classes were taken into consideration, associated with the classes NLoA, MLoA, and LLoA, respectively. The choice fell on these types of expressions because they represent the opposites of the emotional sphere, including a neutral expression, that most influences the operator's LoA. The expressions were associated with the LoAs according to the behavior that an operator could assume with the cobot, as typically occurs among human colleagues: a *neutral* expression was considered proper for the work shift when an operator should not be emotionally involved but focused only on the task; a *happy* one suggests that positive influences could distract the operator but do not compromise the teamwork; an *angry* one suggests that negative influences could compromise the teamwork. Images of FER-2013 were not manipulated.

Of the original HANDS dataset (made of 24,000 RGB images,  $960 \times 540$  pixel sized, and organized into 29 static hand-gesture classes), only the *Horiz* (*HBL*, *HFL*, *HBR*, *HFR*) and *Punch* (*VFR*, *VFL*) classes, associated with the MLoA and LLoA, respectively, were considered. The gestures were associated with the LoAs according to the behavior that an operator could assume when the cobot moved: *Horiz*, the open hands, suggests that the operator is not moving objects or could gesture so that the level of attention should not compromise the teamwork; *Punch* suggests that the operator could move an object or the operator could be affected by negative influences that could compromise the teamwork. The number of class images remained the same.

Of the original HaGRID dataset (made of 552,992 RGB images,  $1920 \times 1080$  pixel sized, and organized into 18 static hand-gesture classes and a *no-gesture* class), only the *no-gesture* class, associated with the NLoA class, was considered. According to the same criteria mentioned above, *no gesture* suggests that the operator has their hands out of the cobot workspace, focused on waiting for the end of the cobot motion. The number of images was reduced to balance the number of ones in the modified HANDS dataset.





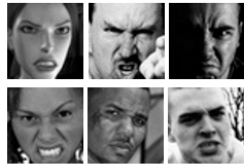

The original images of the hand-gesture datasets were cropped of insignificant information (background, clothes, faces); resized to  $48 \times 48$  pixels, satisfactory for the application; and converted to grayscale in the range of 0–255 for each pixel, as shown in Figure 1. The manipulation of the images did not change their quality.



**Figure 1.** Image processing of the hand gestures for the LoAD building: (a) original image; (b) cropped zone of image; (c) resized image; (d) grey-scale image. For clarity, the (c,d) images are the same size of the original one.

The chosen hand gestures must not be intended as a definition of new hand gestures for cobot control, which could contradict some of the ones adopted in the literature [19,20,50]; however, they are the relevant ones at the basis of the differentiation of the proper gestures for the specific application and the wrong ones.

Some images of the LoAD are reported in Figure 2. The amount of images in each class of the LoAD is reported in Table 1.

	FER 2013	HANDS manipulated	HaGRID manipulated
<b>NLoA</b>	 <i>neutral</i>	/	 <i>no gesture</i>
<b>MLoA</b>	 <i>happy</i>	 <i>Horiz</i>	/
<b>LLoA</b>	 <i>angry</i>	 <i>Punch</i>	/

**Figure 2.** Examples of facial expressions and hand gestures associated with the LoAD classes.

**Table 1.** The structure of the LoAD.

Section	Class	Number of Images
Facial expressions	NLoA	6000
	MloA	6000
	LloA	4440
Hand gestures	NloA	2640
	MloA	3564
	LloA	1800

## 2.2. The ViT Architectures

ViT was chosen for the implementation of EI because the image analysis carried out by ViT performs better than that carried out by the most adopted CNNs [52], as required in an industrial application. ViT architecture processes 2D images subdivided into a sequence of patches of fixed pixel dimensions. As pointed out in [52], the training process of ViTs requires millions of images. Since such a large set of images is not available, to demonstrate the feasibility of the present study, three architectures of pre-trained ViTs were considered: the Vision Transformer Base (ViT-B) [52], the ViT-B-based architectures Data-efficient Image Transformers Base (DeiT-B), and the Data-efficient Image Transformers Tiny (DeiT-Tiny), very effective when focused on a single task [53], as in the case of the present work, hence the choice of them. The architectures are available in the *huggingface/transformer library* by Pytorch [54].

For each image section, the training and validation of the ViT architectures were carried out with 80% and 20%, respectively, of the number of available images. Each architecture was trained on six, nine, and twelve epochs, i.e., how many times each architecture passed through the entire training dataset. The results achieved by each architecture were compared. The learning rate was set to  $10^{-6}$  and the *AdamW* optimizer [55] was applied in each phase. The choice of the best ViT architecture depended on the validation accuracy. Better results were achieved for the twelve epochs, reported in Table 2 and plotted in Figures 3–5.

**Table 2.** Results of 12-epoch training and validation of the ViT architectures.

Model Face	Training Loss	Validation Loss	Validation Accuracy
ViT-B	0.40	0.46	0.83
DeiT-B	0.37	0.50	0.81
DeiT-Tiny	0.54	0.57	0.76
Model Hand	Training Loss	Validation Loss	Validation Accuracy
ViT-B	0.07	0.09	0.99
DeiT-B	0.05	0.05	0.97
DeiT-Tiny	0.02	0.01	0.98

Results reveal that the size of the LoAD is appropriate for the present work. Regarding the accuracy of facial expressions, the ViT-B is the best architecture; the accuracy of the hand gestures is substantially similar for all architectures. Furthermore, the training and validation loss of the ViT-B are the lowest for facial expressions; for the hand gestures, on the contrary, they are the highest but reach values lower than 0.1 in correspondence with the twelfth epoch.

The best validation accuracy of the ViT-B architecture was compared to the validation accuracies of several models based on the FER-2013, HANDS, and HaGRID datasets found in the literature. As shown in Table 3, the results demonstrate that the ViT-B architecture based on the proposed dataset is more beneficial. The ViT-B (hereinafter called, simply, ViT) was chosen as the most appropriate.

**Table 3.** Comparison among model training results.

Model Face					
Dataset	Classes	ViT-B [43]	SSF-ViT-B [56]	A-CNN [57]	ViT-B (our)
FER-2013	neutral (NLoA)	0.61	0.73	0.80	0.80
	happy (MLoA)	0.77	0.89	0.69	0.91
	angry (LLoA)	0.63	0.69	0.53	0.78
Model Hand					
Dataset	Classes	R-FCN [50]	ViT-B [51]	DenseNet201 [58]	ViT-B (our)
HaGRID	no gesture (NLoA)	-	0.98	0.92	1
HANDS	Horiz (MLoA)	0.85	-	-	0.99
	Punch (LLoA)	1	-	-	1

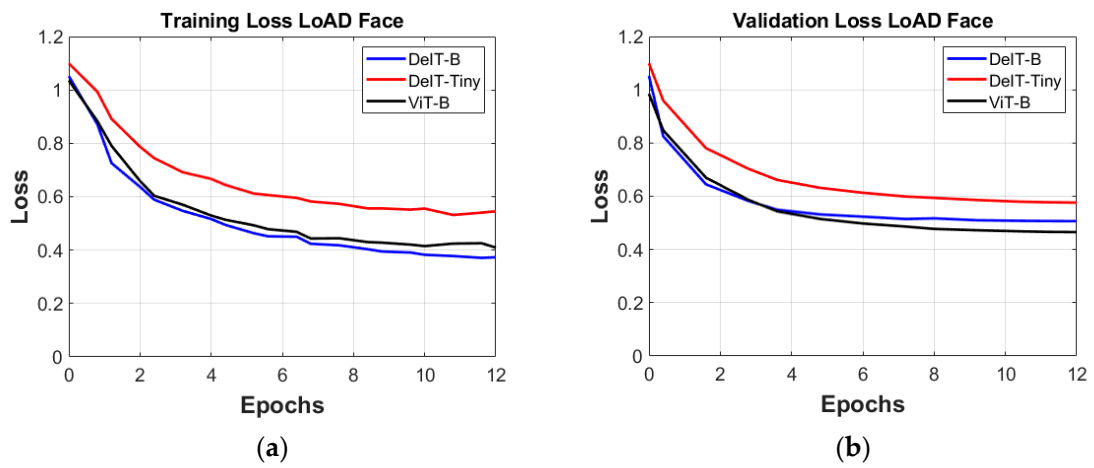


Figure 3. Comparisons among the (a) training loss and (b) validation loss of the ViT architecture applied to the LoAD for facial expressions.

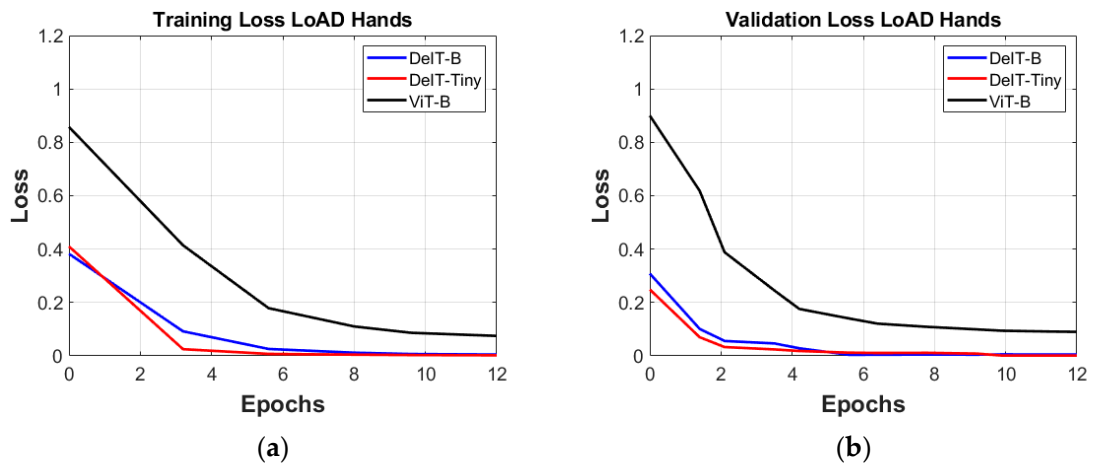


Figure 4. Comparisons among the (a) training loss and (b) validation loss of the ViT architecture applied to the LoAD for hand gestures.

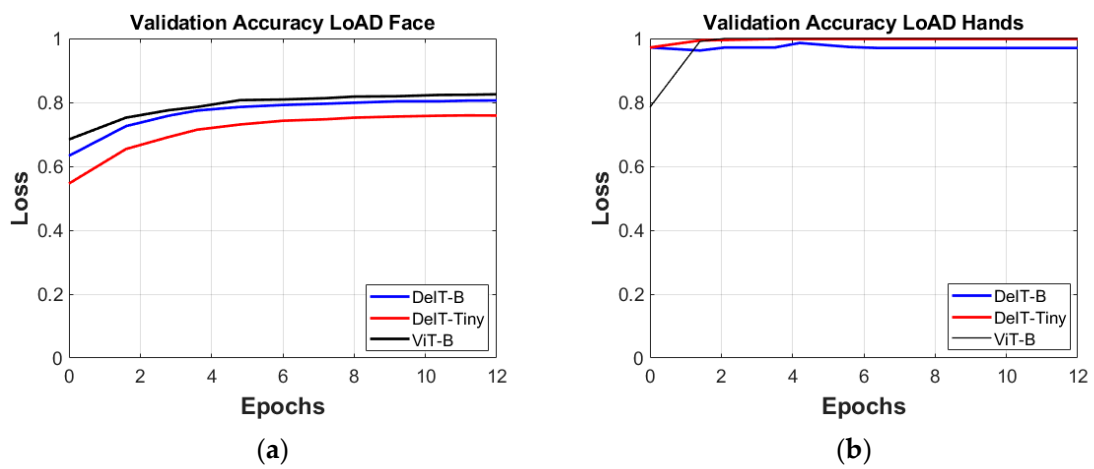
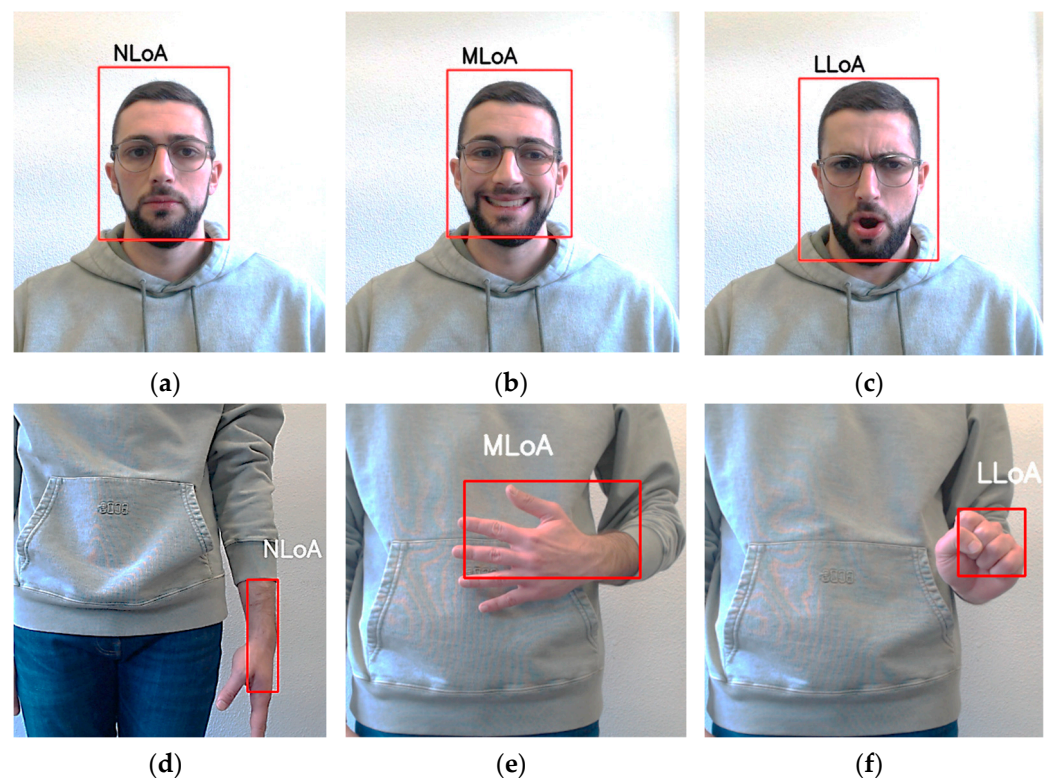


Figure 5. Validation accuracy of the ViT architectures applied to the LoAD for (a) facial expressions and (b) hand gestures.

### 2.3. The Recognition of Facial Expressions and Hand Gestures

To detect facial expressions and hand gestures, a RGB external webcam Logitech C920 HD Pro, with HD 720 p/30 fps, autofocus, and a Field of View (FoV) of 78°, is placed in front of the operator, at face or hand level, and at a distance of 1 m to ensure the image recording inside the FoV of the webcam and maintain the image in focus (Supplementary Materials). The webcam is equipped with auto light correction technology (Rightlight 2) since tests must be carried out with the lighting conditions provided by the lamps of a laboratory, influenced by any other external light disturbances. Such a condition is similar to a real industrial environment, where ceiling lamps provide lighting and can be modified by the sunlight, variable light during the day, and shadows of the operators and moving devices.

A code developed in Python acquires the images from the webcam and sends them to the ViT architecture, which estimates LoAs and generates graphical and numerical outputs. Each output is associated with the corresponding trajectory to be generated. The ViT estimates the LoA at a frequency of 5 Hz. The estimated LoA is the average value of the LoAs estimated in an (adjustable) time window  $t$ , set to 10 s. In this way, the instantaneous effect of external disturbances (sneezing, changes in lighting, gestures not directly related to the LoA) should be reduced. Examples of LoA recognition are reported in Figure 6. The resolution of the images was set to  $640 \times 480$  pixels. For clarity, no blur effects were applied to the face of the subject, whose consent was obtained.



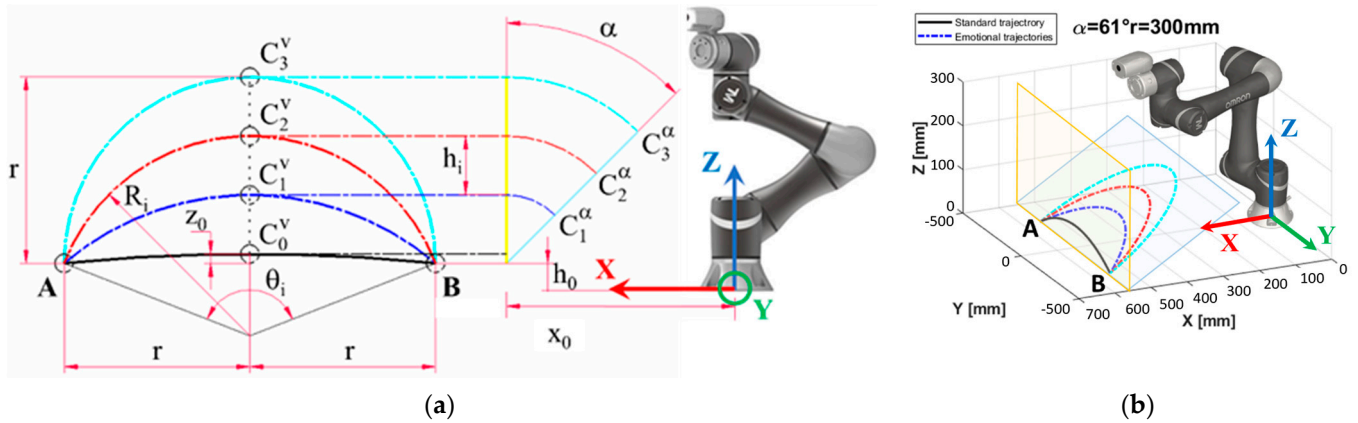
**Figure 6.** Examples of output of the ViT-B architectures applied to the LoAD for facial expression and hand gesture detection: (a) Face-NLoA, (b) Face-MLoA, (c) Face-LLoA, (d) Hand-NLoA, (e) Hand-MLoA, (f) Hand-LLoA.

### 3. The Emotional Intelligence Implementation: An Application

Emotional Intelligence aims to estimate the LoA of the operator in a HRI application and associate the proper trajectory and the speed of the task of a cobot. The greater the deviation from the NLoA, the farther the trajectories will be from the operator and the lower the execution speeds.



A pick-and-place task of a screw from and towards an operator's hand was considered. With reference to Figure 7, after the manual unscrewing, the cobot picks up the screw placed in the operator's hand in correspondence with point A; then, the cobot moves and releases the screw in the operator's hand in correspondence with point B; after the manual tightening of the screw, the cobot moves back to point A, following the trajectory in reverse, to start the task again.



**Figure 7.** Schematic of the standard and emotional trajectories in the pick-and-place task: (a) the parameters; (b) the spatial development of the trajectories. In (a), four trajectories are reported; although, only three were considered in the present study.

With respect to the robot base reference system XYZ, A and B points are placed by  $h_0$  height. The operator is positioned in front of the cobot on the opposite side of the trajectory to be executed. The standard trajectory is an arc of circumference placed in a vertical plane (the yellow one of Figure 7b), passing through points A, B, and  $C_0^V$ . The latter is placed at the half point of the AB segment, whose length is  $2r$ , by the height  $z_0$ .

For the sake of simplicity, the trajectories depending on the LoA, called emotional trajectories, are in the shape of arcs of circumferences passing through points A, B and  $C_i^\alpha$  placed in an  $\alpha$ -slope plane (the transparent blue one of Figure 7b) with respect to the vertical plane. Additionally,  $i$  is the integer number associated with one of the  $n$  estimated LoAs;  $C_i^\alpha$  are the rotated  $C_i^V$ , except  $C_0^V$ .

The slope  $\alpha$  depends on the  $r$  value to maximize the human–cobot distance and  $x_0$ , the distance between the base center of the robot and the vertical plane of the standard trajectory; the  $z_i$  heights, computed from the horizontal plane, including A and B, are achieved according to the number  $n$ : the greater the rate of negligence (due, for example, to stress, inattention, indolence, or drowsiness), the higher the  $z_i$  will be. Additionally,  $C_i^V$  are vertically equidistant by a distance  $h$ . The maximum value of  $z_i$ , corresponding to the most critical LoA, is set to  $r$ : in this case, the arc of circumference degenerates in a semicircle.

The choice for planning arc trajectories along the  $\alpha$ -slope plane provides for changing the trajectory from the standard one to summon the operator's attention and for moving away the distal part of the cobot from the operator to avoid any accidental collision. If the LoA is not detected correctly, the cobot performs the distal trajectory. As a first step, the computation of the emotional trajectories requires the evaluation of the  $\alpha$ -slope; then, the  $C_i^V$  coordinates, the radius  $R_i$ , and the angle at the center  $\theta_i$  of the corresponding arc of circumference; and finally, the length of the latter,  $L_i$ , and the coordinates of the set of 100 points, spaced by the angle  $\beta_i$ , belonging to the computed trajectory. All the points are rotated on the  $\alpha$ -slope plane and their coordinates are calculated in the XYZ base-centered system.

The set of equations to be adopted for the abovementioned computing is the following:

$$\alpha = \left[ \text{arctg} \left( \frac{x_0}{r} \right) \right] \quad (1)$$

$$h = \frac{r - z_0}{n} = h_i \quad (2)$$

$C_0^V$  and  $C_i^V$  points have the following coordinates in the robot base reference system:

$$C_0^V \{x_0, y_0, z_0 + h_0\} \quad (3)$$

$$C_i^V \{x_0, y_0, z_i\}, \quad z_i = z_0 + h_0 + i \cdot h_i, \quad i = 1, \dots, n - 1 \quad (4)$$

$$R_i = \frac{r^2 + z_i^2}{2z_i} \quad (5)$$

$$\theta_i = 2 \arcsin\left(\frac{r}{R_i}\right) \quad (6)$$

$$\beta_i = \frac{\theta_i}{100} \quad (7)$$

Each  $s$ -th point belonging to the  $i$ -th arc of circumference, placed in the vertical plane  $V$ , has the following coordinates in the robot base reference system:

$$P_s^{iV} \{x_s^{iV}, y_s^{iV}, z_s^{iV}\} = P_s^{iV} \left\{ x_0, R_i \cdot \sin\left(-\frac{\theta_i}{2} + (s-1) \cdot \beta_i\right), R_i \cdot \cos\left(-\frac{\theta_i}{2} + (s-1) \beta_i\right) - R_i \cdot \cos\left(\frac{\theta_i}{2}\right) + h_0 \right\}, \quad s = 1, \dots, 101 \quad (8)$$

Hence, each  $s$ -th point belonging to the  $i$ -th arc of circumference, placed in the  $\alpha$ -slope plane, has the following coordinates in the robot base reference system:

$$P_s^{i\alpha} \{x_s^{i\alpha}, y_s^{i\alpha}, z_s^{i\alpha}\} = P_s^{i\alpha} \left\{ x_0 - z_i \cdot \sin(\alpha), R_i \cdot \sin\left(-\frac{\theta_i}{2} + (s-1) \cdot \beta_i\right), \left( R_i \cdot \cos\left(-\frac{\theta_i}{2} + (s-1) \beta_i\right) - R_i \cdot \cos\left(\frac{\theta_i}{2}\right) + h_0 \right) \cdot \cos(\alpha) \right\} \quad (9)$$

Moreover,

$$L_i = 2\pi \cdot R_i \cdot \frac{\theta_i}{360^\circ} \quad (10)$$

Considering a symmetrical trapezoidal speed profile imposed on the Tool Center Point (TCP) of the cobot, the speed corresponding to the constant segment of the speed profile is defined as  $v_{\text{standard}}$  for the standard trajectory. For the other trajectories, the speed,  $v_i$ , is imposed as:

$$v_i = \frac{v_{\text{standard}}}{i+1}, \quad i = 0 \dots n - 1 \quad (11)$$

The acceleration/deceleration ramps are set in the form of the required time to reach the imposed speed value.

In the current case study,  $n$  is set to 3: normal (corresponding to the standard trajectory), medium, and low LoAs, corresponding to the  $i$  values 0, 1, and 2, respectively.

## 4. The Digital Twin of Cobot Omron TM5-700

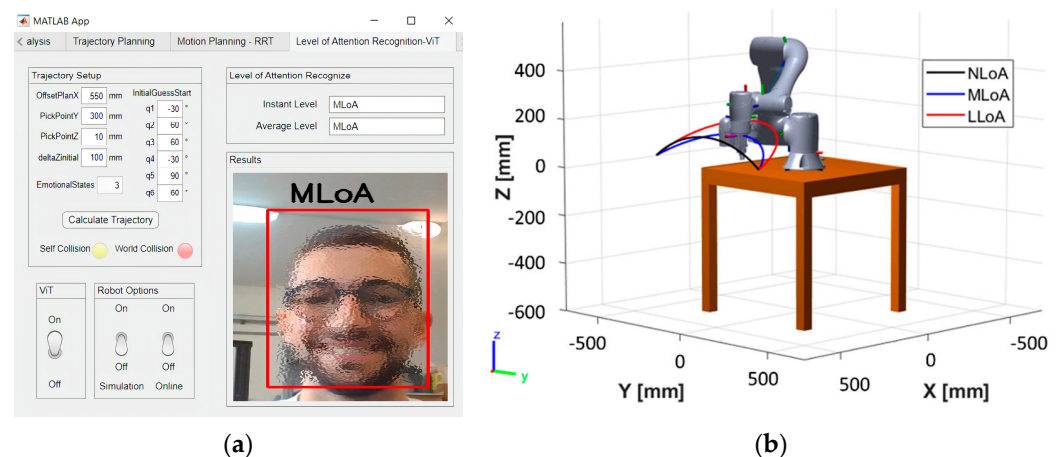
### 4.1. The Aim

Since the EI-based solution, a simulator of the cobot is necessary before testing the proposed system with a physical cobot in a real application. The novel DT was created because the Omron TM5-700 lacks software for simulating planned tasks. Moreover, a DT is able to exchange data with the ViT architecture and the control unit of the real cobot does not exist. The DT aims to acquire the numerical output generated by the ViT for simulating the trajectory associated with the estimated LoA and, in the application with the cobot, to manage an Interface board for creating the communication between the ViT and the Omron TM5-700 control unit.

### 4.2. The Structure

The DT is based on forward and inverse kinematics and trajectory generation algorithms available in the MATLAB libraries [59]. Moreover, it is equipped with a code developed for communication with the ViT by a socket between Pytorch and MATLAB.

Finally, it is provided with a developed code for the communication between MATLAB and the interface board. The construction of the DT started with creating the Unified Robotics Description Format (URDF) model of the cobot, which is not available in commercial libraries. Data about the cobot's kinematic, dynamic, and geometric parameters were assessed in its datasheet [60]. The URDF model includes the cobot support bench and the gripper for the application, used as an end-effector. Then, the previously mentioned algorithms were applied to the URDF by the MATLAB Robotic System Toolbox and a Graphical User Interface (GUI), customized for the application, was developed by the MATLAB App Designer. The *Level of Attention Recognition-ViT* page (Figure 8a) allows setting points A and B of the standard trajectory ( $x_0, y_A, z_A, h_0, r$ ) to define the initial configuration of the cobot, according to the joints coordinates; to set the number of possible LoAs (parameter  $n$ ); to visualize the LoA estimated by the ViT (parameters  $i$ ); and to calculate the corresponding trajectory. When the ViT is on, the DT calculates the trajectory and the speed by (1–11). The page shows the simulated cobot motion (Figure 8b) in a pop-up window. It highlights any robot collisions with itself and the surrounding environment (yellow and red round lines in Figure 8a). Moreover, the page allows the user to choose two work modes: offline, for the simulation of the calculated trajectories, and online, for cobot control, by managing the interface board that generates voltage analog output signals associated with the estimated LoA and, hence, to the trajectory to be executed.



**Figure 8.** The GUI of the DT: (a) the Level of Attention Recognition-ViT page; (b) the motion simulation pop-up window based on the URDF of the cobot.

#### 4.3. The Validation

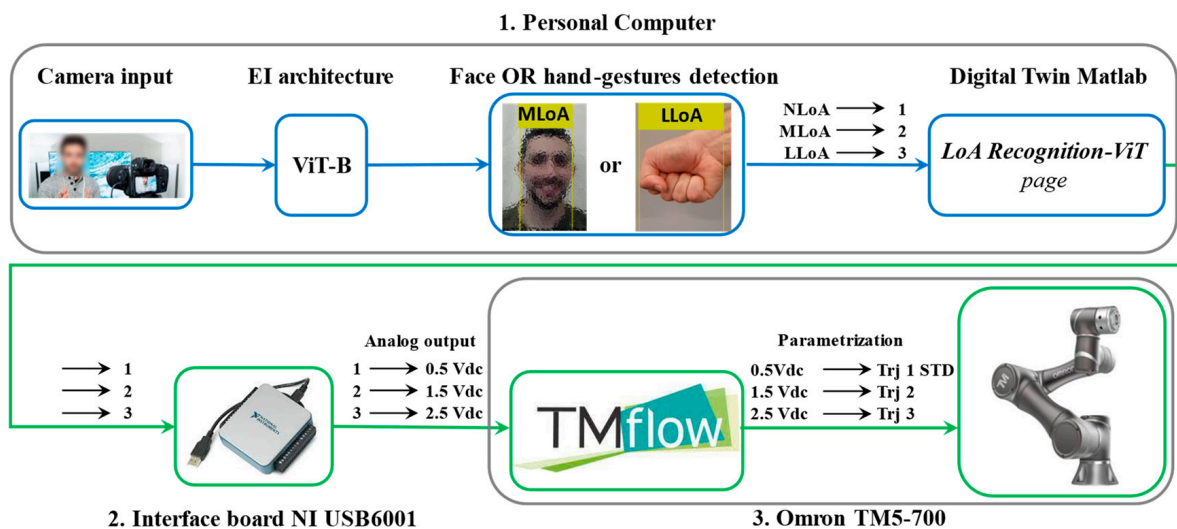
It was carried out by verifying the correspondence of known facial expressions and hand gestures with the estimated LoA; hence, the correspondence of the estimated LoA with the computed trajectory, as shown in Figure 8b; and, finally, the correspondence of the estimated LoA with the analog output value generated by the interface board. In the validation tests, the same test volunteer performed 50 facial expressions, kept constant within the time window, to be referred to as NLoA, MLoA, and LLoA, respectively, for a total amount of 150 tests. Similarly, the same number of tests based on hand gestures were carried out. Each test was conducted by imitating the content of a suitable image selected from the dataset. Therefore, the subject was not emotionally involved in the application. Face and hands were placed in front of the webcam.

All tests gave positive results: there was always a correspondence between the estimated facial expressions/hand gestures with the simulated trajectory of the cobot and the value of the analog output signal generated by the interface board. Under uncertain conditions (two or more faces, two or more hands recorded, or no recognition), the cobot was adjusted to the safest trajectory.

## 5. The Experimental Activity

The experimental activity aims to validate the implementation of EI to the cobot in a real collaborative working condition.

With reference to Figure 9, the testbench is made of three components: (1) a OMEN HP Inc. (Palo Alto, CA, USA) PC (i7-11th generation laptop, NVIDIA RTX2060 6Gb GPU, 32 Gb Ram 1Tb SSD HDD) equipped with the external webcam and the ViT-DT system; (2) a National Instruments (Austin, TX, USA) Data Acquisition Board (NI-DAQ) USB6001 adopted as an interface board for the ViT–cobot communication; (3) the Omron TM5-700 cobot (OMRON Corp., Taipei, Taiwan) managed by TMflow programming software, release 1.86.2300.



**Figure 9.** The schematic of the experimental testbench: (1) the PC; (2) the NI USB6001 board; (3) the Omron TM5-700 cobot.

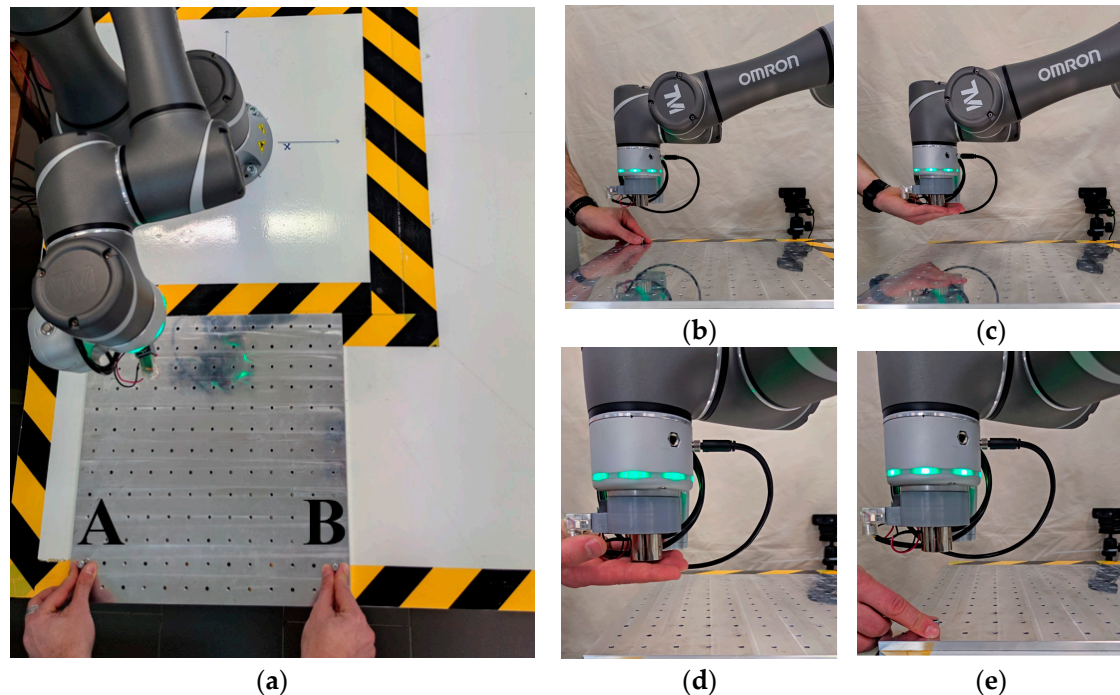
Except for the default emergency stop button connected to the control unit of the cobot, no external safety laser scanners, curtains, or gate systems were adopted. The end-effector is a cylindrical electro-magnet without sharp edges. Cobot was set in the collaborative function mode. No redundancy occurs between the proposed solution and the safety controller of the robot: if an accidental human–cobot collision occurs during the motion, the cobot immediately stops.

With reference to Figure 10, tests were carried out as described in Section 3: manual removal of a screw from a plate mounted on the cobot support bench takes place; the cobot picks the screw placed in the human hand at Point A and moves to Point B, where it releases the screw in the human hand; manual placement of the screw into the plate takes place; then, the cobot moves to Point A, following the previous trajectory in reverse. To reduce the risks of accidental collisions, human–cobot contact occurs when the cobot is stopped during the phase of the screw exchange. This task is a typical occurrence of HRIs where the human operator must pay attention and be actively involved in the experimental validation described below.

For the trajectory planning, according to the 90° clockwise rotated cobot reference system, the following values were set:  $x_0$  equal to  $-647$  mm,  $h_0$  equal to 145 mm,  $z_0$  equal to 145 mm, and  $r$  equal to 170 mm; moreover,  $y_A$  was equal to  $-170$  mm.

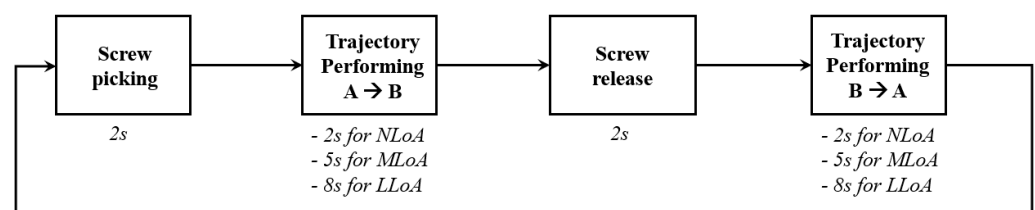
A symmetrical trapezoidal speed profile was set on the TCP: the acceleration and deceleration ramps were defined in terms of ms to reach the set or the zero speed values, respectively (according to the requirements of the cobot software). For each trajectory, the value of the speed, corresponding to the constant value of the speed profile, was set equal to 0.2 m/s, 0.1 m/s, and 0.07 m/s associated with the NLoA, MLoA, and LLoA trajectories, respectively. The highest value of the speed (0.2 m/s) was lower than the highest value of

the speed of the cobot working in collaborative mode (0.243 m/s), as imposed by ISO/TS 15066 in the case of accidental contacts between the cobot and the upper arm, forearm, and wrist of the operator. On a page of the cobot software, it was possible to set the body area subjected to accidental contacts and obtain the kinematic and dynamic constraints according to the standard. The other speed values were set based on Equation (11).



**Figure 10.** The experimental activity: (a) overview of the testbench; (b) manual removal of the screw; (c) picking task at Point A; (d) release task at Point B; (e) manual placement of the screw.

In all the trajectories, the ramp-time duration was set equal to 150 ms, corresponding to the acceleration/deceleration equal to  $1.33 \text{ m/s}^2$ ,  $0.67 \text{ m/s}^2$ , and  $0.47 \text{ m/s}^2$  associated with the NLoA, MLoA, and LLoA trajectories, respectively. Due to these values, the time length of each task (from A to B) equaled about 2, 5, and 8 s for the NLoA, MLoA, and LLoA trajectories, respectively. The picking and the release of the screw required 2 s. The task sequence is shown in Figure 11 in the case of no collisions.



**Figure 11.** The sequence of the pick-and-place task. Times refer to the absence of collisions.

Acquisitions of the cartesian poses of the TCP, joint positions, and velocities of the TCP of the cobot over time are shown in Figure 12, Figure 13, and Figure 14, respectively.

Moreover, as imposed by the cobot software according to ISO/TS 15066, the following constraints were applied: the maximum joint speed was set equal to 190, 190, 190, 235, 235, and  $235^\circ/\text{s}$  for the joints J1, J2, J3, J4, J5, and J6, respectively; the maximum joint torque was set equal to 25, 25, 25, 12, 12, and 12 Nm for the joints J1, J2, J3, J4, J5, and J6, respectively.

Finally, for the task execution, the perpendicularity constraint of the TCP, with respect to the robot base, was set. In detail, the rotation angles  $R_x$ ,  $R_y$ , and  $R_z$  of the TCP were set

at  $180^\circ$ ,  $0^\circ$ , and  $90^\circ$ , respectively. Joints J5 and J6 were fixed at  $90^\circ$  and  $86^\circ$ , respectively. A point-to-point control was applied by the TMflow programming software. The relevant points of the trajectory were imposed and calculated according to the parameters described in Section 3. The “Circle Node” command of the mentioned software was applied for the trajectory execution, passing through the relevant points. Other control parameters cannot be adjusted in the cobot software.

As for the test, before its execution, the test manager, hereinafter called the tester, instructed the test volunteer, hereinafter called the subject, about the operations to be performed. The subject carried out preliminary training for 10 min. Neither the purpose of the test nor the change of trajectory nor the change of speed was communicated to the subject. Ten subjects (seven men and three women; average age  $26.6 \pm 4.7$ ) were submitted to the test. Each signed a consent form about the test modality and possible risks before executing the test.

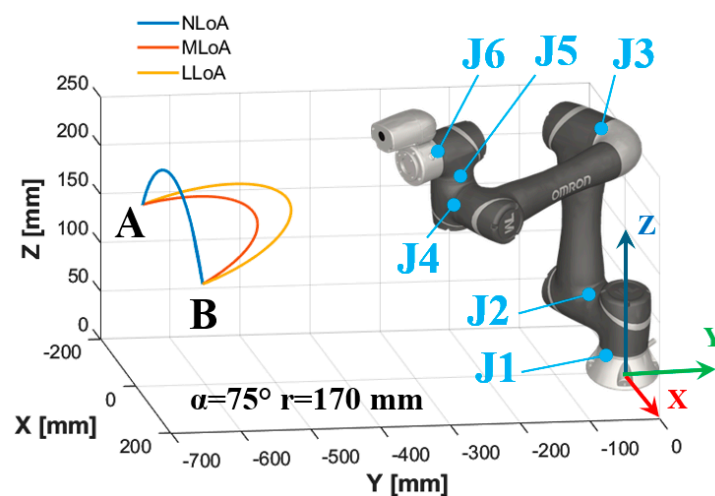


Figure 12. Cartesian poses of the TCP in the experimental activity.

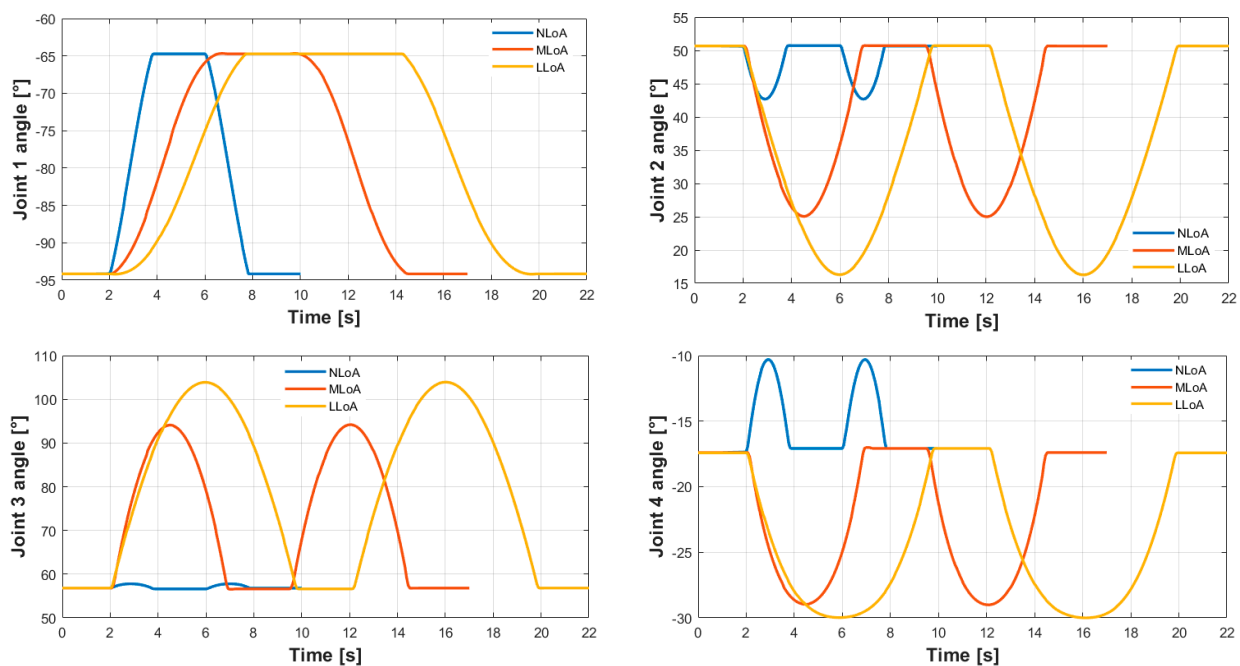


Figure 13. Joint angles of the cobot in the experimental activity. Angles of J5 and J6 are not reported because their constant values equal  $90^\circ$  and  $86^\circ$ , respectively.

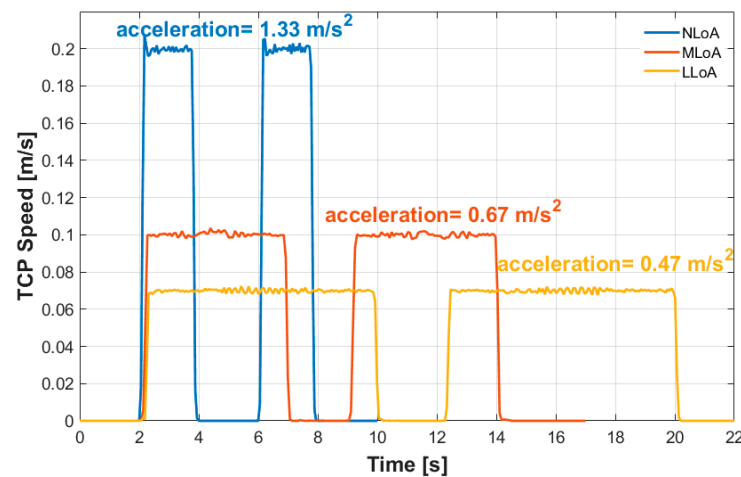


Figure 14. Speed profiles of the TCP in the experimental activity.

The test began when the ViT and DT started to run and the ViT-DT communication was opened. Therefore, the webcam started to record the subject, which began the task. The ViT returned the average LoA every second, computed on the images acquired in the previous floating time window  $t$ , set to 10 s. Moreover, the duration allowed the tester to distract the subject from the in-execution operation and verify the cobot's real-time behavior. The LoA and the webcam images were displayed in the ViT and DT with the related text string. Each LoA was associated with a numerical value: 0 to NLoA, 1 to MLoA, and 2 to LLoA, respectively. Each number was related to a voltage value generated by the NI-DAQ board: 0.5 Vdc, 1.5 Vdc, and 2.5 Vdc with 1, 2, and 3, respectively. Depending on the analog value acquired by the cobot control unit, the TMflow software calculated the trajectory and speed variables according to Equations (1)–(11). The coordinates of points A and B and  $x_0$ ,  $h_0$ ,  $z_0$ , and  $r$  were constant parameters defined in TMflow.

The test was divided into two phases: the first, for estimating only facial expressions and, the second, for only hand gestures. The duration of the single test was 5 min. Each test was repeated 5 times for an overall amount of 25 min. During the test, the subject was distracted and prompted by the tester to talk about various topics. Talking in a friendly and jovial manner was considered a disruptive phase for transitioning from NLoA to MLoA; speaking with pride and irreverence was considered a phase for the transition from NLoA to LLoA. The change of LoA was considered correct if the average value estimated by the ViT remained constant for at least 10 s. Results are shown in Table 4: the circles indicate a correctly estimated change in LoA, on the contrary to the crosses.

Table 4. Results of the experimental activity.

Subject			Facial Expressions (Phase 1)				Hand Gestures (Phase 2)			
No.	Age	Gender	NLoA	MLoA	LLoA	Inference [%]	NLoA	MLoA	LLoA	Inference [%]
1	26	M	o	o	o	100	o	o	o	100
2	31	M	o	o	o	100	X	o	o	66.6
3	27	M	o	o	o	100	X	X	X	0
4	22	M	o	o	o	100	o	o	o	100
5	28	M	o	o	X	66.6	o	o	o	100
6	27	M	o	o	X	66.6	o	o	o	100
7	25	F	o	o	o	100	o	o	X	66.6
8	26	F	o	o	X	66.6	o	o	o	100
9	26	M	o	o	o	100	o	o	o	100
10	28	F	o	o	X	66.6	o	o	o	100
Average Inference						86.6	Average Inference			73.3

The ViT-based EI accurately estimated the transition from NLoA to MLoA for facial expressions and hand gestures. The estimation of the transition from NLoA to LLoA was only sometimes successful. The inference was calculated as the overall success rate for each subject in each phase. The average inferences in Phase 1 and Phase 2 were 86.6% and 73.3%, respectively.

Regardless of the test phase, NLoA, MLoA, and LLoA were properly estimated in 90%, 95%, and 70% of the tests, respectively; in 65% of the tests, the system properly estimated all the LoAs; in 30%, it failed the estimation of one LoA; in 10%, it failed the estimation of all LoAs. As for the facial expressions (Phase 1), in 60% of subjects, LoAs were estimated at 100%; in the remaining 40%, only the LLoA was not properly estimated. As for the hand gestures (Phase 2), in 70% of subjects, LoAs were estimated at 100%; in the remaining 30%, NLoA was not estimated at 10%; and in 10%, all LoAs were not estimated for the same subject.

In detail, in Phase 1, for Subjects 5, 6, and 10, the system did not identify a correct transition from NLoA to LLoA since the generated facial expressions were very close to the neutral ones. For Subject 8, large glasses did not allow an accurate analysis of the eye and eyebrow expression, significant elements for the discrimination of the transition from NLoA to LLoA. However, some tests were repeated without glasses and the transition from NLoA to LLoA was properly detected. In Phase 1, NLoA and MLoA were always correctly estimated for all subjects. This is probably because the dataset is biased towards these two LoAs: the number of images used is significantly higher than in the other class.

Then, in Phase 2, for Subject 2, the estimation of NLoA was invalidated by the presence of a checkered shirt: ViT confused the horizontal lines of the squares with the outline of the fingers. For Subject 3, a change in LoA was never detected due to the color of the shirt being the same as that of the hands. To demonstrate this assumption, a green sheet of paper was placed between the shirt and the hands: some tests were carried out where the change in LoA was detected. For Subject 7, the transition from NLoA to LLoA failed due to the large number of finger rings. This assumption was confirmed since some tests without rings were repeated, providing proper identification.

In all tests, the transition to NLoA was verified: the change of trajectory and speed created doubts in the subjects regarding the malfunctioning of the cobot. From here, they returned to the NLoA. From these tests, it was seen that the average experimental inference for Phase 1 was in line with the results obtained from the validation of the architecture while, in Phase 2, they were 25.4% lower than the validation for the reasons explained above.

The aim of increasing HRIs through EI has been achieved. Indeed, during the execution of the task, no collisions were recorded even when the subject's LoA under examination was low, thus avoiding the robot's downtime and maintaining a high safety rate. Furthermore, from experimental tests, using colored visual protective devices, light-colored work uniforms, or accessories could compromise the system's functionality. To guarantee the maximization of the ViT-based system performance, we suggest limiting the variability of the garments and the presence of accessories. For example, the same uniform could be adopted in a manufacturing environment, avoiding some kinds of accessories, such as rings.

The work showed that EI can be another valuable tool for improving HRIs. In addition, the choice of a camera with a self-balancing light proved correct. Effectively, the instantaneous effect of external disturbances did not affect the LoA recognition. Indeed, environmental conditions related to lighting variability in the laboratory where the tests were conducted did not affect the recognition of the LoA. This was also helped by the dataset constructed from images taken under different lighting conditions rather than under strict conditions, as is typically required in machine vision systems.

Similarly, the positioning of the camera did not affect the subject's LoA for either the facial expression or the hand gestures recorded. This means that applying the proposed system would not change the operator's working conditions and would not be invasive for the operator.



Although brief training had been performed, it is believed that in normal working conditions, the proposed system can be directly implemented at the workstation. As for the algorithm used, it is performant for the type of application considered.

The achievement of an inference very close to 100% is related to the dataset employed for the ViT training. Indeed, the dataset should consider all of the typical occurrences of the application.

The choice to modify the task in trajectory and speed proved correct for the present work's aim. Safety was ensured as the robot effectively moved away from the subject as the LoA was lowered. In addition, the subject was effectively induced to be critical of the robot's behavior and, therefore, to restore the correct LoA. Again, task planning must depend on the peculiarities of the application.

The experimental activity revealed the effective reduction of the cobot downtime. For its estimation, the baseline case refers to the sequence of Figure 12 associated with the NLoA, according to which the cobot should work and the time length of the task, after the picking of the screw, should be the shortest (6 s). Hence, the time taken to restore the cobot if a collision occurs during the baseline case was compared to the time taken to perform the longest task, which lasts 18 s after the picking of the screw, associated with the LLoA.

The time taken to restore the cobot is the sum of the time to execute the following items. When a collision occurs, the cobot stops; then, the operator must first realize the collision, approach and take the robot stick, approach the wrist of the robot to check if the screw is attached (collision occurred in the trajectory performing from A to B) or not (collision occurred in the trajectory performing from B to A) at the end-effector, reset the collision alert by the robot stick, pick the screw (only if it is attached) and provide for its placement, put back the robot stick, and come back to the workplace. Experimentally, this procedure lasts over 16 s if the screw was previously released; on the contrary, it takes over 20 s. Then, the operator must push the start/restore button; the cobot waits for 3 s to move towards Point A with the same motion law associated with the NLoA trajectory. The motion takes a time variable of about 0–2 s. This range depends on the position of the TCP where the collision occurs: if it happens near point A, the cobot requires a very short time to reach point A; if it is near point B, the cobot requires almost 2 s.

The lowest value of the restore time is equal to about 19 s: it means that the collision occurred while performing the trajectory from B to A, near point A. The highest value of the time is equal to about 25 s: it means that the collision occurred while performing the trajectory from A to B, near point B.

In both cases, restore time was higher than the time taken to perform the task associated with the LLoA. Moreover, time measurement was carried out in lab tests where the time to realize the collision is less than 1 s: the subject voluntarily collides with the cobot and rapidly follows the mentioned procedure. Actually, in a real scenario, the operator should not be concentrated and the time to realize the collision should be higher.

## 6. Conclusions

The present work introduced the adoption of a ViT-based EI for the decision-making process of a cobot in a real application typical of an industrial scenario. To simulate the behavior of the cobot, a new DT was designed, developed, and validated numerically and, hence, by an experimental activity, carried out with a real cobot. The proposed approach demonstrated the capability of the EI to induce the operator to pay due attention in the collaborative environment. It could allow better acceptability, comfort, and success in the human–robot colleague interaction. Moreover, it demonstrated that the adoption of a simple camera, or a set of cameras, allows the creation of a less invasive system for the operator (no wearable devices are adopted) and can be easily integrated into an existing robotic cell or in other environments where a cobot can interact with humans. Finally, the operator is not required to train or learn how to use the system; instead, the operator learns how to improve the interaction with the robotic colleague.

However, the proposed system has limitations: it requires the improvement of the hardware, which is made of a personal computer, a webcam, and a microcontroller board near the work area. A structure to place several cameras that do not hinder the work area and simultaneously monitor the operator can be designed. A stereo camera could improve the set-up. Moreover, integrating the ViT architecture algorithm into TMflow should simplify the required hardware.

Improvements can certainly be applied: for the LoAD, a campaign of simultaneous acquisitions of facial expression and hand gestures, truly classified in terms of attention, must be carried out; the LoAD must be enriched with elements that currently act as a disturbance (glasses, rings, accessories). The performance of the ViT-B can be improved by adopting a multimodal architecture that can simultaneously analyze different inputs (fusion sensors), such as signals from external sensors, as in the present application, and suitable biometric signals, such as galvanic skin response signals or data about biometric measurements carried out by a smartwatch. This improvement should allow the use of cobots not only in industrial fields but also in domestic, assistive, and medical fields where the ViT-B must consider more variable factors, such as different types of accessories, garments, and environmental conditions.

The next step of the research will focus on two main aspects: the implementation of algorithms for the dynamic correction of the movement trajectory to increase the operator's safety further and solving the inter- and intra-personal variability of classification, regardless of the field of application of the proposed approach. The last topic is well known for the classification. In the present study, it was not considered because the research focused on developing and testing the approach's feasibility using the pre-trained ViT fine-tuned with a custom dataset. In future works, the inter- and intra-personal variability of classification will be a fundamental issue to analyze and limit to generalize the methodology and apply it to more scenarios.

**Supplementary Materials:** The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/machines12020113/s1>, Video S1: Facial expressions recognition during a task, Video S2: Hand gestures recognition during a task, Video S3: Collisions and recovery time.

**Author Contributions:** Conceptualization, M.G.A. and E.M.; methodology, M.G.A. and E.M.; software, M.G.A. and E.M.; validation, M.G.A., E.M. and N.S.; formal analysis, M.G.A., E.M. and N.S.; investigation, E.M. and N.S.; resources, M.G.A.; data curation, M.G.A., E.M. and N.S.; writing—original draft preparation, E.M. and N.S.; writing—review and editing, M.G.A., P.B.Z. and C.M.; visualization, M.G.A., E.M. and N.S.; supervision, M.G.A., P.B.Z. and C.M. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** In this study, the publicly available datasets FER-2013, HANDS, and HaGRID were considered and can be accessible, respectively, at <https://www.kaggle.com/datasets/msambare/fer2013>, <https://data.mendeley.com/datasets/ndrczc35bt/1>, and <https://github.com/hukenovs/hagrid>, accessed on 18 January 2024.

**Acknowledgments:** The authors want to thank Alessandro Lisci for his precious help in the experimental activity.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Schwab, K. *The Fourth Industrial Revolution*; World Economic Forum: Geneva, Switzerland, 2016.
2. Dzedzickis, A.; Subačiūtė-Žemaitė, J.; Šutinys, E.; Samukaitė-Bubnienė, U.; Bučinskas, V. Advanced Applications of Industrial Robotics: New Trends and Possibilities. *Appl. Sci.* **2022**, *12*, 135. [CrossRef]
3. Kyrarini, M.; Lygerakis, F.; Rajavenkatanarayanan, A.; Sevastopoulos, C.; Nambiappan, H.R.; Chaitanya, K.K.; Babu, A.R.; Mathew, J.; Makedon, F. A Survey of Robots in Healthcare. *Technologies* **2021**, *9*, 8. [CrossRef]
4. Yin, S.; Yuschenko, A. Application of Convolutional Neural Network to Organize the Work of Collaborative Robot as a Surgeon Assistant. In Proceedings of the International Conference on Interactive Collaborative Robotics ICR2019, Istanbul, Turkey, 20–25 August 2019; pp. 287–297.

5. Antonelli, M.G.; Beomonte Zobel, P. Automated screwing of fittings in pneumatic manifolds. *Int. J. Autom. Technol.* **2021**, *15*, 140–148. [CrossRef]
6. Kim, W.; Lorenzini, M.; Balatti, P.; Nguyen, P.D.H.; Pattacini, U.; Tikhanoff, V.; Peternel, L.; Fantacci, C.; Natale, L.; Metta, G.; et al. Adaptable Workstations for Human-Robot Collaboration: A Reconfigurable Framework for Improving Worker Ergonomics and Productivity. *IEEE Robot. Autom.* **2019**, *26*, 14–26. [CrossRef]
7. Lanzoni, D.; Negrello, F.; Fornaciari, A.; Lentini, G.; Ierace, S.; Vitali, A.; Regazzoni, D.; Ajoudani, A.; Rizzi, C.; Bicchi, A.; et al. Collaborative Workcell in Industrial Assembly Process with Online Ergonomics Monitoring. In Proceedings of the I-RIM Conference, Milan, Italy, 13–14 October 2022; pp. 201–203.
8. *ISO/TS 15066:2016; Robots and Robotic Devices-Collaborative Robots*. ISO: Geneva, Switzerland, 2016.
9. Grella, F.; Baldini, G.; Canale, R.; Sagar, K.; Wang, S.A.; Albini, A.; Jilich, M.; Cannata, G.; Zoppi, M. A Tactile Sensor-Based Architecture for Collaborative Assembly Tasks with Heavy-Duty Robots. In Proceedings of the 20th ICAR, Ljubljana, Slovenia, 6–10 December 2021; pp. 1030–1035.
10. AIRSKIN<sup>®</sup>. Available online: <https://www.airskin.io/> (accessed on 4 November 2023).
11. Malekzadeh, M.S.; Queißer, J.F.; Steil, J.J. Multi-Level Control Architecture for Bionic Handling Assistant Robot Augmented by Learning from Demonstration for Apple-Picking. *Adv. Robot.* **2019**, *33*, 469–485. [CrossRef]
12. Manti, M.; Hassan, T.; Passetti, G.; D’Elia, N.; Laschi, C.; Cianchetti, M. A Bioinspired Soft Robotic Gripper for Adaptable and Effective Grasping. *Soft Robot.* **2015**, *2*, 107–116. [CrossRef]
13. Antonelli, M.G.; Zobel, P.B.; D’ambrogio, W.; Durante, F. Design Methodology for a Novel Bending Pneumatic Soft Actuator for Kinematically Mirroring the Shape of Objects. *Actuators* **2020**, *9*, 113. [CrossRef]
14. Antonelli, M.G.; D’Ambrogio, W. Soft Pneumatic Helical Actuator for Collaborative Robotics. In Proceedings of the 4th International Conference of IFToMM, Naples, Italy, 7–9 September 2022; pp. 702–709.
15. Neri, F.; Forlini, M.; Scoccia, C.; Palmieri, G.; Callegari, M. Experimental Evaluation of Collision Avoidance Techniques for Collaborative Robots. *Appl. Sci.* **2023**, *13*, 2944. [CrossRef]
16. Scoccia, C.; Palmieri, G.; Palpacelli, M.C.; Callegari, M. A Collision Avoidance Strategy for Redundant Manipulators in Dynamically Variable Environments: On-Line Perturbations of Off-Line Generated Trajectories. *Machines* **2021**, *9*, 30. [CrossRef]
17. Scalera, L.; Giusti, A.; Vidoni, R.; Gasparetto, A. Enhancing fluency and productivity in human-robot collaboration through online scaling of dynamic safety zones. *Int. J. Adv. Manuf. Technol.* **2022**, *121*, 6783–6798. [CrossRef]
18. Liu, Q.; Liu, Z.; Xu, W.; Tang, Q.; Zhou, Z.; Pham, D.T. Human-robot collaboration in disassembly for sustainable manufacturing. *Int. J. Prod. Res.* **2019**, *57*, 4027–4044. [CrossRef]
19. Sajedi, S.; Liu, W.; Eltouny, K.; Behdad, S.; Zheng, M.; Xiao, L. Uncertainty-Assisted Image-Processing for Human-Robot Close Collaboration. *IEEE Robot. Autom. Lett.* **2022**, *7*, 4236–4243. [CrossRef]
20. Dimitropoulos, N.; Togiias, T.; Zacharaki, N.; Michalos, G.; Makris, S. Seamless Human–Robot Collaborative Assembly Using Artificial Intelligence and Wearable Devices. *Appl. Sci.* **2021**, *11*, 5699. [CrossRef]
21. Neto, P.; Simão, M.; Mendes, N.; Safeea, M. Gesture-based human-robot interaction for human assistance in manufacturing. *Int. J. Adv. Manuf. Technol.* **2019**, *101*, 119–135. [CrossRef]
22. Abdullah, M.; Lihui, W. Advanced Human-Robot Collaborative Assembly Using Electroencephalogram Signals of Human Brains. In Proceedings of the 53rd CIRP Conference on Manufacturing Systems (CIRP-CMS 2020), Chicago, IL, USA, 1–3 July 2020; pp. 1200–1205. [CrossRef]
23. Adel, A. Future of industry 5.0 in society: Human-centric solutions, challenges and prospective research areas. *J. Cloud Comp.* **2022**, *11*, 40. [CrossRef] [PubMed]
24. Kahneman, D. *Thinking, Fast and Slow*; Macmillan: New York, NY, USA, 2011.
25. James, W. *The Principles of Psychology*; Henry Hold and Company: New York, NY, USA, 1890; Volume 1.
26. Norman, G.J.; Necka, E.; Berntson, G.G. 4—The Psychophysiology of Emotions. In *Emotion Measurement*; Herbert, L., Meise Man, A., Eds.; Woodhead Publishing: Sawston, UK, 2016; pp. 83–98.
27. Buodo, G.; Sarlo, M.; Palomba, D. Attentional Resources Measured by Reaction Times Highlight Differences Within Pleasant and Unpleasant, High Arousing Stimuli. *Motiv. Emot.* **2002**, *26*, 123–138. [CrossRef]
28. Palomba, D.; Sarlo, M.; Angrilli, A.; Mini, A.; Stegagno, L. Cardiac Responses Associated with Affective Processing of Unpleasant Film Stimuli. *Int. J. Psychophysiol.* **2000**, *36*, 45–47. [CrossRef] [PubMed]
29. Mayer, J.D.; Salovey, P.; Caruso, D.R. Emotional Intelligence: New Ability or Eclectic Traits? *Am. Psychol.* **2008**, *63*, 503–517. [CrossRef] [PubMed]
30. Marcos-Pablos, S.; García-Peñalvo, F.J. Emotional Intelligence in Robotics: A Scoping Review. In *New Trends in Disruptive Technologies, Tech Ethics and Artificial Intelligence: The DITTET Collection*, 1st ed.; Springer International Publishing: Cham, Switzerland, 2022; pp. 66–75.
31. Spezialetti, M.; Placidi, G.; Rossi, S. Emotion Recognition for Human-Robot Interaction: Recent Advances and Future Perspectives. *Front. Robot. AI* **2020**, *7*, 145. [CrossRef] [PubMed]
32. Kollias, K.-F.; Syriopoulou-Delli, C.K.; Sarigiannidis, P.; Fragulis, G.F. The Contribution of Machine Learning and Eye-Tracking Technology in Autism Spectrum Disorder Research: A Systematic Review. *Electronics* **2021**, *10*, 2982. [CrossRef]

33. Banire, B.; Al Thani, D.; Qaraqe, M.; Mansoor, B. Face-Based Attention Recognition Model for Children with Autism Spectrum Disorder. *J. Healthc. Inform. Res.* **2021**, *5*, 420–445. [[CrossRef](#)] [[PubMed](#)]
34. Geetha, M.; Latha, R.S.; Nivetha, S.K.; Hariprasath, S.; Gowtham, S.; Deepak, C.S. Design of face detection and recognition system to monitor students during online examinations using Machine Learning algorithms. In Proceedings of the 2021 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 27–29 January 2021. [[CrossRef](#)]
35. Tawari, A.; Trivedi, M.M. Robust and Continuous Estimation of Driver Gaze Zone by Dynamic Analysis of Multiple Face Videos. In Proceedings of the 2014 IEEE Intelligent Vehicles Symposium (IV), Dearborn, MI, USA, 8–11 June 2014.
36. Hu, Z.; Lv, C.; Hang, P.; Huang, C.; Xing, Y. Data-Driven Estimation of Driver Attention Using Calibration-Free Eye Gaze and Scene Features. *IEEE Trans. Ind. Electron.* **2022**, *69*, 2. [[CrossRef](#)]
37. Chu, H.C.; Tsai, W.W.J.; Liao, M.J.; Chen, Y.M. Facial emotion recognition with transition detection for students with high-functioning autism in adaptive e-learning. *Soft Comput.* **2018**, *22*, 2973–2999. [[CrossRef](#)]
38. Kotsia, I.; Pitas, I. Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE Trans. Image Process* **2007**, *16*, 172–187. [[CrossRef](#)] [[PubMed](#)]
39. Hua, W.; Dai, F.; Huang, L.; Xiong, J.; Gui, G. Hero: Human emotions recognition for realizing intelligent internet of things. *IEEE Access* **2019**, *7*, 24321–24332. [[CrossRef](#)]
40. Lu, C.T.; Su, C.W.; Jiang, H.L.; Lu, Y.Y. An interactive greeting system using convolutional neural networks for emotion recognition. *Entertain. Comput.* **2022**, *40*, 100452. [[CrossRef](#)]
41. Heredia, J.; Lopes-Silva, E.; Cardinale, Y.; Diaz-Amado, J.; Dongo, I.; Graterol, W.; Aguilera, A. Adaptive Multimodal Emotion Detection Architecture for Social Robots. *IEEE Access* **2022**, *10*, 20727–20744. [[CrossRef](#)]
42. Vazquez-Rodriguez, J.; Lefebvre, G.; Cumin, J.; Crowley, J.L. Transformer-Based Self-Supervised Learning for Emotion Recognition. In Proceedings of the 26th International Conference on Pattern Recognition (ICPR 2022), Montreal, QC, Canada, 21–25 August 2022.
43. Chaudhari, A.; Bhatt, C.; Krishna, A.; Mazzeo, P.L. ViTFER: Facial Emotion Recognition with Vision Transformers. *Appl. Syst. Innov.* **2022**, *5*, 80. [[CrossRef](#)]
44. Siriwardhana, S.; Kaluarachchi, T.; Billingham, M.; Nanayakkara, S. Multimodal Emotion Recognition with Transformer-Based Self Supervised Feature Fusion. *IEEE Access* **2020**, *8*, 176274–176285. [[CrossRef](#)]
45. Karatay, B.; Bestepe, D.; Sailunaz, K.; Ozyer, T.; Alhaji, R. A Multi-Modal Emotion Recognition System Based on CNN-Transformer Deep Learning Technique. In Proceedings of the 7th International Conference on Data Science and Machine Learning Applications (CDMA 2022), Riyadh, Saudi Arabia, 1–3 March 2022; pp. 145–150.
46. Toichoa Eyam, A.; Mohammed, W.M.; Martinez Lastra, J.L. Emotion-Driven Analysis and Control of Human-Robot Interactions in Collaborative Applications. *Sensors* **2021**, *21*, 4626. [[CrossRef](#)]
47. Lagomarsino, M.; Lorenzini, M.; Balatti, P.; De Momi, E.; Ajoudani, A. Pick the Right Co-Worker: Online Assessment of Cognitive Ergonomics in Human-Robot Collaborative Assembly. *IEEE Trans. Cogn. Dev. Syst.* **2022**, *15*, 1928–1937. [[CrossRef](#)]
48. Brandizzi, N.; Bianco, V.; Castro, G.; Russo, S.; Wajda, A. Automatic RGB inference based on facial emotion recognition. In Proceedings of the 2021 Scholar’s Yearly Symposium of Technology, Engineering and Mathematics, Catania, Italy, 27–29 July 2021.
49. Goodfellow, I.J.; Erhan, D.; Carrier, P.L.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.-H.; et al. Challenges in representation learning: A report on three machine learning contests. In Proceedings of the International Conference on Neural Information Processing, Daegu, Republic of Korea, 3–7 November 2013; pp. 117–124. [[CrossRef](#)]
50. Nuzzi, C.; Pasinetti, S.; Pagani, R.; Coffetti, G.; Sansoni, G. MEGURU: A gesture-based robot program builder for Meta-Collaborative workstations. *Robot. Comput.-Integr. Manuf.* **2021**, *68*, 102085. [[CrossRef](#)]
51. Kapitanov, A.; Makhlyarchuk, A.; Kvanchiani, K. HaGRID-H and Gesture Recognition Image Dataset. *arXiv* **2022**, arXiv:2206.08219.
52. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
53. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training Data-Efficient Image Transformers & Distillation through Attention. *arXiv* **2020**, arXiv:2012.12877.
54. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (System Demonstrations), Virtual Conference, 16–20 November 2020; pp. 38–45.
55. Loshchilov, I.; Hutter, F. Decoupled weight decay regularization. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019; pp. 94–156.
56. Chen, X.; Zheng, X.; Sun, K.; Liu, W.; Zhang, Y. Self-supervised vision transformer-based few-shot learning for facial expression recognition. *Inf. Sci.* **2023**, *634*, 206–226. [[CrossRef](#)]
57. Minaee, S.; Minaei, M.; Abdolrashidi, A. Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network. *Sensors* **2021**, *21*, 3046. [[CrossRef](#)] [[PubMed](#)]
58. Padhi, P.; Das, M. Hand Gesture Recognition using DenseNet201-Mediapipe Hybrid Modelling. In Proceedings of the International Conference on Automation, Computing and Renewable Systems (ICACRS 2022), Pudukkottai, India, 13–15 December 2022.

- 
59. Mathworks. Available online: <https://it.mathworks.com/products/robotics.html> (accessed on 5 November 2023).
  60. Omron Industrial Automation. Available online: <https://industrial.omron.eu/en/products/collaborative-robots> (accessed on 6 November 2023).

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.