Hindawi
Mobile Information Systems
Volume 2017, Article ID 7932529, 11 pages
https://doi.org/10.1155/2017/7932529

Research Article

Analyzing and Predicting Empathy in Neurotypical and Nonneurotypical Users with an Affective Avatar

Esperanza Johnson,[1] Ramón Hervás,[1] Carlos Gutiérrez-López-Franca,[1] Tania Mondéjar,[1,2] and José Bravo[1]

[1] *MAmI Research Lab, University of Castilla-La Mancha, Paseo de la Universidad 4, Ciudad Real, Spain*
[2] *eSmile, Psychology for Children & Adolescents, Calle Toledo 79 1°E, Ciudad Real, Spain*

Correspondence should be addressed to Ramón Hervás; ramon.hlucas@uclm.es

Received 7 February 2017; Revised 20 April 2017; Accepted 9 May 2017; Published 8 June 2017

Academic Editor: Jinglan Zhang

In recent times, diagnosing and treating different health issues have improved greatly with the help of technology, with an example being cognitive health issues. Despite this, there is still a difference between how the technology is working towards it and the actual potential that can be achieved. In this paper, we propose a mobile application with an affective avatar, encompassed in the area of serious games, which will obtain information related to the interactions performed by the users. There are a total of 50 users, of neurotypical and nonneurotypical backgrounds, with the latter being people with Down syndrome and intellectual disability. Based on collected data from the different users interacting with the avatar in a mobile device, we analyzed the results to obtain a ground truth about prototypic empathic interactions and feed those interactions to a learning algorithm to support the diagnosis process and therapy treatment of empathy and socialization issues.

1. Introduction

Communication is an important part of the human experience, and it is formed by a variety of processes, with emotions being among them. Because of this, affective computing has come from technologies that show or influence emotion to coach human behavior in regard to communication through emotions, so as to help any issues such as social communication disorders (SCD). These disorders are usually conformed by problems with social interaction and understanding [1]. With this in mind, assistive technologies can be used to help people lead more independent lives, by helping them with difficulties by improving their social skills and understanding of their emotional state.

In this paper, which is an extension of a previous one [2], we propose an affective avatar, which will engage the user by means of interaction through a tablet, having the result of a human-avatar interaction based on the different emotions shown by the avatar to the different interactions performed by the user. The interactions themselves will later be provided to a learning algorithm to detect empathic or not empathic interaction, which will in turn make it possible to assist in SCD diagnosis. We have decided on this approach as usual diagnosis of SCDs is performed through psychological assessments and observation, which require time and human intervention. On top of this, empathy issues can be hard to detect due to patients dealing with other pathologies which make it complicated to fulfill assessments or understand and follow the instructions to perform on those tests.

The paper will have the following distribution. In Section 2 we will discuss the works related to our own and what kind of knowledge we have extracted from them. Section 3 will discuss the avatar's design, the methodology followed to work on it, the psychological fundamentals, the design fundamentals, and the validation of the model. Section 4 will focus more on the conducted experiment with the users, as well as the extracted data and the results, whereas Section 5 will discuss how a machine learning algorithm was fed this data and what kind of algorithm was used. Section 6 will end this paper presenting the conclusions extracted from previous sections, as well as future work that will be done based on this model.

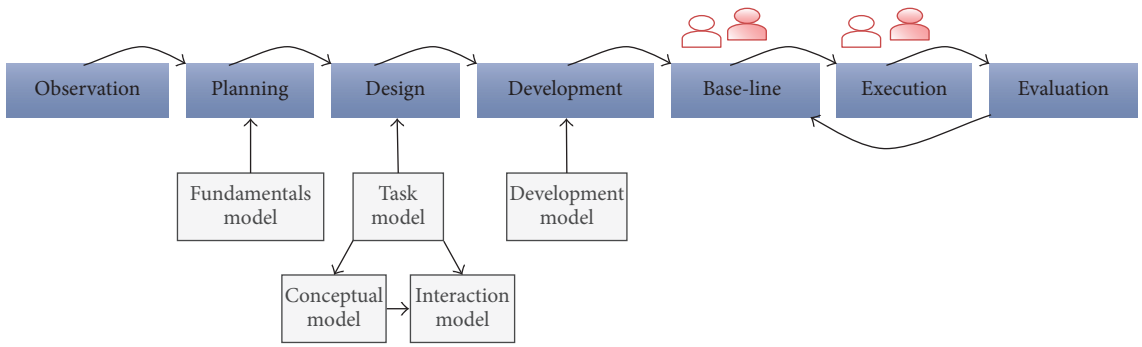


Figure 1: Methodology stages and involved submodels. Observation phase identifies the problems to address in the assistive system. Planning phase consists of (a) formal problem definition, (b) description of results to achieve, and (c) collecting background literature. Design stage obtains the design guidelines to create the assistive system. Development phase corresponds to the implementation of the system. Base-line stage collects information to deploy the initial trials. Execution phase gathers feedback from end users improving next evaluation stages. Evaluation phase assesses the suitability of the analysis system. The last three stages comprise the evaluation circle.

2. Related Work

Human-avatar interaction has been studied for several years, with it being applied to several different areas, such as assistance to older adults to conduct activities of daily living [3], assistance for people with severe motor disabilities [4], and the support of elderly people with mental disorders and/or physical disabilities [5]. It was determined that when the avatar is perceived as friendly by the users, they have a more positive reaction [6], as well as the preferences of children in regard to the design of the avatar [7]. Finally, other studies have looked into how to enact intelligent reactive behavior in avatar [8].

As for previous work in the area of machine learning, we have looked into proposals that aim to recognize user emotions. There are several ways to achieve this goal, among them, facial recognition through active shape models [9], histograms of oriented gradients [10], voice recognition based on support vector machines [11] and ontological mechanisms to identify emotions and human behavior [12], and EEG-based cognitive activity [13]. As for general applications in health, it can be used for physical rehabilitation [14] or prevention of health issues [15], as well as applications in game-based approaches [16].

Our contribution for human-avatar interaction would be the focus on its potential to assist in SCD evaluation, as related work did not look into analysis of human-avatar interaction with diagnostic purposes, and we are not aware of previous work on the application of avatars powered with machine learning techniques to assist in dealing with SCDs.

3. Avatar Design

For the design of the avatar we have used a user-centered approach which involved end users during the lifecycle of the project. This was done to get a better requirement acquisition and an effective evaluation and to, in theory, ensure the success of the final product.

Different types of experts have participated in this project, such as a psychologist, occupational therapists, and computer science engineers, who belonged to "eSmile" Psychology center for children and adolescents, Association of Down syndrome "Caminar," and the University of Castilla-La Mancha (Modelling Ambient Intelligence research group), respectively.

The methodology applied for this project was designed and applied in previous works [17], which focuses on a pipelines process that includes seven steps, and generates five models:

(i) Psychological model: focuses on the psychological fundamentals of the system.

(ii) Task model: encompasses activities supported by the system.

(iii) Conceptual model: concepts and the relationships between them, which are significant to the system.

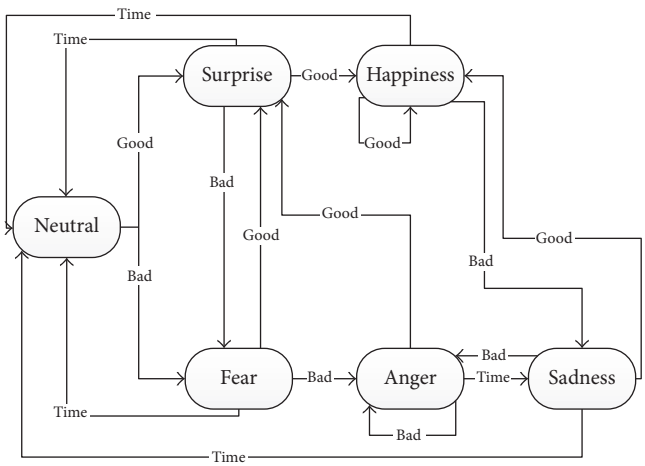
(iv) Interaction model: defines how the users can interact with the system and the feedback they receive.

(v) Development model: guidelines to the developer based on previous models.

This can be seen in Figure 1, which shows the seven steps of the process and the models involved in each step.

3.1. Psychology Fundamentals. The main goal of the avatar is to aid users in regard to affective cognitive process. The principles of affective cognitive function, like the use of mirror neurons and the theory of the mind, are psychological fundamentals of this work, and they are important research lines in SCD [1]. These theories will be important to our future work where the avatar will assist people with SCD to face their socialization and empathy issues, though this proposal focuses on the functionalities of the avatar to help the diagnosis of SCDs, so we will not further explore these theories at this time.

We also defined a taxonomy [18] to support the cognitive processes related to SCDs, and then we based ourselves in said taxonomy to adopt some of the interactions the prototype would have to support a particular cognitive process



Good (action) → caress
Bad (action) → poke
Time (no action) → wait

Figure 2: State machine representing the transition between emotions according to user input.

related to communication, which are emotional states and empathic behavior. We also employed this taxonomy to help to understand how the human-avatar interactions affect different cognitive processes, and to establish reference models to aid in the development of multi-interactive avatars for SCD.

3.2. Design Fundamentals. After designing the taxonomy, we decided to implement some aspects of it into our avatar. At this particular moment, the functionality that was implemented was tactile interaction, due to the nature of the device where the experiment would be taking place, a pc-tablet. The interactions are classified as good, bad, and Neutral, where a caress would be good, a poke would be bad, and waiting would be considered Neutral. With each interaction the user has with the avatar, its emotion will change, with the transition time varying from one emotion to the next when the user chooses to wait, according to difference in arousal between the emotions. The way interactions and emotions are tied can be seen in Figure 2.

The way the avatar portrays the emotions was through designing a three-dimensional model in Blender, which consisted of extending a two-dimensional plane over a three-dimensional axis to model the face, and the use of spheres and half-spheres for the eyes and hair, respectively. Other elements were added, such as a more complete version of the mouth with teeth and tongue, as well as eyebrows and other more superficial characteristics, to add realism and more human-like traits to the avatar.

To be able to animate the avatar, an armature was added, which had bones controlling the different vertices of the face. These bones are called drivers and are set in a hierarchy in the form of a tree, in which there is a Master bone, from which an EyeControl and a Head bone hang, and all the other facial bones hang from the Head bone. Shapes were assigned to the driver bones, with the objective of their area of control being more intuitive and clear at first glance.

The changes from a sketch model to the fully realized three-dimensional model can be observed in Figure 3.

The way the interactions and the avatar animations work is that the avatar's face will change with each interaction from the user to portray a different emotion, as seen in Figure 2. The range of emotions presented by the avatar is extracted from Paul Eckman's definition of basic emotions, and barring disgust for it not being relevant to the current iteration of the prototype, the emotions used are Happiness, Sadness, Anger, Surprise, Fear, and a Neutral state. These emotions are reflected on the avatar's face, which we decided to model as androgynously as possible, as one of the works we studied showed the gender preference of younger users [7]. While we tried to keep it as such, when translating the model from a two-dimensional sketch to a three-dimensional model, some of the androgynous characteristics were lost. The emotions as presented by the avatar, with some loss of quality from screen-capture, can be seen in Figure 4.

To achieve this, we imported the 3D model to Unity, taking advantage of Unity's Mecanim, which is the state machine that can be used to control animation cycles. Each animation created in Blender was assigned to a state in Mecanim, and the connectors have the actions that connect different states. This is done with a type of variable called *trigger* which, when combined with a C# script, enables the transition between the different emotions according to the actions detected by the tablet, which is done in the script. As Unity is a game engine that is oriented towards multiplatform options, we were able to easily export the file to an .apk to install in the tablet to conduct the experiments, an example of which can be seen in Figure 5.

3.3. Validation of the Model. A preliminary experiment was carried out to validate the model, with 30 participants, divided into groups by age (C1: under 12; C2: 12–21 and C3: 22–30), with 13 males and 17 females, with the range and cohorts determined by the usual ages for diagnosis and treatment of SCDs [19]. The experiment went as follows: participants were told about the purpose of the experiment, the information that would be collected, and what they would have to do as part of the experiment. The users would have to perform 20 interactions with the avatar, which would be in its starting emotion of Neutral. Every time the user interacted with the avatar, it would be recorded and they would be asked to identify the perceived emotion after each interaction, their own emotional response to the avatar's emotion, and if they thought the emotion shown by the avatar was logical considering the interaction they performed and the emotion before the interaction.

We extracted information from the results, such as overall accuracy, which in this case is the correct identification of the emotion shown by the avatar. Overall accuracy for the whole test was 73.69%, and when only taking into account the last 10 interactions it increased to 79.33%. We decided that differentiating between the total number of interactions and the last 10 interactions could provide useful information

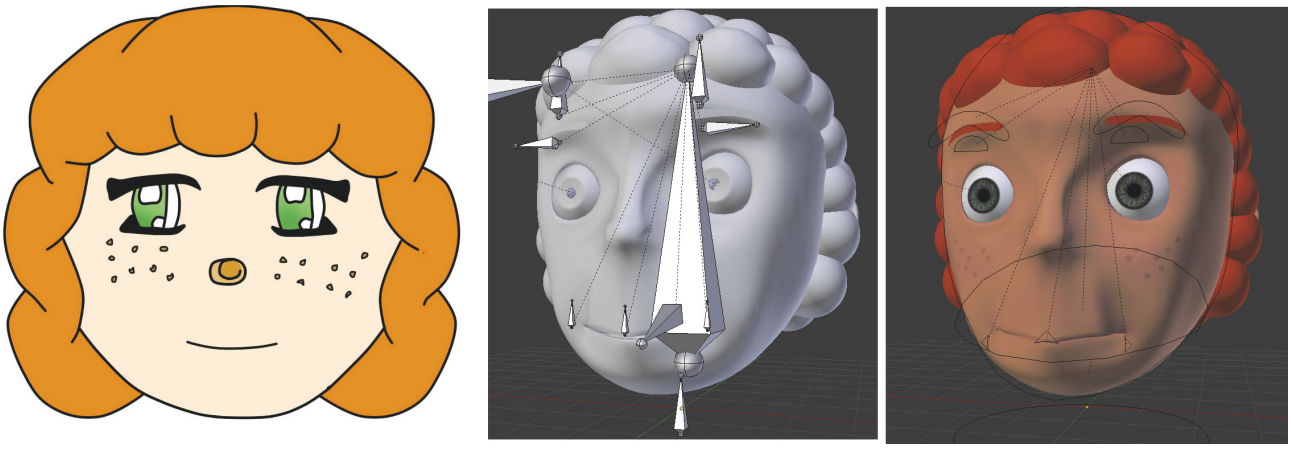


Figure 3: Transition of the model from a 2D sketch to the 3D model with armature and finally the model with the applied textures and visual changes to the armature that controls facial movement.

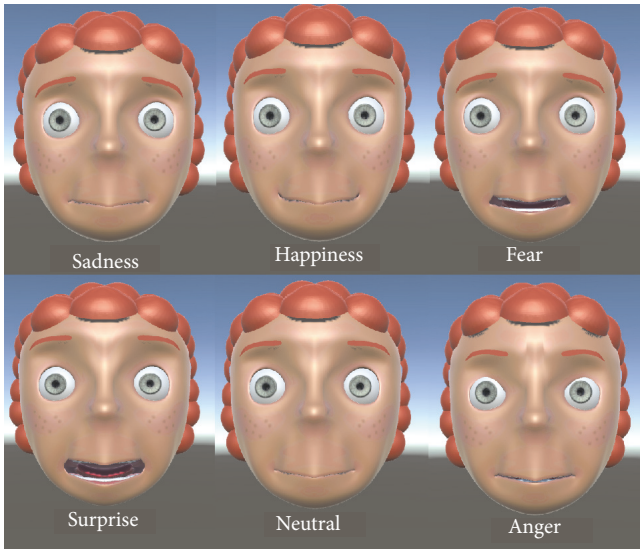


Figure 4: Final emotional states.

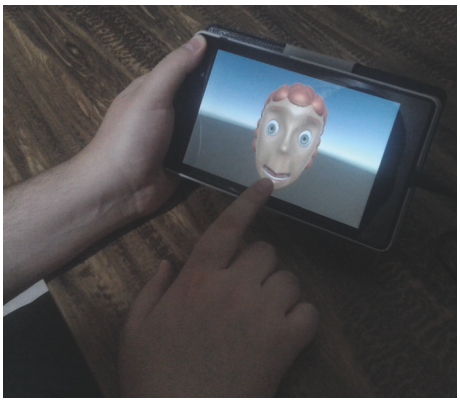


Figure 5: Avatar running in a Nexus 7 tablet, used for the experiments and validation of the model.

regarding any possible learning curve on the part of the users while they are getting used to the facial expressions of the avatar. Table 1, which works with preliminary data, shows the accuracy percentages for each emotion, both for 20 and for 10 interactions, in which the differences after the user has become more familiar with avatar can be seen.

We can see that Neutral is the worst in terms of recognition, and Happiness is the emotion with the best recognition results. We can also observe a general improvement after 10 interactions, which supports the theory that a need for familiarity with the avatar will improve the results. To be able to better observe the reason for these numbers, confusion matrixes for the results are in Table 2, extracted from the total number of times each emotion was shown by the avatar, divided into *successes* (the user correctly identified the emotion portrayed by the avatar) and *failures* (the user's prediction did not line up with the portrayed emotion). Successes are shown in the matrixes' diagonal (in bold), whereas failures are shown in the italic cells of the matrix, indicating an incorrect identification of the emotion. The row shows the emotion from which we are gathering the data, and the column is the emotion with which it was confused. For example, on a global scale, Neutral was correctly identified 98 times, and it was confused with Happiness 28 times. It shows any confusion between emotions, helping us better understand how the confusion occurred, giving us more information to improve the avatar.

As it can be seen in both matrixes, the emotion with most confusion is Neutral, which is confused with all the other emotions, though improvement can be seen after 10 interactions. It can be seen that it was confused with Happiness most often, but this does not hold up the other way around. This was due to the design of Neutral, which we approached to present itself in a generally friendly manner, which many people interpreted as Happiness, but once they saw Happiness presented in the avatar, they were quick to realize the difference. As it is, Happiness was the emotion with the least confusion, and it was just confused with Neutral, and

Table 1: Global accuracy percentages (top) and for the 10 last interactions (bottom) for each emotion.

	Neutral	Fear	Anger	Happy	Surprise	Sadness
Accuracy	59.88	65.26	61.90	94.16	82.05	73.68
Accuracy-10	65.15	75.56	64.10	94.27	92.00	79.31

Table 2: Global (top) and last 10 interactions (bottom) confusion matrixes.

	N	Fe	An	Ha	Su	Sa
Actual/predicted						
Neutral	**98**	*13*	*1*	*28*	*10*	*12*
	60.5%	*8.02%*	*0.62%%*	*17.28%*	*6.18%%*	*7.40%*
Fear	*0*	**62**	*1*	*1*	*3*	*26*
	0%	**66.67**%	*1.07%*	*1.07%*	*3.23%*	*27.96%*
Anger	*7*	*1*	**39**	*0*	*0*	*16*
	11.11%	*1.59%*	**61.9**%	*0%*	*0%*	*25.4%*
Happiness	*11*	*0*	*0*	**129**	*0*	*0*
	7.85%	*0%*	*0%*	**92.15**%	*0%*	*0%*
Surprise	*1*	*12*	*2*	*5*	**97**	*0*
	0.1%	*10.6%*	*1.8%*	*4.5%*	**83**%	*0%*
Sadness	*4*	*1*	*7*	*0*	*2*	**42**
	7.1%	*1.8%*	*12.5%*	*0%*	*3.6%*	**75**%
Actual/predicted - 10						
Neutral	**43**	*0*	*0*	*15*	*3*	*5*
	65.15%	*0%*	*0%*	*22.73%*	*4.54%*	*7.58%*
Fear	*0*	**34**	*1*	*0*	*1*	*9*
	0%	**75.6**%	*2.2%*	*0%*	*2.2%*	*20%*
Anger	*4*	*1*	**25**	*0*	*0*	*11*
	9.75%	*2.44%*	**60.98**%	*0%*	*0%*	*26.83%*
Happiness	*4*	*0*	*0*	**67**	*0*	*0*
	5.64%	*0%*	*0%*	**94.36**%	*0%*	*0%*
Surprise	*0*	*1*	*1*	*0*	**46**	*0*
	0%	*2.08%*	*2.08%*	*0%*	**95.84**%	*0%*
Sadness	*0*	*1*	*4*	*0*	*1*	**23**
	0%	*3.45%*	*13.79%*	*0%*	*3.45%*	**79.31**%

N (Neutral), Fe (Fear), An (Anger), Ha (Happiness), Su (Surprise), and Sa (Sadness).

it was still a small percentage compared to the times it was correctly identified.

From the accuracy results, the response by the users indicates that the emotions shown by the avatar were logical given the previous emotion and performed interaction. By the general feel of the users, we concluded that the presented model's emotions and state machine were adequate and proceeded to future experiments with this model. A more in depth study of all the results extracted from the data can be seen in previous work [19].

4. Experiment to Analyze Affective Interaction

Using an empirical approach, we had 18 more participants in an experiment just as the one previously explained, but with the participants being nonneurotypical, in this case, two groups of people, one with Down syndrome (Cn2, $n = 8$) and the other with intellectual deficiency (Cn1, $n = 10$), which means the total amount of data for the participants of this part of the experiment was 48 people, with the parameters of the experiment the same as the ones explained in Section 3.3. This also brings the total number of interactions to 960.

The overall accuracy of this group when identifying emotions is of 53.61%, much lower than the neurotypical group. The reason for such a low rate can be seen more clearly in Table 3, with the calculations done with preliminary data.

As it can be seen, Neutral has an extremely low accuracy rate, followed closely by Fear, whereas Anger and Happiness are the two with the best accuracy rates. It is interesting to note that while generally the percentages increase after the user has become familiar with the avatar it does not become true for Fear or Sadness. To better understand the reasons for the low percentages, we can observe the confusion matrixes for the emotions, given in Table 4.

The first thing we can address is the reason why accuracy did not seem to improve after 10 interactions for Fear, and we

Table 3: Accuracy in the recognition of the avatar emotions by nonneurotypical users.

	Neutral	Fear	Anger	Happy	Surprise	Sadness
Accuracy	12.77	25.58	72.92	69.70	47.69	72.00
Accuracy-10	21.43	17.65	76.00	69.86	60.87	64.29

Table 4: Global (top) and last 10 interactions (bottom) confusion matrixes for nonneurotypical users.

	N	Fe	An	Ha	Su	Sa
Actual/predicted						
Neutral	**6**	5	8	4	2	22
	12.77%	10.64%	17.02%	8.51%	4.26%	46.8%
Fear	1	**11**	6	0	0	25
	2.33%	**25.58**%	13.95%	0%	0%	58.14%
Anger	2	3	**35**	0	1	5
	4.35%	6.52%	**76.09**%	0%	2.17%	10.87%
Happiness	10	7	7	**92**	15	1
	7.58%	5.3%	5.3%	**69.7**%	11.36%	0.76%
Surprise	2	13	2	11	**31**	5
	3.13%	20.31%	3.13%	17.19%	**48.44**%	7.8%
Sadness	0	1	5	0	0	**18**
	0%	4.17%	20.83%	0%	0%	**75**%
Actual/predicted - 10						
Neutral	**6**	2	4	2	1	13
	21.43%	7.14%	14.29%	7.14%	3.57%	46.43%
Fear	0	**3**	1	0	0	13
	0%	**17.65**%	5.88%	0%	0%	76.47%
Anger	0	2	**19**	0	0	2
	0%	8.7%	**82.6**%	0%	0%	8.7%
Happiness	9	5	2	**51**	7	0
	12.16%	6.76%	2.7%	**68.92**%	9.46%	0%
Surprise	0	5	0	4	**14**	0
	0%	21.74%	0%	17.39%	**60.87**%	0%
Sadness	0	0	4	0	0	**9**
	0%	0%	30.77%	0%	0%	**69.23**%

can observe that while the number of errors has decreased somewhat so has the number of times it was correctly identified. As for Neutral, it can be seen that it is confused with all the other emotions, most frequently Sadness, and the number of times it was correctly identified stays the same in both cases, which is the reason for the extremely low accuracy percentage, a similar problem to Fear's low accuracy percentage. We can see that Anger and Happiness have confusion, but Anger is less prone to be confused with other emotions, resulting in a higher accuracy rate than Happiness, which is often confused with Surprise of Neutral.

When looking at the interactions the users performed, they showed a general inclination towards caressing the avatar, though some groups had a greater preference than others towards this kind of interaction, as can be seen in Table 5.

As it can be seen in Table 5, groups C1, Cn1, and Cn2 behaved similarly when it came to performing interactions, where they did not opt to wait as frequently as performing an interaction. This contrasts with groups C2 and C3, which have a more even distribution of the performed interactions.

Given the nature of the amount of interactions performed by each group, we proceeded to analyze more closely how the users interacted with the avatar, and if there was any tendency towards the repetition of those interactions. For this, we decided on classifying 4 consecutive interactions of the same kind as a single repetitive interaction, and adding a unit for every time the interaction was repeated after a repetitive interaction. This classification was done after analyzing the data, and observing a tendency towards repeating an action 3 times across all groups of users, but any more than that was unusual in users without an SCD. Table 6 shows the collected data and how it differs between groups.

The most notable conclusion that can be drawn from observing the table is that nonneurotypical users have more repetitive interactions than neurotypical, despite the difference in number of users. Upon closer observation, it can be seen that caressing is the interaction that has the greatest

Table 5: Total actions done by neurotypical users (above) and non-neurotypical users (below), as well as by cohort. Between parenthesis is the percentage for how often each interaction was performed.

	Poke	Caress	Wait
C3	64 (32%)	70 (35%)	66 (33%)
C2	72 (36%)	70 (35%)	58 (29%)
C1	69 (34.5%)	113 (56.5%)	18 (9%)
Total	205 (34.2%)	253 (42.2%)	142 (23.6%)
Cn2	71 (35.5%)	97 (48.5%)	32 (16%)
Cn1	42 (26.3%)	93 (58.1%)	25 (15.6%)
Total	113 (31.4%)	190 (52.8%)	57 (15.8%)

Table 6: Repetitive interactions made by the users divided by groups (NT: neurotypical, NNT: nonneurotypical) and cohorts, per interaction, as well as the percentages of those repetitive interactions over the total interactions.

	Poke	Caress	Wait
NT-C3	2 (1%)	2 (1%)	5 (2.5%)
NT-C2	1 (0.5%)	0 (0%)	0 (0%)
NT-C1	1 (0.5%)	24 (12%)	0 (0%)
NT-total	*4 (0.6%)*	*26 (4.3%)*	*5 (0.8%)*
NNT-Cn2	13 (8.1%)	20 (12.5%)	0 (0%)
NNT-Cn1	9 (4.5%)	15 (7.5%)	2 (1%)
NNT-total	22 (6.1%)	35 (9.7%)	2 (0.5%)
Total	*26 (2.7%)*	*61 (6.4%)*	*7 (0.7%)*

number of repetitive interactions, followed by poking, though the increase is more notable in poking ($4 \rightarrow 26$) than in caressing ($26 \rightarrow 35$). Neurotypical users in the youngest cohort (C1) are the ones which add to the number of repetitive interactions for the groups of neurotypicals, as the older cohorts tend not to perform repetitive interactions, and even then, the oldest cohort (C3) evenly distributes these kinds of interactions.

On the other hand, nonneurotypical users rarely wait enough for it to be considered a repetitive interaction, and they prefer direct interaction with the avatar, repeating the same interaction several times. They also prefer caressing to poking, but there is not as big of a difference as compared to group C1.

5. Learning Algorithm to Infer Empathy

After obtaining the data from the previous section and basing ourselves in the extracted ground truth on the difference between neurotypicals and nonneurotypicals when it comes to interacting with the avatar, we decided on implementing an algorithm that could be trained with the gathered data to help the diagnosis of SCDs. For this part of the experiment, we had 30 neurotypical users and 20 nonneurotypical users classified as explained in Section 4. Therefore, we have a total of 50 participants and 1000 interactions from the users to use as starting data to train an algorithm.

We used a logistic regression model that has been built to predict whether a set of interactions presents an empathic behavior or not. As this is a simple classification problem, we have just two output classes: 1 when empathic and 0 when not empathic.

We selected a learning algorithm based on the characteristics of the problem, such as a small dataset of training examples ($m = 50$) and a large number of features, which in this case are interactions with the avatar and the corresponding reaction ($n = 40$). Logistic regressions and support vector machines (SVM) have been proven successful in these kinds of problems, hence ruling out other options that are not suitable, such as neural networks or lineal regression with normal equation.

The hypothesis in logistic regression is defined as a sigmoid function of the model parameters (θ) and the system input (x), shown in

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T \cdot x}}. \tag{1}$$

The cost function of our logistic regression algorithm determines how well the algorithm maps the output. The cost function in logistic regression regarding the ith training example follows the formula shown in

$$\begin{aligned} &\mathrm{Cost}\left(h_\theta\left(x^{(i)}, y^{(i)}\right)\right) \\ &\quad = \begin{cases} -\log\left(h_\theta(x)\right), & \text{if } y = 1 \\ -\log\left(1 - h_\theta(x)\right), & \text{if } y = 0. \end{cases} \end{aligned} \tag{2}$$

The cost function based on the whole training set can be computed as shown below in

$$\begin{aligned} J(\theta) = -\frac{1}{m}\Bigg[&\sum_{i=1}^{m} y^{(i)} \log h_\theta\left(x^{(i)}\right) \\ &+ \left(1 - y^{(i)}\right) \log\left(1 - h_\theta\left(x^{(i)}\right)\right)\Bigg]. \end{aligned} \tag{3}$$

The suitability of the logistic regression model is achieved through the minimization of the cost function. This process has been performed through the named gradient descent

with a learning rate $\alpha = 0.1$ (obtained experimentally using convergence tests). This method consists in finding the global minimum of a function in terms of θ parameters. The gradient descent was implemented using an iterative algorithm to compute, shown in

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right) x_j^{(i)}. \quad (4)$$

5.1. Algorithm Optimization Decision. Since the optimization of parameters using gradient descent provided acceptable convergence results, it returned a local minimum. For this reason, we used an optimization solver provided by MATLAB and called *fminunc* to optimize the cost function $J(\theta)$ with parameters θ. This optimization function, running for 200 steps, gives the optimal parameters of θ.

Additionally, the characteristics of this problem, as mentioned before, are that a large number of features and small data set typically can entail an overfitted model. To reduce this problem, the algorithm was provided with regularization mechanisms. Equation (5) shows the formula to regularize the cost function and (6) the regularization of the gradients obtained with *fminunc:*

$$J(\theta)\,\text{reg} = J(\theta) + \frac{\lambda}{2m} \sum_{j=1}^{n} \theta_j^2, \quad (5)$$

$$\theta_j \text{reg} = \theta_j + \frac{\lambda}{m} \theta_j. \quad (6)$$

Finally, it is important to remark that scaling mechanisms were not necessary. However, the data collected in the experiment were discrete inputs, both the kind of interaction performed by users (wait, caress, and poke) and the resulting emotions (Anger, Fear, Surprise, Happiness, Sadness, and Neutral). For this reason and in order to better handle the data, these inputs were "binarized"; that is, the mentioned categorical features were represented as multiple Boolean features.

Also, due to the nature of the algorithm and the small training set we had available, we decided on creating new data from scratch based on the real dataset. Because each user performed 20 interactions with the avatar, each user test was divided into four sets of five sequenced interactions (1–5, 6–10, 11–15, and 16–20), which were shuffled generating 23 new training examples for user tests. The synthetic dataset now has 1150 new training examples, 1200 including the original dataset, so the total number of interactions analyzed was 2400. The new dataset can include some noise as we are losing one particular characteristic of the user interaction: the user needs the first interactions to get used to the avatar's behavior. However, the results obtained and explained seem to denote that this characteristic is not critical in the algorithm's performance.

5.2. Results. The training examples were split into two groups: a training set which had 80% of the examples of each cohort, and a test set with the remaining 20% to evaluate the algorithm. The data of a particular individual is always part of one, and only one, set (training or testing, never part of both). This restriction ensures that the real data of a user does not train the algorithm that will be tested with a shuffled set of the same data, which can corrupt the evaluation. We evaluated the algorithm's performance using precision, recall, and F-score, taking into account true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) in terms of classification under empathy-related disorder (1) or neurotypical users (0).

Table 7: Table with average precision, recall, and f-score values for the two types of synthetic data.

	Performance metrics		
	Precision	Recall	F-score
Free-shuffled synthetic data	0.69	0.91	0.78
Restricted synthetic data	0.66	0.82	0.72

Several combinations of training sets and test sets of the real data were generated and used to evaluate the algorithm. The performance results were very diverse, giving a perfect fit in some cases and very bad results in others, which was expected due to the small training set that resulted in an overfitted model with some kind of training and set examples. This evaluation process gave us an interesting finding, which showed most of the classification errors corresponded to young users from C1, at around 6 years of age, and which the algorithm classified as empathy-related disorders, a result that will be discussed more in depth later.

The experiments using synthetic data resulted in more conclusive scores, and we evaluated the algorithm with two kinds of synthetic data:

(i) Free-shuffled synthetic data: no restrictions were applied to the dataset; thus any sequence of interaction (real or synthetic) can be finally used as part of the training set or the testing set.

(ii) Restricted synthetic data: in this case, a subset of 80% of users forms the training set and the other 20% will be the testing set. In other words, the real data and synthetic data of a particular individual is always part of one, and only one, set (training or testing, never part of both). This restriction ensures that the real data of a user does not train the algorithm that will be tested with a shuffled set of the same data, which can corrupt the evaluation.

Table 7 shows averaged precision, recall, and f-score values of experiments with combinations of free-shuffled synthetic data on one hand and restricted synthetic data on the other.

Tables 8 and 9 show the confusion matrixes of the whole test set from the free-shuffled and the restricted synthetic data, respectively. Each cohort's confusion matrix is shown individually for a better visualization of how the global confusion matrix came to be. As we explained above, C1, C2, and C3 correspond to the three sets of neurotypical users grouped by age; Cn1 corresponds to the cohort of Down syndrome, and Cn2 was the cohort of intellectual deficiency.

Table 8: Confusion matrixes using free-shuffled synthetic data

		Confusion matrixes											
		Global		C1(0)		C2(0)		C3(0)		Cn1(1)		Cn2(1)	
TP	FP	84	37	0	18	0	10	0	9	44	0	40	0
FN	TN	8	101	0	28	0	36	0	37	2	0	6	0

Table 9: Confusion matrixes using restricted synthetic data.

		Confusion matrixes											
		Global		C1(0)		C2(0)		C3(0)		Cn1(1)		Cn2(1)	
TP	FP	76	40	0	19	0	12	0	9	39	0	37	0
FN	TN	16	98	0	27	0	34	0	37	7	0	9	0

It is important to remark that the expected output of the learning algorithm to neurotypical users should be "Negative" and in the case of Cn1 and Cn2, should be "Positive" in general. This is because the algorithm aims to identify social communication disorders. Consequently, result retrieved to C1, C2, and C3, that is, neurotypical users, only can be classified as False-Positive (FP), when the user is erroneously classified as person with SCD and True-Negative (TN) if the user was correctly classified. In the case of Cn1 and Cn2, the only possible situations are True-Positive (TP) if the user was correctly classified as SCD person or False-Negative (FN) in the other case.

Focusing on the restricted synthetic data, the obtained results show higher errors when classifying neurotypical users of cohort C1 where 41% of test examples were erroneously classified as empathy-related disorder, most of them users aged around 6 years old. This can be due to the original data collected, in which it was seen that, after a certain amount of time (usually after 10 interactions), there were several cases in which the children had finally figured out the action that made the avatar smile and therefore continued to repeat the action, which coupled with the logical/not logical reaction data collected from children meant that we have a considerable amount of data from one group (C1) that has several more instances of repetitive interactions and emotions. This coupled with the logical/not logical question that had to be simplified, as some young children had difficulty understanding the concept of logical behavior, meant that there were instances of FP, which in our study were incorrectly identified emotions, but where the children claimed the reaction of the avatar as logical. On the other hand, examples from people with intellectual disorders resulted slightly worse classified (19%) than people with Down syndrome (15%).

6. Conclusions and Future Work

In this paper we present an interactive avatar, able to collect information about affective management in terms of interactions. The performed experiments allowed us to collect data with neurotypical and nonneurotypical users with SCD related to empathy and emotion management. This data can establish a ground truth about empathic interactions with avatars and allow the training of learning algorithms to identify empathic and not empathic behavior.

We can conclude several things from the different points seen throughout this paper:

(i) The state machine that controls the avatar's transitions between emotions was accepted by the users as a logical representation of reactions towards their interactions.

(ii) Younger users will lean more towards direct interaction, as well as users with SCDs. They will also be more prone towards repetitive interactions.

(iii) Context was important to more accurately understand the portrayed emotion by the avatar. In some occasions, when seeing Happiness after Neutral, the emotion became clearer than when it happened after other emotions, such as seeing Neutral after Sadness.

(iv) The use of synthetic data based on shuffling a set of sequenced interactions keeps the nature of the data and seems to be a good solution where the real dataset is not large enough.

(v) There is a significantly worse performance on the classification when analyzing interactions of children under 7 years old. This issue questions the suitability of this kind of system for very young people, at least in terms of diagnosis.

(vi) In general, the classification of people with empathy and socialization difficulties performed well. Performance can be worse for people with intellectual deficiency than with Down syndrome, with the main reason seeming to be the diversity of pathologies and capabilities of the first group.

(vii) Given the results and the kind of problem presented, we have seen the limitations of fully automatic machine learning algorithms in cases where there is little training data, though there are approaches that seek to solve this issue [20, 21].

In general, this paper contributes to the field of serious games for health, in particular, those games that help to diagnose and treat cognitive issues. The final aim is to make invisible the psychological tools into videogames (an avatar-driven game in this case) in order to make this task easier and more enjoyable for the patient, and to reduce external factors that typically occur in traditional therapies (unfriendly environment, exhausting and complex tests, etc.).

Future work will focus on increasing the dataset and improving the information of each subject with quantitative data about empathy and socialization. We would also like to include a wider array of tools to help improve the interactive capabilities of our proposal, such as EEG headset, or include overall movement of the body [22].

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

Special thanks are due to "Down Caminar Association" and all families involved in this project.

References

[1] J. Gibson, C. Adams, E. Lockton, and J. Green, "Social communication disorder outside autism? A diagnostic classification approach to delineating pragmatic language impairment, high functioning autism and specific language impairment," *Journal of Child Psychology and Psychiatry and Allied Disciplines*, vol. 54, no. 11, pp. 1186–1197, 2013.

[2] R. Hervás, E. Johnson, C. Gutiérrez López de la Franca, J. Bravo, and T. Mondéjar, "A learning system to support social and empathy disorders diagnosis through affective avatars," in *Proceedings of the International Conference on Ubiquitous Computing and Communications and 2016 International Symposion on Cyberspace and Security (IUCC-CSS)*, IEEE, Granada, Spain, December 2016.

[3] S. Hanke, E. Sandner, A. Stainer-Hochgatterer, C. Tsiourti, and A. Braun, "The technical specification and architecture of a virtual support partner," in *Proceedings of the European Conference on Ambient Intelligence*, vol. 1528, November 2015.

[4] M. Adjouadi, A. Sesin, M. Ayala, and M. Cabrerizo, "Remote eye gaze tracking system as a computer interface for persons with severe motor disability," in *Computers Helping People with Special Needs*, vol. 3118, pp. 761–769, Springer, Berlin, Germany, 2004.

[5] H. Plischke and N. Kohls, "Keep it simple! assisting older people with mental and physical training," in *Universal Access in Human-Computer Interaction. Addressing Diversity*, vol. 5614, pp. 278–287, Springer, Berlin, Germany, 2009.

[6] D. M. Cereghetti, S. Kleanthous, C. Christophorou, C. Tsiourti, C. Wings, and E. Christodoulou, "Virtual partners for seniors: analysis of the users' preferences and expectations on personality and appearance," in *Proceedings of the European Conference on Ambient Intelligence*, vol. 1528, November 2015.

[7] Y. Inal, H. Sancar, and K. Cagiltay, "Childrens avatar preferences and their personalities," in *Proceedings of the Society for Information Technology & Teacher Education International Conference*, Fla, USA, March 2006.

[8] A. Iglesias and F. Luengo, "AI framework for decision modeling in behavioral animation of virtual avatars," in *Computational Science*, Y. Shi, G. D. van Albada, J. Dongarra, and et al, Eds., vol. 4488 of *Lecture Notes in Computer Science*, pp. 89–96, Springer, Berlin, Germany, 2007.

[9] E. Lozano-Monasor, M. T. López, A. Fernández-Caballero, and F. Vigo-Bustos, "Facial expression recognition from webcam based on active shape models and support vector machines," in *Ambient Assisted Living and Daily Activities*, vol. 8868, pp. 147–154, Springer International Publishing, 2014.

[10] P. Guerrero, M. Pavez, D. Chávez, and S. F. Ochoa, "Landmark-based histograms of oriented gradients for facial emotion recognition," in *Ambient Assisted Living. ICT-based Solutions in Real Life Situations*, vol. 9455, pp. 288–299, Springer International Publishing, December 2015.

[11] V. Rojas, S. F. Ochoa, and R. Hervás, "Monitoring moods in elderly people through voice processing," in *Ambient Assisted Living and Daily Activities*, vol. 8868, pp. 139–146, Springer International Publishing, December 2014.

[12] O. Banos, J. Bang, T. Hur et al., "Mining human behavior for health promotion," in *Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2015*, pp. 5062–5065, ita, August 2015.

[13] T. Mondéjar, R. Hervás, E. Johnson, C. Gutierrez, and J. M. Latorre, "Correlation between videogame mechanics and executive functions through EEG analysis," *Journal of Biomedical Informatics*, vol. 63, pp. 131–140, 2016.

[14] I. Raso, R. Hervás, and J. Bravo, "M-Physio: personalized accelerometer-based physical rehabilitation platform," in *Proceedings of the 4th International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies UBICOMM*, pp. 416–421, Florence, Italy, 2010.

[15] R. Hervás, J. Fontecha, D. Ausín, F. Castanedo, D. López-de-Ipiña, and J. Bravo, "Mobile monitoring and reasoning methods to prevent cardiovascular diseases," *Sensors*, vol. 13, no. 5, pp. 6524–6541, 2013.

[16] A. Holzinger, "Interactive Machine Learning (iML): a challenge for Game-based approaches," in *Proceedings of the Challenges in Machine Learning: Gaming and Education*, vol. 39 of *NIPS CiML Workshop*, Barcelona, Spain, 2016.

[17] J. Fontecha, R. Hervás, T. Mondéjar, I. González, and J. Bravo, "Towards context-aware and user-centered analysis in assistive environments:: a methodology and a software tool," *Journal of Medical Systems*, vol. 39, no. 120, 2015.

[18] E. Johnson, R. Hervás, T. Mondéjar, J. Bravo, and S. F. Ochoa, "Improving social communication disorders through human-avatar interaction," in *Proceedings of the Ambient Intelligence for Health*, vol. 9456 of *Lecture Notes in Computer Science*, pp. 237–243, Springer International Publishing, December 2015.

[19] E. Johnson, R. Hervás, C. Gutiérrez López de la Franca, T. Mondéjar, S. F. Ochoa, and J. Favela, "Assessing empathy and managing emotions through interactions with an affective avatar," *Health Informatics Journal*, 2016.

[20] A. Holzinger, "Interactive machine learning for health informatics: when do we need the human-in-the-loop?" *Brain Informatics*, vol. 3, no. 2, pp. 119–131, 2016.

[21] A. Holzinger, M. Plass, K. Holzinger, G. Crisan, C. Pintea, and V. Palade, "Towards interactive machine learning (iml): applying ant colony algorithms to solve the traveling salesman problem with the human-in-the-loop approach," in *Availability, Reliability, and Security in Information Systems*, vol. 9817 of *Lecture Notes in Computer Science*, pp. 81–95, Springer International Publishing, Berlin, Germany, 2016.

[22] C. Gutiérrez López de la Franca, R. Hervás, E. Johnson, T. Mondéjar, and J. Bravo, "Extended body-angles algorithm to recognize activities within intelligent environments," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–19, 2017.