Article

Facial Emotion Recognition from an Unmanned Flying Social Robot for Home Care of Dependent People

Anselmo Martínez [1], Lidia M. Belmonte [1,2], Arturo S. García [1,3], Antonio Fernández-Caballero [1,3,4,*] and Rafael Morales[1,2]

1 Instituto de Investigación en Informática de Albacete, Universidad de Castilla-La Mancha, 02071 Albacete, Spain; Anselmo.Martinez1@alu.uclm.es (A.M.); Arturosimon.Garcia@uclm.es (A.S.G.)
2 Departamento de Ingeniería Eléctrica, Electrónica, Automática y Comunicaciones, Universidad de Castilla-La Mancha, 02071 Albacete, Spain; LidiaMaria.Belmonte@uclm.es (L.M.B.); Rafael.Morales@uclm.es (R.M.)
3 Departamento de Sistemas Informáticos, Universidad de Castilla-La Mancha, 02071 Albacete, Spain
4 Biomedical Research Networking Center in Mental Health (CIBERSAM), 28016 Madrid, Spain
* Correspondence: Antonio.Fdez@uclm.es; Tel.: +34 967599200

Abstract: This work is part of an ongoing research project to develop an unmanned flying social robot to monitor dependants at home in order to detect the person's state and bring the necessary assistance. In this sense, this paper focuses on the description of a virtual reality (VR) simulation platform for the monitoring process of an avatar in a virtual home by a rotatory-wing autonomous unmanned aerial vehicle (UAV). This platform is based on a distributed architecture composed of three modules communicated through the Message Queue Telemetry Transport (MQTT) protocol: the UAV Simulator implemented in MATLAB/Simulink, the VR Visualiser developed in Unity, and the new emotion recognition (ER) System developed in Python. Using a face detection algorithm and a convolutional neural network (CNN), the ER System is able to detect the person's face in the image captured by the UAV's on-board camera and classify the emotion among seven possible ones (surprise, fear, happiness, sadness, disgust, anger or neutral expression). The experimental results demonstrate the correct integration of this new computer vision module within the VR platform, as well as the good performance of the designed CNN, with around 85% in the F1-score, a mean of the precision and recall of the model. The developed emotion detection system can be used in the future implementation of the assistance UAV that monitors dependent people in a real environment, since the methodology used is valid for images of real people.

Keywords: Flying Social Robot; Autonomous Unmanned Aerial Vehicle (UAV); Emotion Recognition; Convolution Neural Network (CNN); Virtual Reality (VR); Unity; MATLAB/Simulink; Python

1. Introduction

According to a now classic definition of social robots [1], these are robots that exhibit human social features like expressing and/or perceiving emotions; communicating with high-level dialogue; learning/recognising models of other agents; establishing/maintaining social relationships; using natural cues (gaze, gestures, etc.); exhibiting distinctive personality and character; learning/developing social competencies.

Within the field of social robots, our research group has opted for the use of autonomous unmanned aerial vehicles (UAVs) as a promising alternative for home care of dependent people, mainly elderly people living alone. The role of UAVs in Ambient Assisted Living for the elderly is a hot topic [2–4]. Such UAVs, equipped with an on-board camera, provide a novel solution for navigating around the inhabitant and monitoring him/her. The aim is to take images for analysis using computer vision techniques to determine the state of the person and, thus, to decide on the possible assistance or help required at any given moment [5]. For this, we consider emotion analysis as a fundamental tool. This is a non-invasive technique, in which computer vision algorithms are used to analyse facial images and possibly determine the mood of the person or the robot's intent as a social signal.

In this way, our UAVs can be considered flying social robots that must incorporate the capacity of people detection and tracking through perceiving human features, in addition to the perception required for the localisation, navigation and obstacle avoidance functions. Moreover, the recognition of facial expressions is also present in these flying social robots [6–8]. It is our intention to incorporate more human social features in a close future. For the development of such vision-based assistance UAVs for monitoring of dependent people, it is necessary to solve all the technical challenges for the safe navigation of the UAV in a house, as well as to implement the computer vision algorithms for the emotion analysis (or other algorithms to add new features or services in the future, such as a fall detection alarm), in the presence of occlusions, variable illumination, moving camera, and varying background [1].

Apart from this, to provide a solution that is accepted by the dependent person and can thus be of real use to him/her, it is also necessary to approach the design of the assistant UAV from the point of view of the assisted person [9,10]. Therefore, we have decided to design a virtual reality (VR) simulation platform that provides a safe and realistic environment to develop and test the different engineering solutions [11,12], and also allows us to conduct studies with potential end-users and people who know the problem of care for dependent people in order to adapt the design to their preferences.

This article focuses on the description of the new implementation of the VR platform, which initially had two modules [13], the UAV Simulator implemented in MATLAB/Simulink® and the VR Visualiser developed in Unity, to which a new emotion recognition (ER) system programmed in Python has been added. This computer vision module is the most important novelty of this work and its objective is to analyse the aerial images received from the UAV camera in order to determine the emotion of the person. It is worth mentioning that the VR Visualiser has also been updated to add new functionalities during the monitoring process so that the integration of the ER System is complete. In this way, through the Message Queue Telemetry Transport (MQTT) protocol, the three modules communicate to exchange the necessary information for the correct operation of each one, thus emulating the behaviour of an assistant UAV monitoring an avatar in a virtual home.

The structure of the article is as follows: Section 2 presents an overview of the VR platform, detailing the architecture and the MQTT communication used. Sections 3 and 4 briefly describe the UAV Simulator and the VR Visualiser, respectively. The new ER System is detailed in Section 5, while experimental results are discussed in Section 6. Finally, Section 7 presents conclusions and future work.

2. General Description of the VR Platform

This work is framed in an ongoing research project aimed to improve the quality of life of dependent persons by means of the design of a social flying robot for home care of dependants. The main task of the UAV is to monitor the patient at home, that is, to carry out flights to capture images of the person from the on-board camera. These photographs will be sent to a base station for analysis in order to determine the person's condition and, based on this, be able to provide the necessary assistance in each case or situation.

In this context, we have developed a VR platform, which initial version has already been presented [13], capable of providing a realistic simulation environment for the validation of the different algorithms required for the operation of the assistance UAV, both at the level of navigation and control of the aircraft itself during the monitoring of the patient, as well as those related to image processing to determine the person's state. This computer vision features added by means of the new ER System are the main novelty of this work. In addition, this platform allows us to carry out studies with participants who realistically experience the monitoring task performed by the assistance UAV. This means that it is possible to carry out studies to evaluate different options and adjust the operation of the UAV to the preferences of the users, thus moving towards total user acceptance, which is fundamental in any social robot.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2021 doi:10.20944/preprints202103.0189.v1

2.1. High-Level Architecture

The VR simulation platform is based on the distributed architecture detailed in Figure 1. In its current implementation, this platform is composed of three modules; UAV Simulator, VR Visualizer and ER System, which are respectively in charge of: (i) simulating the UAV's dynamics, including its control algorithm and the monitoring trajectory planner, (ii) recreating the virtual environment where the UAV monitors the dependent person, and (iii) processing the images grabbed by the UAV's on-board camera to determine the person's mood.

The modules have been developed using different software programs according to the requirements of each one, and they can be executed on the same PC or on different ones, since all communicate with each other by means of the MQTT protocol as detailed later. Finally, it is worth mentioning that the distributed architecture used provides us with versatility during the design phases, and will make it easier in the future to replace the software modules with the hardware systems that will be developed for the final implementation of the assistance UAV in real environments.

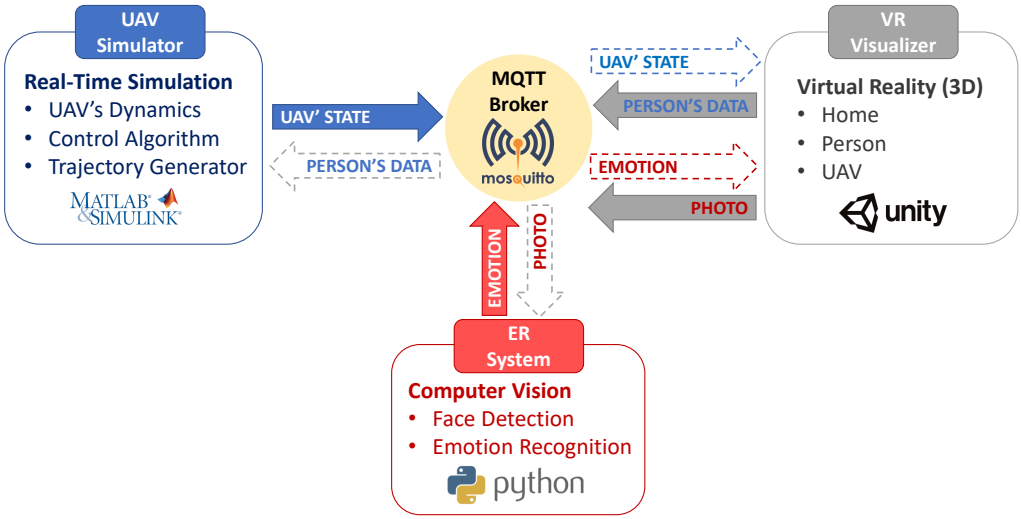

Figure 1. High level architecture of the VR simulation platform.

2.2. Communication

As mentioned above, the MQTT (Message Queuing Telemetry Transport) message protocol is used to communicate the different modules that make up the VR platform. It consists of a simple and light communication protocol based on TCP/IP with a system of publication and subscription of messages in topics. These topics are essential because the receivers need to be subscribed to one so that the MQTT server (also called *broker*) can send them the information published by the senders. As shown in Figure 1, the Mosquitto open source server is used for this purpose. Inside the VR platform, there are two main communication paths (see Figure 1): the first one is used to exchange information between the UAV Simulator and the VR Visualiser, and the second one between the latter and the vision-based ER System used in facial emotion analysis.

The UAV Simulator implemented in MATLAB/Simulink® models the dynamic behaviour of a quadrotor, including the trajectory planner and the control algorithm. The UAV will perform the process of monitoring the person, in this case the avatar in the VR application in Unity, for which it needs to know the information relating to his/her pose. This information is published from Unity in a topic called "Sim_UAV/Person/Data" at which the MATLAB/Simulink® client is subscribed. Based on this information, the planner calculates the reference path for the position and orientation of the quadrotor. This data is used by the control algorithm to determine the input to be applied to the UAV's model to minimise the error in tracking the trajectories, and thus get to guide the UAV as expected in the monitoring process. On the other hand, the information of the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2021 doi:10.20944/preprints202103.0189.v1

state of the UAV (position and orientation) is used to represent the flight of the UAV in the virtual environment, which is published from the MATLAB/Simulink® in the topic "Sim_UAV/UAV/State". This way, the Unity client (subscribed to the previous topic) periodically receives the information published by the UAV Simulator, and updates the pose of the 3d model of the UAV inside the virtual home.

Regarding the image processing and emotion recognition, it has been established that the UAV's camera (within the VR application in Unity) will be taking pictures of the avatar at one-second intervals. These images are automatically sent to a specific topic (called "Sim_UAV/UAV/Camera") to which the ER System is subscribed, so it can receive them at the moment of their publication. When the ER System finishes classifying the emotion, it publishes the results in a different topic (called "Sim_UAV/Person/Emotion") to which the Unity application is subscribed, receiving the data and showing them.

Finally, the MQTT communication in the VR platform is schematically summarised in Figure 2.

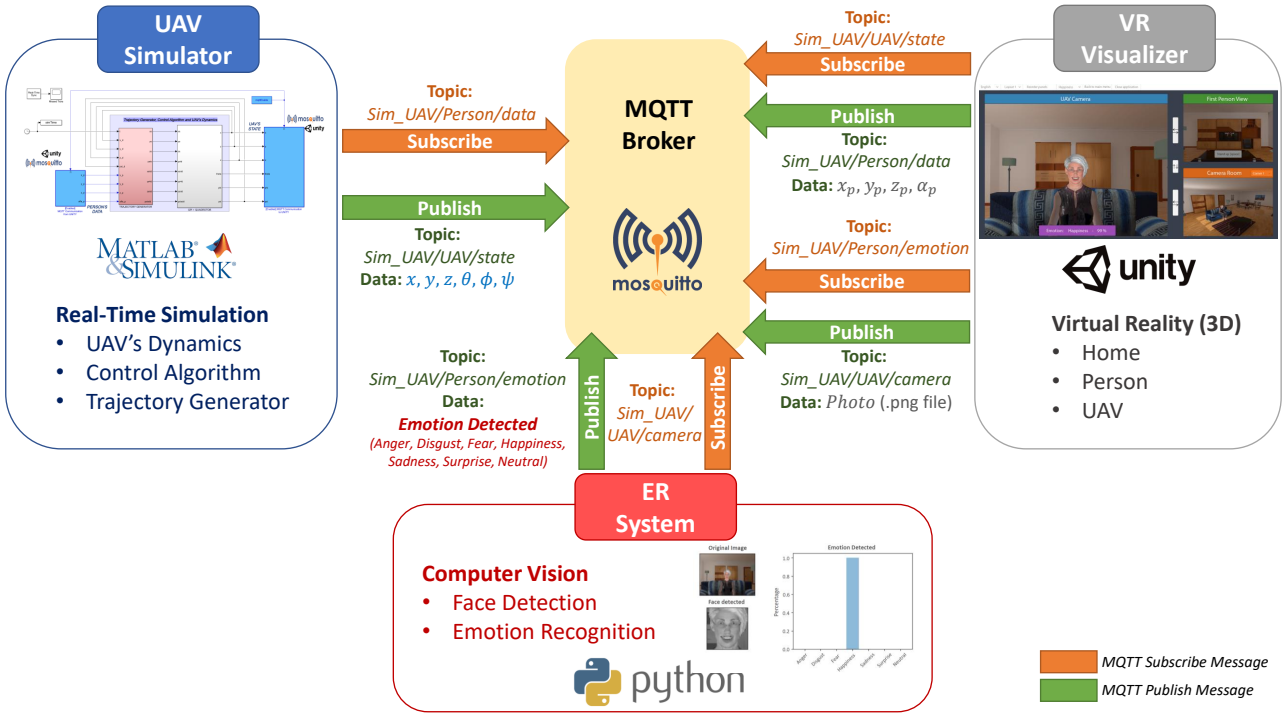

Figure 2. MQTT communication in the VR platform.

3. UAV Simulator

This section briefly describes the UAV Simulator module, whose main objective is to represent the dynamic behaviour of the assistant UAV, a rotatory-wing aircraft of type quadrotor which is in charge of flying at home to monitor the person. This way, the UAV Simulator has to calculate at each moment the state of the aircraft, that is the position and orientation, considering its dynamics, the action of the controller and the trajectory planner, which in the end determines the reference trajectories for the position and orientation of the UAV in order to perform the monitoring process. To do that, the following three components were implemented in MATLAB/Simulink software [13]:

- Quadrotor's Dynamic Model: It mathematically represents how the lift forces of the quadrotor changes when the rotational speed of its four propellers are modified, thus achieving the three possible motions; pitch, roll and yaw. It was obtained following the Euler Lagrange formulation according to [14] and can be consulted in [13,15]. Please note that the output of this component is the state of the UAV, its position and

 doi:10.20944/preprints202103.0189.v1

orientation, information that is sent to the VR Visualiser in order to reproduce the flight of the UAV by updating the position and orientation of a 3D virtual quadrotor.

- Control Algorithm: it is used to calculate what inputs should be applied to the quadrotor model in order to follow a specific trajectory reference, thereby ensuring that the UAV is correctly positioned and oriented during the monitoring flight. For this component, we designed a generalised proportional integral (GPI) controller based on the flatness theory that demonstrated good results in both stabilisation and tracking tasks, even in the presence of atmospheric disturbances and noise measurements, improving the performance of a traditional PID controller. The theoretical details of the GPI control scheme are described in [15].
- Trajectory Planner: In base of the person's position and orientation, which is received from the avatar in the VR Visualiser, a state-machine-based planner generates the references for the position and yaw angle of the UAV for each of the manoeuvres that make up the monitoring process. In the current implementation of the planner, it is possible to configure the tracking trajectory to define the height at which the UAV flies, as well as the trajectory (circular or elliptical) it describes around the person, to suit the user's preferences. Details of this planner can be found in [16].

4. VR Visualiser

The game engine Unity (also referred as Unity 3D) was used to develop the VR Visualiser in which the virtual home, the quadrotor UAV and the user's representation (avatar) are rendered. This module was initially included in the VR platform [13] to reproduce the flight of the assistant UAV by updating the position and orientation of the quadrotor 3D model according to the information received from the UAV Simulator, which in turn determines the monitoring trajectory based on the avatar's information received from Unity application.

As described in [13], two main user interfaces were implemented for non-immersive and immersive VR setups. The former makes use of a keyboard and mouse control mode. This mode is useful to assess the behaviour of the assistant UAV, since it is easier to visualise the UAV trajectory on a PC screen and using a third-person perspective with free-camera movements. The latter uses the HTC VIVE headset and its controllers, so the user can walk and look inside the virtual house, while the assistant UAV performs the monitoring process. This mode enables immersive studies with participants to be conducted while they experiment the labour of the social flying robot in first person.

The traditional interface is divided into three interchangeably frames to display different camera views (see Figure 11); the first-person view of the avatar, the on-board UAV camera view, and a selection of camera positions placed around the room. Other features or options added were: (i) two main selectable characters (avatar of an elderly male or an elderly female), (ii) the colour representation of the path followed by the quadrotor within the virtual environment to assess the correct functioning of the system (as this can be compared with the monitoring trajectory calculated by the UAV Simulator), and (iii) spatial sound to simulate the buzzing sounds of the UAV, thus increasing the realistic of the platform.

Now, with the integration of the new ER System into the VR platform, the VR Visualiser has been updated. Dynamic facial expressions have been added to the avatars, as will be described below (Subsections 5.1, 5.2). In addition, it has been established that the UAV's on-board camera captures images regularly during the monitoring process. These images are sent to the ER System for analysis. Once the image processing is finished, the avatar's emotion is displayed in the Unity app's user interface. More details about the bidirectional communication of these two modules are provided in the experimental results (Subsection 6.1).

 doi:10.20944/preprints202103.0189.v1

5. ER System

The emotion recognition (ER) system is the new computer vision module of the VR platform that simulates the operation of the assistant UAV for home care of dependent persons. This module is responsible for analysing the images received from the camera on board the UAV to determine the mood of the person by analysing their facial expression. It is mainly composed of two elements: the set of cascade classifiers that analyse the image to detect the person's face (or avatar's in this case), and the convolutional neural network (CNN) that analyses the facial expression to detect the person's emotion.

This way, the ER System analyses the images captured by the camera on board the UAV. This has made it necessary to update the application developed in Unity in order to provide the characters with the ability to express emotions during the simulation. Therefore, we begin by describing the design of the emotions in the avatars, as well as the implementation of the transition between the different emotions. This is followed by a description of the face detection algorithm, based on the aforementioned cascade classifiers, and then a description of the design of the convolutional neural network which classifies the avatar's emotion. This section ends with a description of the neural network training process.

5.1. Design of Emotions in Avatars

Two models were designed to represent a male and a female avatar. Figure 3 shows the faces of the two virtual characters. They initially show a neutral expression.

Figure 3. The faces of the avatars with neutral expression (left - old man, right - old woman).

Each model has a series of blendshapes, which consist of a set of coefficients that allow the modification of specific groups of vertices of their face. They focus on facial characteristics such as eyebrows, eyelids or lips. In addition, these blendshapes designed individually can be combined, enabling the generation of facial expressions in a simple way. The desing of the blendshapes that compose the six basic emotions defined by Ekman (anger, disgust, fear, happiness, sadness and surprise) [17] has been carried out following a procedure previously detailed [18], which is based on the well-known Facial Action Coding System [17]. Figure 4 shows how these emotions are seen in the faces of the two characters.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2021 doi:10.20944/preprints202103.0189.v1

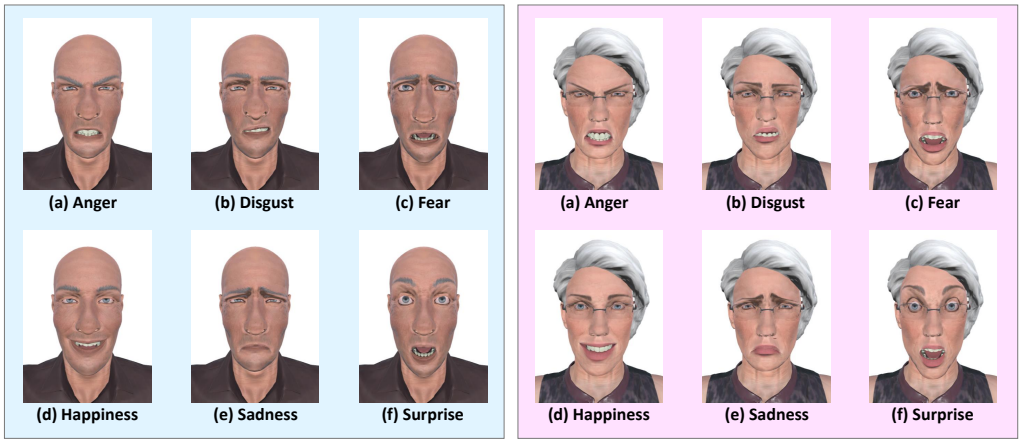

Figure 4. Basic emotions in the avatars (left - old man, right - old woman): (a) *anger*, (b) *disgust*, (c) *fear*, (d) *happiness*, (e) *sadness*, and (f) *surprise*.

5.2. Transitions between Emotions

The option to choose the desired emotion has been added to the Unity application, and can be established before starting the simulation. In addition, a drop-down menu has also been added to the user interface so that it can be changed at any time, even if the simulation has already started. Apart from the six basic emotions plus the neutral expression, a random mode that selects a different one every few minutes has also been added.

When moving from one emotion to another, a transition is made in which the facial expression changes little by little so that there is no sudden change. As mentioned before, each emotion is formed by the combination of a set of blendshapes that modify a specific area of the face. Therefore, to create a fluid transition between two expressions, the blendshape values of the previous one are increased or decreased by one unit each frame until they coincide with those of the following one.

5.3. Face Detection Algorithm

Cascade classifiers are used to recognise a person's face from a photograph. Next, the characteristics and operation of the cascade classifiers are described to later specify how this machine learning methodology has been implemented in the emotion recognition module.

5.3.1. Cascade Classifiers

Description

In machine learning, a *classifier* is an algorithm that can identify which of several categories a new example belongs to through pattern recognition, because it has been trained with a set of previously labelled cases. On the other hand, a *cascade classifier* is a combination method or *ensemble* that consists of the concatenation of several classifiers, where the information produced by each one of them as output is used as additional information for the next entry.

Cascade classifiers were initially proposed by Viola and Jones [19], and they supposed a new way of detecting objects in machine learning that is characterised by processing images extremely fast and with a high detection rate. This approach can be divided into three parts: (i) Transformation of the original image into a new representation called "integral image", which allows the identification of features to be computed very quickly; (ii) A machine learning algorithm based on *AdaBoost*, which consists of selecting a few of the most important characteristics to reduce their number and produce much more efficient classifiers; (iii) The cascade classifier to quickly discard unnecessary areas of the image and focus on the regions where the figure to find can be.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2021 doi:10.20944/preprints202103.0189.v1

Functioning

A cascade classifier has several stages, each of which is a combination of weak classifiers. This is based on the machine learning meta-algorithm called *Boosting*, which consists of putting together the results of several weak classifiers to form a robust one. The weak ones are those predictors with a low precision but slightly higher than a random one (that is, greater than 50% in a binary classification problem), while the robust ones are those with a high percentage of correctness.

The operation is as follows, each area of the image is inspected by means of a sliding window and, for each stage of the cascade classifier, it is checked whether the object being sought is found in that subsection (checking the characteristics of the area). When it fails in one stage (that is, the value given by one of the weak classifiers does not exceed a threshold), it no longer goes to the next stage, which greatly saves the calculation. Only the area that manages to pass all the stages is finally considered positive. The training of a cascade classifier requires images of positive and negative cases of the object or figure to be detected, being necessary that all the images have the same size. Once trained it can then be used to detect the same type of objects in other images.

5.3.2. Implemented Solution

For the recognition of the person's face from a photograph, cascade sorters based on Haar filters have been used. This has been done using OpenCV (Open Source Computer Vision Library) [20], an open source software library with functions and algorithms for use in machine vision and machine learning, originally developed by Intel and introduced in 1999 [21].

OpenCV provides several trained models of classifiers (available at its GitHub page [22]) that can be loaded through the library itself. Three different models have been used for this application in order to maximise the probability of finding the person's face correctly in the analysed image. Specifically, the following models have been used: "haarcascade_frontalface_default", "haarcascade_frontalface_alt", and "haarcascade_frontalface_alt2". With them, the probability of finding the face is practically guaranteed, although some punctual errors can always arise due to shadows, blurs or turns. In this way, the three models are used to search for a face in the image, and it is saved if found. Finally, the photo of the face is cropped to keep only the area of interest and remove everything else. The resulting images are then introduced into the convolutional neural network for classification into one of the possible emotions.

5.4. Design of the Convolutional Neural Network (CNN)

After detecting the person's face in the photograph, the next step focuses on recognising the person's emotion based on their facial expression. To do this, the image is analysed using a convolutional neural network (CNN). The main characteristics and operation of this type of network are described below, as well as the details of the application of the neural network in the specific case of detecting the avatar's emotion.

5.4.1. Convolutional Neural Network

Description

An artificial neural network is a computational model inspired by the behaviour of networks of neurons in a brain. They consist of a set of units called neurons that transmit certain data to each other. The most famous type of network is the multilayer perceptron (MLP), which is divided into several layers of neurons connected one to the next layer. The initial information enters through the neurons of the input layer and is transformed through the intermediate or hidden layers. Each link has a value called weight that is multiplied by the value of the previous neuron and reaches the next one, where an operation is performed with all the incoming values. Finally, results are returned in the output layer. The weights of the links are adjusted during the process to give more importance to certain inputs that help in the classification.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2021 doi:10.20944/preprints202103.0189.v1

A convolutional neural network (CNN) is a type of artificial neural network designed to recognise visual patterns by mimicking the primary visual cortex of the brain. They are named after the mathematical concept of "convolution", which is a linear transformation of two functions into a new one representing the magnitude at which they overlap. This type of network is similar to MLP, but applied to two-dimensional arrays to classify or segment images, and is able to capture the spatial dependencies of the image through the use of various filters. It contains a hierarchy of layers that become more specialised, i.e. the first layers detect lines and curves while the last ones are able to recognise complex shapes such as an entire object.

Functioning

In order for a neural network to learn to recognise objects and shapes on its own, it must first be trained with a large number of images. In this way it will be able to pick up the most important features of the sample.

In the first layer, the pixel values of the image are taken as input (it is desirable that they are normalised between 0 and 1). That is, when entering the value of each one, if the image is 48 pixels wide and 48 pixels high, it would be necessary to have $48 \times 48 = 2,304$ entries. This is true for images that have only one colour channel (such as a black and white images). If they have three channels (such as an RGB colour images), $48 \times 48 \times 3 = 6,912$ entries would be needed.

Next, groups of neighbouring pixels are taken and operated on with a small matrix called *kernel*. Its size can be specified (e.g. 3×3) and it is filled with weight values. It cycles through all input values (from left to right and from top to bottom, moving one or more units for each step as specified), generating a new matrix which will be the next hidden layer. For example, for a 48×48 image (and with only one colour channel), a 3×3 kernel (with one shift unit) can generate a 46×46 matrix. An example of this transformation can be seen in Figure 5, using a 3×3 kernel moving from pixel to pixel in an original 5×5 image, resulting in a matrix of size 3×3. This process can be applied multiple times on the same layer, having a set of kernels (called a filter) that will produce several output matrices (this set is called *feature mapping*). Thus, a filter of 32 kernels would result in 32 output matrices.

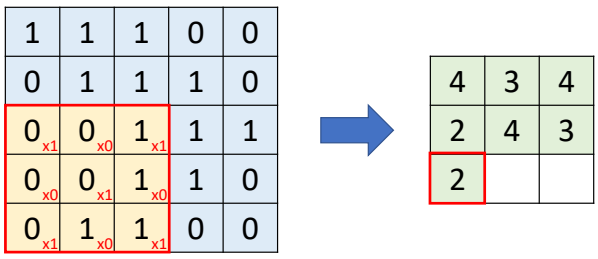

Figure 5. Example of kernel traversing an image.

It is possible that some values in the matrix turn out to be negative, so an activation function is used to rectify the data. The most commonly used is the ReLU (Rectified Linear Unit) function, which returns the maximum value between that data and 0. It is defined as: $f(x) = max(0, x)$.

Each of these new matrices represents specific features of the original image, which will help with object distinction later on. If we have generated 32 of them and each is 46 by 46, we would need $46 \times 46 \times 32 = 67,712$ entries for the next layer. The initial image is too small and yet too many values are formed in the first layer alone. Many of these are not even important, so it is advisable to reduce their number in order to lighten the processing in the next layers. This is done in the *subsampling* part, where a kernel-like process is performed by windowing through each matrix and forming a new, reduced one by keeping a smaller number of values. There are several types, such as *MaxPooling* (which

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2021 doi:10.20944/preprints202103.0189.v1

takes the highest values from each region) or the *AveragePooling* (which generates a new value from the average). In Figure 6 you can see how MaxPooling is performed with a window size of 2 by 2 (and with 2 displacement units) to a 4 by 4 matrix to reduce it by half.

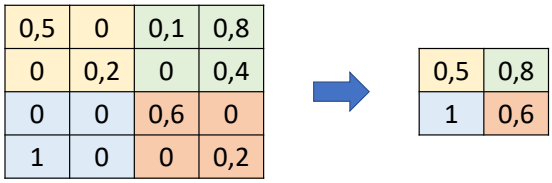

Figure 6. Example of subsampling with MaxPooling.

In this way a convolution is completed. As a summary of the previous steps: the image values are entered, the kernel filter is applied, the set of matrices or feature mapping is obtained, *MaxPooling* is applied and the reduced matrices are generated as a result. The values of the latter matrices would be used as input in the next convolution, repeating the whole process successively according to the number of convolutions carried out.

The output of the last convolution is then linked to another neural network (a fully connected multilayer perceptron). This is done by a layer called *Flatten*, which transforms the matrices into a vector so that their values can be used as inputs. In addition, a *Dropout*, a process that randomly discards a percentage of neurons to avoid overfitting during training, can be performed on each layer of neurons. Finally, in the last layer, an activation function called *Softmax* is run to convert the output into a probability distribution with as many values as there are classes. Once trained, classification is carried out in this part. When classifying, the output returns a set of data where each one represents the probability that the image belongs to that class.

It must be taken into account that the training of a convolutional neural network requires a lot of memory because a large number of images must be processed. Therefore, this process has to be done step by step, dividing the set of images into several batches or batches. Moreover, not only all data has to be passed once, but multiple times in order to improve feature capture. Thus, when all batches are passed until an iteration is completed, a *epoch* is said to be completed. Performing too few can cause the model to underfit the data while too many can lead to overfitting, so the right number must be found. However, there is no definitive solution as this number can vary greatly depending on the type of problem and the amount of data set.

5.4.2. Implemented Solution

The CNN network for emotion classification has been created using Keras [23], the deep learning Application Programming Interface (API) written in Python. It is the high-level API of TensorFlow 2 [24] (end-to-end open-source machine learning platform developed by Google) in which the model starts from zero and the layers that will form its structure must be added one by one. Among the layers that can be included are:

- Convolutional layers: These are responsible for deriving features from the spatial dependencies between pixels in the image, generating multiple filters that produce a feature map. Several of these layers are usually included (sometimes back-to-back) to capture as much information as possible. The first layers detect simple shapes such as lines and curves while the later layers are more specialised and can recognise complex shapes. However, it is not advisable to add too many layers because, at some point, they do not significantly improve the model and only increase its complexity and computational time.
- Subsampling layers (such as MaxPooling or AveragePooling): These are included after the convolutional layers to reduce the number of parameters generated and subsequently reduce the overfitting of the model.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2021 doi:10.20944/preprints202103.0189.v1

- Flatten layer: It converts the output of the convolutions into a vector used as the input of the final stage of the network, the fully connected layers.
- Fully connected layers (FCL): These are typically used to calculate probabilities and have an input layer, one or more hidden layers and an output layer.
- Dropout layers: These are placed between the fully connected layers to remove a percentage of their neurons and reduce overfitting.

It has been decided to add three sets of convolutional layers (the first with only one layer and the others with two) so that the model is able to detect a large amount of information from the images. After each set, a subsampling layer is attached to reduce the size of its parameters before they are introduced into the final fully connected layers. The diagram of Figure 7 depicts the structure of the convolutional neural network with all the layers that have been included, while Table 1 compiles the size of the outputs and the number of parameters generated in each one.

A short description of the layers that make up the CNN network is given below:

- "Conv (1)": The first convolutional layer has 64 5×5 kernels and uses ReLU as an activation function. Its input size is 48x48x1 as the input images are 48 pixels wide by 48 pixels high with only one colour channel (black and white).
- "MaxP": A 5×5 MaxPooling layer with 2 displacement units.
- "Conv (2)": A second convolutional layer with 64 3×3 kernels and ReLU as activation function.
- "Conv (3)": A third convolutional layer just like the previous one.
- "AvgP (1)": A first 3×3 AveragePooling layer with 2 displacement units.
- "Conv (4)": A fourth convolutional layer with 128 3×3 kernels and ReLU as activation function.
- "Conv (5)": A fifth convolutional layer identical to the previous one.
- "AvgP (2)": A second 3×3 AveragePooling layer with 2 displacement units
- "Flatten": A layer which takes the output from the convolutional layers and converts it to an input vector for the fully connected layers where the classification is finished.
- "Input (FCL)": The first fully connected layer with 1,024 neurons which takes the inputs from the feature analysis and applies weights to predict the correct label.
- "Drop (1)": A first Dropout layer to get rid of 20% of the neurons and reduce overfitting.
- "Hidden (FCL)": A hidden fully connected layer with the same number of neurons as the input.
- "Drop (2)": A second layer of Dropout with the same characteristics as the previous one.
- "Output (FCL) ": The output layer where a Softmax function is run to convert the output into a probability distribution of size 7 (equal to the number of classes to be classified, i.e. the six basic emotions plus neutral).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2021 doi:10.20944/preprints202103.0189.v1

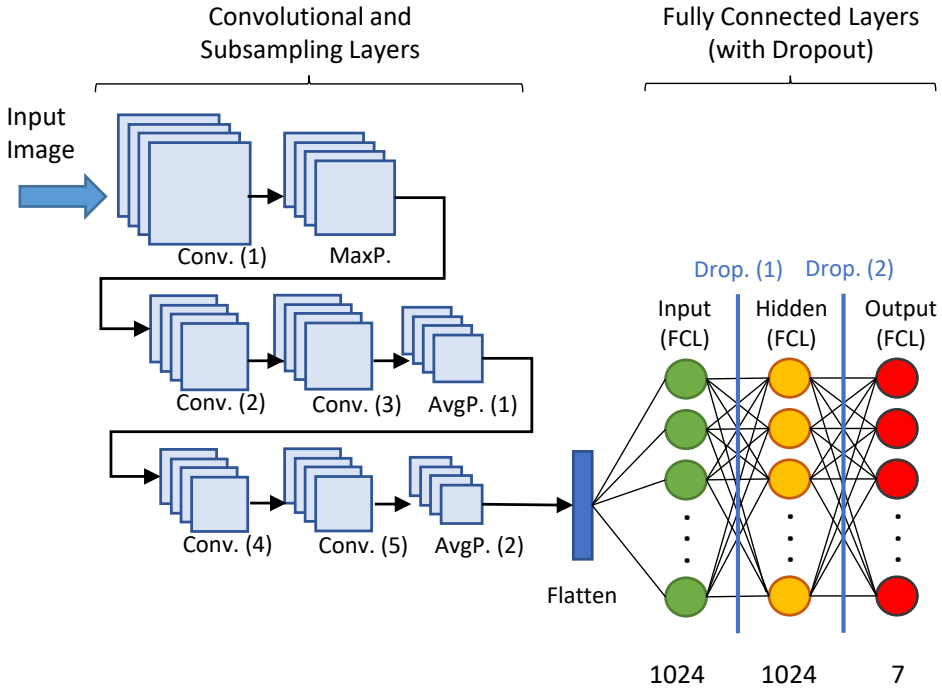

Figure 7. Structure of the CNN implemented.

Table 1. Details of the layers that make up the CNN.

Label	Type	Output Size	Params
Conv (1)	Convolutional	44×44×64	1 664
MaxP	MaxPooling	20×20×64	0
Conv (2)	Convolutional	18×18×64	36 928
Conv (3)	Convolutional	16×16×64	36 928
AvgP (1)	AveragePooling	7×7×64	0
Conv (4)	Convolutional	5×5×128	73 856
Conv (5)	Convolutional	3×3×128	147 584
AvgP (1)	AveragePooling	1×1×128	0
Flatten	Flatten	128	0
Input (FCL)	Fully Connected	1 024	132 096
Drop (1)	Dropout	1 024	0
Hidden (FCL)	Fully Connected	1 024	1 049 600
Drop (2)	Dropout	1 024	0
Output (FCL)	Fully Connected	7	7 175

5.5. Definition of Neural Network Training

A series of new virtual characters have been designed so that the CNN can learn to extract the basic characteristics of emotions from photos of their faces. If the network were trained with the characters included in the application (the male and female avatars shown above in Figure 3) then the model would end up over-adjusting to them, memorising their faces instead of learning to obtain the particularities of each expression independently of the person. It would not serve then to generalise and it would not work well if new characters were added.

For this reason, 10 new avatars have been created exclusively to use their photos in the training: 5 male and 5 female (they can be seen in Figure 8). All of them are very different from each other, varying in their features and skin colour. For each one,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2021 doi:10.20944/preprints202103.0189.v1

several photographs of their emotions have been taken, slightly modifying some features at random so that the image is not always the same (although without changing enough to be considered different emotions). In addition, they have been taken in different settings and varying the intensity of the light. This diverse database will help the network to focus on the most important features rather than unnecessary areas.

Figure 8. Set of avatars used for convolutional neural network training.

These images have been passed through the cascade sorter models to detect the face areas and cut them out. The size of the photos have a significant influence on the processing time of the convolutions, so they have been resized to 48 pixels wide by 48 pixels high (a small resolution but sufficient to distinguish the features). The number of colour channels also affect the processing time, although fortunately this is not a necessary element to detect facial expressions. Therefore, images are converted to black and white (single colour channel). This initial transformation of the images considerably lighten the computational load of the training.

Figure 9 shows the distribution of the samples with the amount of photos available for each emotion. There was a total amount of 8 566 images composing the training set. The differences in the number of images per emotion are related to problems with the initial detection of the face. Nevertheless, a similar proportion is maintained between the classes. This will help the model to learn all emotions in the same way and not to undermine any of them.

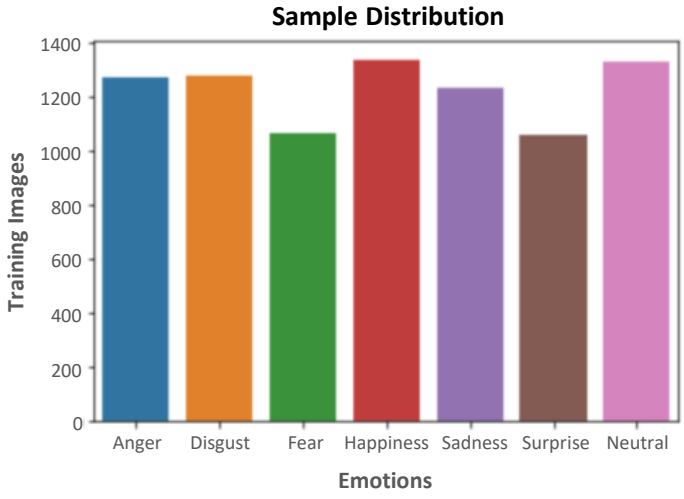

Figure 9. Distribution of samples used for training the convolutional neural network.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2021 doi:10.20944/preprints202103.0189.v1

From this data set, a stratified sample (with the same proportion of emotions as the main one) of 10% has been taken to be used as a validation of the model during the training (i.e. to check if it improves or worsens when trying different combinations of parameters). In this way, the model has been trained in batches of 100 images for 40 epochs. A success rate of 98.25% has been obtained for the training set and 89.64% for the validation set. Increasing the number of epochs did not improve the model, and over-adjustment occurred.

The resulting configuration was be saved in a file including: its architecture, the values of the weights learned, training configuration and its status (to follow the training process where it has been left off). This configuration file must be loaded at the start of the emotion recognition application in order to classify new images.

6. Experimental Results

After completing the development of the new ER module, several tests have been carried out to check the correct functioning of the module on the VR platform, which are detailed below. Firstly, a test is presented that focuses on the communication between the VR Visualiser and the ER System where it is verified that the emotion detection system correctly classifies the avatar's emotion in the image taken from the UAV camera in the virtual environment, while the second test analyses the performance of the emotion detection using a set of metrics.

6.1. Integration of the ER System into the VR platform

The following is an example to verify that the communication between the VR Visualiser and the ER System works as expected. After receiving the images from Unity and performing the classification process, the model generates a probability distribution for all the emotions. An example of the emotion detection interface is shown in Figure 10. The original image is in the upper left corner, underneath it is the face that has been detected (already transformed to enter the model) and on the right there is a graph with the percentages for each emotion. At the end of the classification, the recognition system returns a message with the detected emotion together with its percentage of certainty.

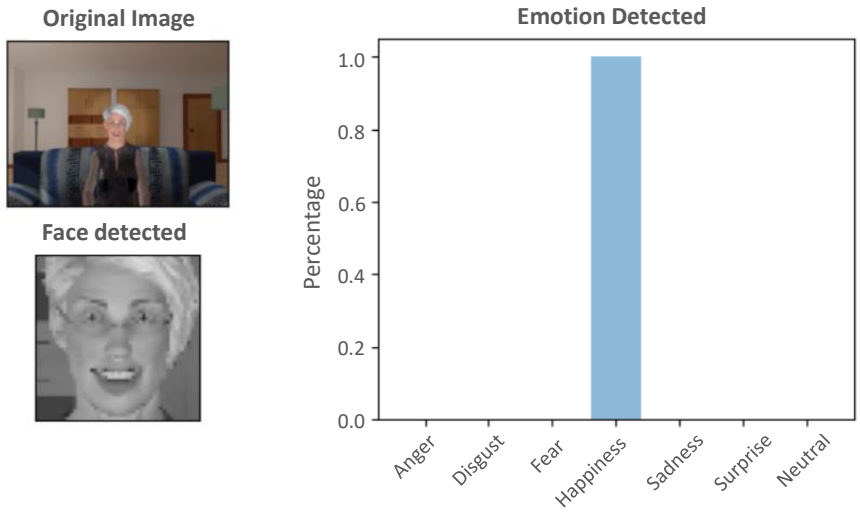

Figure 10. Test results displayed on the emotion detection interface.

Figure 11 shows the VR simulator implemented in Unity when the message with the information of the emotion is received. It is divided into different sections that show various views of the environment. On the UAV's camera section, the data received from the detection system being shown on the purple panel at the bottom. The application correctly displays the detected emotion together with its percentage, so the communication between both programs works as expected.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2021 doi:10.20944/preprints202103.0189.v1

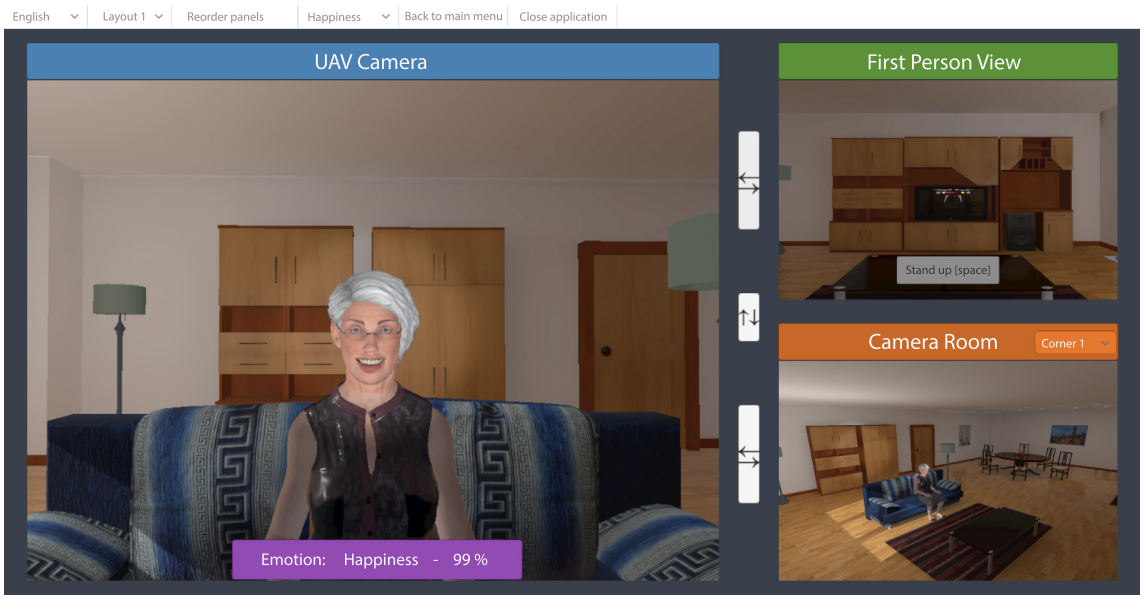

Figure 11. Test results displayed on the VR Visualiser interface.

6.2. Performance test of the emotion detector

From an initial set of 616 images (44 images per 7 classes per 2 avatars), 502 images with faces of the male and female avatar emotions were identify correctly by the face detection algorithm and were used to test the emotion classifier. It is important to remember that no images of these two characters have been used for the training of the convolutional neural network, so it will be the first time the CNN finds them. If they are correctly classified, it means that the neural network has successfully extracted the most important features from the training images and can generalise to new cases.

Confusion matrices have been used in order to analyse the performance of the classifier. The rows of the matrix represent the real classes (to which each case really belongs) and the columns indicate the classification done by the model. The main diagonal represents correct classifications, while other cells are classification errors. Thus, Figure 12 shows the confusion matrix for the test performed with a combination of the man and woman images. In addition, the right side of this image shows the *Recall or True Positive Rate (TPR)* and the *False Negative Rate (FNR)*. They are the percentages of images classified correctly (left column) and incorrectly (right column) for each emotion, respectively. Regarding anger, 100% of the images were correctly classified, while this percentage drops to 53% for the case of fear, the worst case. On the other hand, the *Precision or Positive Predictive Value (PPV)* and the *False Discovery Rate (FDR)*, that is, the percentages of correct (first row) versus incorrect (second row) predictions are shown at the bottom. In this case, the best result is for surprise, as 98.1% of the predictions about this emotion were correct, compared to 73.2% and 74.7% for the cases of anger and disgust.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 March 2021 doi:10.20944/preprints202103.0189.v1

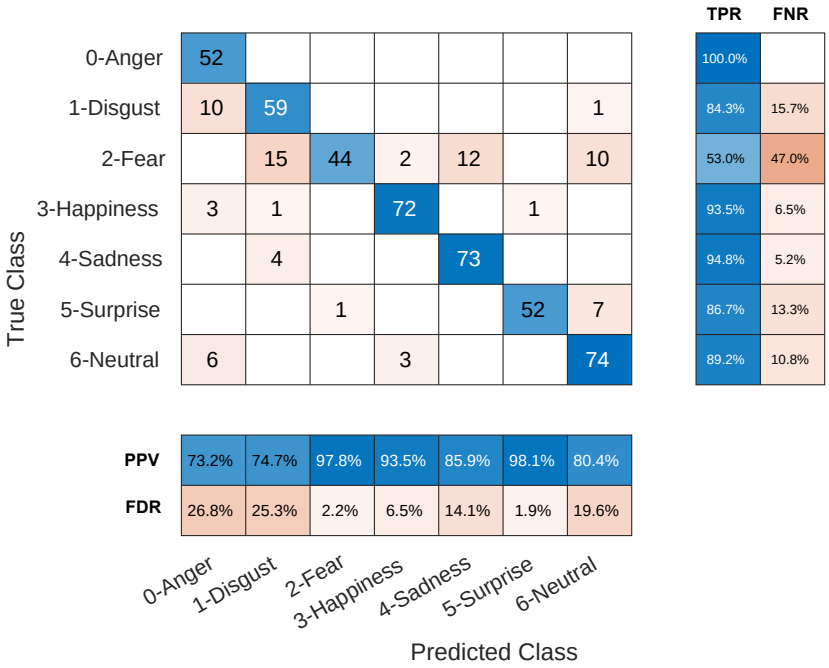

Figure 12. Confusion matrix for the tested image set.

To complete the data analysis, several metrics are summarised in Table 2; the *Accuracy (ACC)*, the above-mentioned *Recall (TPR)* and *Precision (PPV)*, the *Specificity or True Negative Rate (TNR)*, and the *F1 score*, the harmonic mean of precision and recall. All these parameters have been calculated for each class (emotion), and also globally, using the arithmetic mean and the weighted average (taking into account the number images of each class). According to the metrics, over 84% in all average parameters, it can be concluded that the model works satisfactorily and it is able to correctly classify the emotions in a high percentage of the cases.

Table 2. Model metrics for the tested image set.

Class	Images	Accuracy, ACC $\frac{TP+TN}{TP+TN+FP+FN}$	Recall, TPR $\frac{TP}{TP+FN}$	Precision, PPV $\frac{TP}{TP+FP}$	Specificity, TNR $\frac{TN}{TN+FP}$	F1 Score $\frac{2 \cdot PPV \cdot TPR}{PPV+TPR}$
0-Anger	52	96,22%	100,00%	73,24%	95,78%	84,55%
1-Disgust	70	93,82%	84,29%	74,68%	95,37%	79,19%
2-Fear	83	92,03%	53,01%	97,78%	99,76%	68,75%
3-Happiness	77	98,01%	93,51%	93,51%	98,82%	93,51%
4-Sadness	77	96,81%	94,81%	85,88%	97,18%	90,12%
5-Surprise	60	98,21%	86,67%	98,11%	99,77%	92,04%
6-Neutral	83	94,62%	89,16%	80,43%	95,70%	84,57%
Total	**502**	**-**	**-**	**-**	**-**	**-**
Average (AVG)	**-**	**95,67%**	**85,92%**	**86,23%**	**97,48%**	**84,68%**
Weighted AVG	**-**	**95,53%**	**84,86%**	**86,71%**	**97,53%**	**84,32%**

TP - True Positive; FP - False Positive; TN - True Negative; FN - False Negative

7. Conclusions and Future Works

The main long-term objective of our research project is the development of an assistant aerial vehicle for monitoring dependent people at home. This social robot will have the mission of flying every so often to take snapshots of the person in order to determine their

state and, based on this, the assistance or help needed. To determine the person's state, emotion analysis has been selected as the main technique in our proposal.

In the development of this project, we have designed a VR platform that allows us to have a safe environment in which to develop and validate the different solutions at an engineering level, as well as to carry out studies with participants in a realistic environment in order to adapt the solutions to the preferences of future users and their families. This VR platform is based on the real-time communication of three modules using the MQTT message protocol; the UAV Simulator implemented in MATLAB, the VR Visualiser developed in Unity and the new ER System programmed in Python.

The computer vision ER module is the great novelty of this work and is composed of two main parts: a set of cascade classifiers used to detect the face of the person in the image received from the camera on board the UAV, and a CNN designed for classifying the person's emotion in one of the six plus one basic emotions (surprise, fear, happiness, sadness, disgust, anger or neutral expression).

The integration of the new ER System within the VR platform has been evaluated and fulfils the expected function. The VR Visualiser sends the images captured by the on-board camera of the virtual quadrotor to the ER System, while the latter sends back the information about the avatar's emotion after the image analysis is finished. Apart from this, within the VR platform, the UAV Simulator and the VR Visualiser need to exchange information for the correct navigation of the UAV during the monitoring process. This part has not been directly evaluated in this paper, but can be found in [13,16].

On the other hand, the results of the tests carried out to evaluate the behaviour of the CNN demonstrate that the system is able to classify the emotions correctly in a great number of cases. The metrics used to measure the performance are higher than 84% in all the cases and the F1 score, which is a balance between the precision and recall, is near 85%. In this way, the results of the CNN can be considered satisfactory, however it should be noted that inaccuracies have been observed because the set of cascade classifiers sometimes fail to correctly detect the face of the person in the aerial images. Therefore, this makes the overall performance of the ER System decrease. This is one of the points of improvement for future work, to achieve a higher percentage of success in detecting the person's face so that more images can be analysed by the CNN.

Other points to debug and refine in order to improve the CNN results will be to increase the set of CNN training images with new characters, different degrees of emotions and with pictures taken without being fully frontal (with the head slightly rotated). In addition, it should be mention that the ER System could be used in the future implementation of the assistant UAV that monitors dependent people in a real environment, as the methodology used is valid for images of real people. However, some calibrations would probably be needed to fine-tune and adapt this module to a real environment.

Author Contributions: L.M.B., R.M. and A.F.-C. conceived the proposal. L.B.M. and R.M. designed and evaluated the proposed trajectory planning algorithm. A.M. and A.S.G. designed and evaluated the virtual reality environments and the computer vision algorithm. A.M. managed the study with participants. Additionally, L.M.B., A.S.G., R.M. and A.F.-C. analysed the data and participated in writing the paper.

Funding: This work was partially supported by VALU3S, a European co-funded innovation project that has been granted by the ECSEL Joint Undertaking (JU) [grant number 876852]. The funding of the project comes from the Horizon 2020 research programme and participating countries. National funding is provided by Germany, including the Free States of Saxony and Thuringia, Austria, Belgium, Finland, France, Italy, the Netherlands, Slovakia, Spain, Sweden, and Turkey. The Spanish co-funded innovation project has been granted by Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación (AEI) [grant number PCI2020-112001]. This work was also partially supported by Spanish Ministerio de Ciencia, Innovación y Universidades, Agencia Estatal de Investigación (AEI) / European Regional Development Fund (FEDER, UE) under EQC2019-006063-P and PID2020-115220RB-C21 grants, and by CIBERSAM of the Instituto de Salud Carlos III.

Conflicts of Interest: The authors declare no conflict of interest. The funding sources had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results

Abbreviations

The following abbreviations are used in this manuscript:

ACC	Accuracy
CNN	Convolutional Neural Network
ER	Emotion Recognition
FCL	Fully Connected Layer
FDR	False Discovery Rate
FN	False Negative
FNR	False Negative Rate
FP	False Positive
GPI	Generalised Proportional Integral
MLP	Multilayer perceptron
MQTT	Message Queue Telemetry Transport
PPV	Positive Predictive Value (or Precision)
TN	True Negative
TNR	True Negative Rate (or Specificity)
TP	True Positive
TPR	True Positive Rate (or Recall)
UAV	Unmanned Aerial Vehicle
VR	Virtual Reality

References

1. Fong, T.; Nourbakhsh, I.; Dautenhahn, K. A survey of socially interactive robots. *Robotics and Autonomous Systems* **2003**, *42*, 143–166. doi:10.1016/S0921-8890(02)00372-X.
2. Calderita, L.V.; Vega, A.; Barroso-Ramírez, S.; Bustos, P.; Núñez, P. Designing a Cyber-Physical System for Ambient Assisted Living: A Use-Case Analysis for Social Robot Navigation in Caregiving Centers. *Sensors* **2020**, *20*. doi:10.3390/s20144005.
3. Loza-Matovelle, D.; Verdugo, A.; Zalama, E.; Gómez-García-Bermejo, J. An Architecture for the Integration of Robots and Sensors for the Care of the Elderly in an Ambient Assisted Living Environment. *Robotics* **2019**, *8*. doi:10.3390/robotics8030076.
4. Sokullu, R.; Balcı, A.; Demir, E., The Role of Drones in Ambient Assisted Living Systems for the Elderly. In *Enhanced Living Environments: Algorithms, Architectures, Platforms, and Systems*; Ganchev, I.; Garcia, N.M.; Dobre, C.; Mavromoustakis, C.X.; Goleva, R., Eds.; Springer International Publishing: Cham, 2019; pp. 295–321. doi:10.1007/978-3-030-10752-9_12.
5. Wang, J.; Spicher, N.; Warnecke, J.M.; Haghi, M.; Schwartze, J.; Deserno, T.M. Unobtrusive Health Monitoring in Private Spaces: The Smart Home. *Sensors* **2021**, *21*. doi:10.3390/s21030864.
6. Lee, W.; Kim, J.H. Social Relationship Development between Human and Robot through Real-Time Face Identification and Emotional Interaction. Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction; Association for Computing Machinery: New York, NY, USA, 2018; HRI '18, p. 379. doi:10.1145/3173386.3177531.
7. Malliaraki, E. Social Interaction with Drones Using Human Emotion Recognition. Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction; Association for Computing Machinery: New York, NY, USA, 2018; HRI '18, p. 187–188. doi:10.1145/3173386.3176966.
8. Liu, S.; Watterson, M.; Mohta, K.; Sun, K.; Bhattacharya, S.; Taylor, C.J.; Kumar, V. Planning Dynamically Feasible Trajectories for Quadrotors Using Safe Flight Corridors in 3-D Complex Environments. *IEEE Robotics and Automation Letters* **2017**, *2*, 1688–1695.
9. Giansanti, D. The Social Robot in Rehabilitation and Assistance: What Is the Future? *Healthcare* **2021**, *9*. doi:10.3390/healthcare9030244.
10. Belmonte, L.M.; Morales, R.; García, A.S.; Segura, E.; Novais, P.; Fernández-Caballero, A. Assisting Dependent People at Home Through Autonomous Unmanned Aerial Vehicles. In *Advances in Intelligent Systems and Computing*; Springer International Publishing, 2019; pp. 216–223. doi:10.1007/978-3-030-24097-4_26.
11. Berni, A.; Borgianni, Y. Applications of Virtual Reality in Engineering and Product Design: Why, What, How, When and Where. *Electronics* **2020**, *9*. doi:10.3390/electronics9071064.
12. de la Cruz, M.; Casañ, G.; Sanz, P.; Marín, R. Preliminary Work on a Virtual Reality Interface for the Guidance of Underwater Robots. *Robotics* **2020**, *9*. doi:10.3390/robotics9040081.
13. Belmonte, L.; Garcia, A.S.; Segura, E.; Novais, P.J.; Morales, R.; Fernandez-Caballero, A. Virtual Reality Simulation of a Quadrotor to Monitor Dependent People at Home. *IEEE Transactions on Emerging Topics in Computing* **2020**, pp. 1–1. doi:10.1109/TETC.2020.3000352.

14. Castillo, P.; Dzul, A.; Lozano, R. Real-time stabilization and tracking of a four-rotor mini rotorcraft. *IEEE Transactions on Control Systems Technology* **2004**, *12*, 510–516. doi:10.1109/TCST.2004.825052.
15. Fernández-Caballero, A.; Belmonte, L.M.; Morales, R.; Somolinos, J.A. Generalized Proportional Integral Control for an Unmanned Quadrotor System. *International Journal of Advanced Robotic Systems* **2015**, *12*, 85. doi:10.5772/60833.
16. Belmonte, L.M.; García, A.S.; Morales, R.; de la Vara, J.L.; López de la Rosa, F.; Fernández-Caballero, A. Feeling of Safety and Comfort towards a Socially Assistive Unmanned Aerial Vehicle That Monitors People in a Virtual Home. *Sensors* **2021**, *21*. doi:10.3390/s21030908.
17. Ekman, P.; Friesen, W. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*; Consulting Psychologists Press: Palo Alto, CA, 1978.
18. García, A.S.; Fernández-Sotos, P.; Vicente-Querol, M.A.; Lahera, G.; Rodriguez-Jimenez, R.; Fernández-Caballero, A. Design of reliable virtual human facial expressions and validation by healthy people. *Integrated Computer-Aided Engineering* **2020**, *27*, 287–299. doi:10.3233/ICA-200623.
19. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, 2001, Vol. 1, pp. I–I. doi:10.1109/CVPR.2001.990517.
20. Bradski, G. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* **2000**.
21. Kaehler, A. *Learning OpenCV 3 : computer vision in C++ with the OpenCV library*; O'Reilly Media: Sebastopol, CA, 2016.
22. OpenCV (https://opencv.org/) - GitHub Page. https://github.com/opencv/opencv, 2021.
23. Chollet, F.; others. Keras. https://keras.io, 2015.
24. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard, M.; Jia, Y.; Jozefowicz, R.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mané, D.; Monga, R.; Moore, S.; Murray, D.; Olah, C.; Schuster, M.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P.; Vanhoucke, V.; Vasudevan, V.; Viégas, F.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y.; Zheng, X. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, 2015. Software available from https:/www.tensorflow.org.