JEMImE: a Serious Game to Teach Children with ASD How to Adequately produce Facial Expressions

Arnaud Dapogny*, Charline Grossard †, Stephanie Hun‡, Sylvie Serret‡, Jérémy Bourgeois‡, Hedy Jean-Marie§, Pierre Foulon§, Huaxiong Ding¶, Liming Chen¶, Severine Dubuisson‖, Ouriel Grynszpan*, David Cohen† and Kevin Bailly*

*Sorbonne Université, CNRS, ISIR, F-75005 Paris, France
†AP-HP, Groupe Hospitalier Pitié-Salpêtrière, Sorbonne Université, F-75013, Paris, France
‡Centre de Ressource Autisme, Université de Nice, Fondation Lenval, F-06200, Nice, France
§Genious Healthcare, F-34000, Montpellier, France
¶Ecole Centrale de Lyon, LIRIS, F-69134, Lyon, France
‖Aix-Marseille Université , LSIS, F-13397, Marseille, France

***Abstract*—Being able to produce facial expressions (FEs) that are adequate given a social context is key to harmonious social development, particularly in the case of children plagued with autism spectrum disorder (ASD). In this paper, we introduce JEMImE, a serious game solution that aims at teaching children how to produce FEs. JEMImE is based on a FE recognition module that is learned on a large video corpus of children performing FEs. This module is validated and incorporated through multiple scenarios of gradual difficulty, ranging from a training phase where children have to perform the FEs on request, with or without an avatar model, to an in-context phase that involves many emotion-eliciting social situations with virtual characters.**

***Keywords*-Facial expression recognition, serious game, autism spectrum disorders, emotion production**

INTRODUCTION

Autism spectrum disorder (ASD) is a neurodevelopmental disorder that affects communication and socialization of individuals with deficits in social emotion reciprocity, in non-verbal communication and in developing and maintaining relationships. All these social skills are important in enabling a person to achieve social competence [1] and are factors of integration in the society at all ages of life. Among them, emotional skills are essential to communicate our emotions to others and to adapt our behavior according to their reaction. Within these emotional skills, facial expressions (FEs) are key components of emotional signal and allow people to express and understand emotions [2]. Their correct recognition and production is essential. Moreover, they have to be adapted to the social context, requiring people to take care of the situation and social rules that apply to it [3]. Teaching social skills to individuals with ASD is a considerable challenge. Recently, many studies have considered the use of information communication technologies (ICTs) in therapy [4]. In fact, some studies showed that ICTs improve interest and motivation of children with ASD [5]. ICTs present information in a sequential way, making them predictable and reassuring [6]. Moreover, they allow working on social skills thanks to virtual environment allowing the therapist to place the child in many different situations close to reality but in a safe place [7]. The review of Grynszpan et al. (2014) [8] reports that results of intervention based on ICTs are promising. In a recent review [9], 31 serious game solutions were identified that aim at teaching social skills to people with ASD. Among them, there are only 4 games that deal with targeted FE production (LifeIsGame, CopyMe, SmileMaze, and the serious game of Park et al. (2012)) [10], [11], [12], [13]. Only the game LifeIsGame includes emotion production exercises in a social context with no visual support while the others games proposed only to work on FEs without linking these to a social environment. None of these games includes a feedback of the facial expression produced by the player; however, feedback on the facial production improves the capability of self-correction for people with and without ASD [14]. There was no assessment conducted in order to evaluate the efficiency of the interventions; only 2 games were proposed to children with ASD, but without evaluating their progress after playing. A summary of these contributions can be found in Table I.

Regarding the gameplay of these 4 games, they seem not to fully exploit the potential of ICTs. Only LifeIsGame proposes a dynamic support with a virtual avatar. The other games use 2D and static supports, which could be presented on paper. Moreover, playful aspects such as rewards are not always present to increase the urge to play. The personalization possibilities of the player's character are limited, which further limits the interest in the game. To wrap it up, only a handful of games exist for the purpose of teaching FE production, and they do not use all the entertainment characteristics of a game, which is critical to motivate the children to play in the first place. Moreover,

978-1-5386-2335-0/18/$31.00

Table I
SUMMARY OF GAMES TEACHING FACIAL EXPRESSION PRODUCTION

game	support	clinical study	main results	feedback on FE quality
CopyMe [12]	Pictures of real persons	Game was used with children aged 8 to 10 years in a childcare centre in Sydney. However, no assessment was reported.	No	No
LifeIsGame [11]	3D avatar	Only a qualitative assessment of the design by 9 participants. No evaluation.	Participants enjoyed playing this game. The children seemed to match images more than they recognised expressions.	No
SmileMaze [13]	Smileys	No	Informal field-testing showed that children with ASD enjoy playing the game.	No
Theory driven serious game framework [10]	Photos and writings	No	No	No

they do not offer enough information (such as feedback on players productions, or taking into account social situations) to help the child producing suitable FEs.

I. JEMIME OVERVIEW

JEMImE is a French acronym that stands for "educative multimodal game for emotional imitation". The goal of this project is to develop a serious-game platform on which children with ASD can receive feedback on the expression that they display, so as to produce FEs adequately, given a social context. The proposal of adding a feedback regarding FE follows a clinical study conduct with a sister game JeSTIMULE based on the same principals but without feedback [15] A lot of emphasis is put into developing a fun environment to create an incentive for the children to be interested in the game, with appealing visuals and thorough personalization possibilities (see Section V). By doing so, we hope improving playability of the serious game [16]. The game is primarily geared towards being used by ASD children aged from 6 to 12 years old. An overall flowchart of this game is shown on Figure 1. First and foremost, we have gathered and labeled a large database of videos depicting typical children's FEs (Section II). For each video, during an offline phase, we extract a number of frames that we use to feed a machine learning algorithm (see Section III), which extract low-level features and map those features to the corresponding FE label. The learned predictive model are validated (Section IV), then can be used as a FE recognition module and integrated into the serious game solution (Section V). While playing, a child can go through multiple modes to try and learn to produce FEs with gradual difficulty, by first imitating and producing FEs on request, then by producing FEs adequately given a specific scenario in a virtual environment.

II. GATHERING THE DATABASE

A. Data collection

A total of 157 volunteer children aged between 6 and 11 years were recorded in Paris (63 children) and Nice (94 children). Among this pool of children, 52% were boys and 48% girls. Moreover, 77% were Caucasian, 8.3% Black-African, 7% Asian and 7% Nord-African. Each child was asked to produce 4 facial expressions: *neutral*, *happiness*, *anger* and *sadness* following two tasks: the *on request* and *imitation* FE production tasks.

More specifically, children were put in front of a computer that was recording the emotional display. An examiner stand behind this screen in order to encourage children to keep their heads in front of the screen. For the *on request* task, the screen was explicitly displaying the FE that the child had to produce ("can you show me *happiness*?"). For the *imitation* task, the child was presented an avatar displaying the desired FE, and was asked to imitate it. Each child was asked to perform each FE 6 times total, 2 times for the *on request* task and 4 times for the *imitation* task, each corresponding to either *visual* or *audiovisual* modalities, and with avatars of both genders. The modality and avatar presentation order were randomized to avoid any learning effect. Children were roughly 1 meter away from the recording sensor, hence the face crops are approximately 300×400 pixels.

B. Annotation and extraction

Thus, each child was recorded 24 times in total, making a total of 3768 videos of 3s average length. As explained in Section I, the JEMImE project is geared towards assessing, through a serious game platform, whether the FEs produced by children with ASD are adequate given a social context. Therefore, we not only have to recognize FEs produced by children, but also to guess to what extent the recognized FE is credible.

For that matter, 3 judges blindly labeled the videos in terms of FE *quality*. FE quality was measured on a 0-

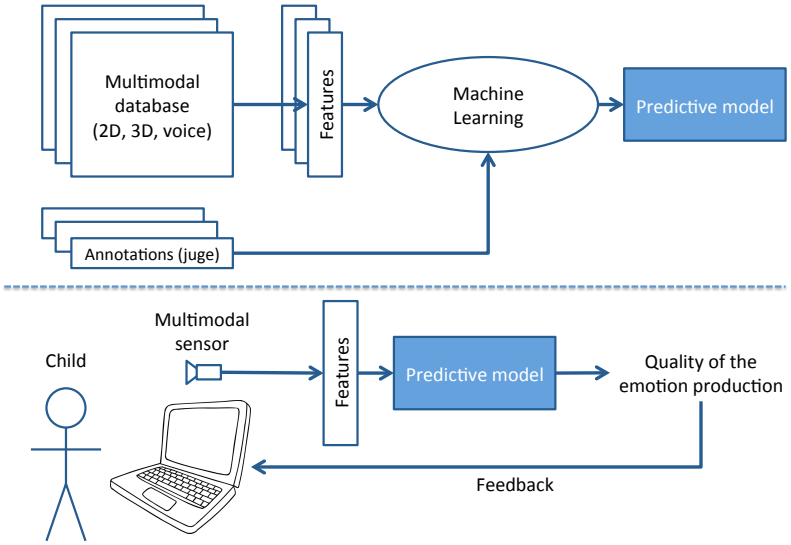

Figure 1. Overall flowchart of JEMImE development. First, during an offline phase (top row), we extract high-dimensional, low-level facial features and train a predictive model using machine learning techniques. This model learns a mapping between the features and the corresponding facial expression labels. Then, this predictive model is applied to the raw video stream of a child playing JEMImE in front of the camera, and provide feedback on the child's expressions in real-time.

10 continuous interval with the following convention: a 0 corresponds to an unrecognized FE, a 5 corresponds to a recognized but not credible FE, and a 10 corresponds to a completely credible and well identified FE. While training regressors to directly model the quality of FEs is an interesting research direction that will be addressed in future works, in this study we extract only a subset of videos for which the FE quality is superior or equal to 7. The reason for this is that we consider that a suitable expression display for one child (and, ultimately, a child with ASD) shall look similar to a high-quality (*i.e.* well recognized, and credible) FE produced by typical children.

For each video, we converted the first frame to grayscale levels, and applied OpenCV Viola & Jones face detector [17]. Then, we applied the intraface feature point tracker [18] to locate a set of 49 feature points. Then, we tracked the feature points on the remaining frames of the video. We selected the last frame of each video for training and testing the FER models, as it usually depicted the peak (apex) of the FE. We discarded some videos for which the feature point tracker could not follow the head motion and extracted a total of 1458 images for children from Paris and 2323 images from Nice, each associated to a FE quality label, a children ID number and a set of aligned feature points. In what follow, we respectively refer to those datasets as JEMImE-Paris and JEMImE-Nice. The concatenation of those two datasets is referred as JEMImE-All (3781 images total).

The data repartition for JEMImE-All is showed on Table II, in terms of FE labels and FE qualities, respectively. As it can be seen on Table II, the database is heavily imbalanced in favor of classes *neutral* as compared to *anger* and *sadness*, as there are roughly 3 times more examples of the former than of the latter. Thus the proposed FER pipeline shall be adapted to be robust to data imbalanced to a certain extent.

Table II
EXPRESSION LABEL REPARTITION (%) FOR JEMIME-ALL

Expression	Repartition(%)
Neutral	36.5
Happiness	28.5
Anger	21.5
Sadness	13.5

III. DISCRIMINATING CHILDREN'S FACIAL EXPRESSIONS

A traditional FER pipeline consists in first extracting a set of candidate features upon which a prediction model can be trained. As it will be discussed in the following subsections, we use random forests (RFs) for the purpose of classifying or regressing the facial expressions. This RF framework allows to generate a large pool of features on-the-fly at the node level (Section III-A). Relevant features among those large collections are then selected by minimizing a purity criterion. To perform FER, we essentially extract heterogeneous features (*i.e.* geometric and appearance) from multiple generic templates, as it provide high-end accuracies on multiple state-of-the-art databases [19].

A. Facial feature extraction from multiple templates

Each of these feature templates $\phi^{(i)}$ have different input parameters that are randomly generated during training. More specifically, for each template $\phi^{(i)}$, the upper and lower bounds are estimated from the training data and candidate thresholds are sampled from uniform distributions within this range. Those features are then associated with a set of candidate thresholds θ to produce a sety of binary split candidates for each split node.

We use two different geometric feature templates which are generated from the set of facial feature points $f(x)$ aligned on image x with SDM [18]. The first geometric feature template $\phi^{(1)}_{a,b}$ is the distance between feature points f_a and f_b, normalized w.r.t. inter-ocular distance $iod(f)$ for scale invariance (Equation 1).

$$\phi^{(1)}_{a,b}(x) = \frac{||f_a - f_b||_2}{iod(f)} \tag{1}$$

Because any information relative to orientation is discarded in $\phi^{(1)}$, we also use the angles between feature points f_a, f_b and f_c as a second geometric feature $\phi^{(2)}_{a,b,c,\lambda}$. In order to ensure continuity for angles around 0, we use the cosine and sine instead of the raw angle value. Thus, $\phi^{(2)}$ outputs either the cosine or sine of angle $\widehat{f_a f_b f_c}$, depending on the value of a boolean parameter λ (Equation (2)):

$$\phi^{(2)}_{a,b,c,\lambda}(x) = \lambda \cos(\widehat{f_a f_b f_c}) + (1-\lambda)\sin(\widehat{f_a f_b f_c}) \tag{2}$$

As for appearance features, we use Histogram of Oriented Gradients (HOG) for their descriptive power and robustness to illumination changes. To allow fast HOG feature extraction, we use pre-computed integral channels as discussed in [20]. First, images are rescaled to a constant size of 250×250 pixels. Then, we compute horizontal and vertical gradients on the image and use these to generate 9 feature maps, the first one containing the gradient magnitude, and the 8 remaining correspond to a 8-bin quantization of the gradient orientation. Then, integral images are computed from these feature maps. From here, we define the appearance feature template $\phi^{(3)}_{\tau,ch,s,\alpha,\beta,\gamma}$ as an integral histogram computed over channel ch within a window of size s normalized by inter-ocular distance. Such histogram is evaluated at a point defined by its barycentric coordinates α, β and γ within a triangle τ defined over feature points $f(x)$. Also, we store the gradient magnitude in the first channel to normalize the histograms. Thus, HOG features can be computed with only 4 access to the integral channels (plus normalization).

B. The random forest framework

Random Forests (RFs) is a popular learning framework introduced in [21]. It has been ubiquitously used in computer vision as they are suited to handle very high-dimensional data (such as images) and can be easily parallelized for fast training and evaluation.

A RF is traditionally built from the combination of T decision trees grown by only examining a subset of the whole feature pool (*random subspace*), and using data bootstraps sampled from the whole training dataset (*bagging*). In our case, we use bootstraps generated at the level of subject IDs, which allows extra tree randomization as well as faster evaluation using out-of-bag error estimate [22]. Formally, a tree can be defined recursively as either a split or a leaf node. Split nodes contains information about a binary split function which consists in a feature and an associated threshold.

During training, split nodes are set using a greedy procedure. For each node n. We denote $l(n)$ and $r(n)$ the left and right subtrees associated with node n. x_n, $x_{l(n)}$ and $x_{r(n)}$ with class labels y_n, $y_{l(n)}$ and $y_{r(n)} \in \mathcal{Y}$ denote the data at node n, $l(n)$ and $r(n)$, respectively. At node n we generate $k^{(i)}$ binary feature candidates for each template $\phi^{(i)}$.

For each candidate ϕ and threshold θ we compute the information gain induced by this candidate, defined as:

$$G(y_n, y_{l(n)}, y_{r(n)}) = H(y_n) - H(y_{l(n)}) - H(y_{r(n)}) \tag{3}$$

Then, we select the "best" binary feature ϕ^n among all features from the different templates, *i.e.* the one that minimizes the impurity criterion H, and use it to set a split at node n. Then, those steps are recursively applied for the left and right subtrees with accordingly routed data until the label distribution at each node is homogeneous, where a leaf node is set. Depending on the purpose of the predictive model (e.g. classification or regression), the nature of the impurity criterion H and the nature of data stored in leave nodes vary. In our case, we use Shannon's entropy as the impurity criterion. Thus, at node n we have:

$$H(y_n) = -m \sum_{y=1}^{\mathcal{Y}} \frac{card(y_n = y)}{m} \log(\frac{card(y_n = y)}{m}) \tag{4}$$

Where m denotes the number of training examples at node n and $card(y_n = y)$ the number of elements with label y. Moreover, as it is also classical in the literature covering RFs, the leave nodes contains the class distributions. During evaluation, an image x is successively routed left or right of each tree according to the outputs of the binary tests, until it reaches a leaf node. Each tree t thus returns the class distribution $p_t(y|x)$. The output prediction $\hat{y}$ is thus given by averaging among the T trees of the forest:

$$\hat{y} = argmax_y \frac{1}{T} \sum_{t=1}^{T} p_t(y|x) \tag{5}$$

Note that given the highly skewed label distribution showed in Table II, balancing the dataset to train the classifiers is essential. For that matter, we apply class-wise downsampling of the bootstraps prior to learning each tree.

As indicated in [23], downsampling leads to similar results compared to other alternatives (e.g. oversampling or class weighting), with a significantly reduced runtime.

IV. Validation of the FE recognition module

A. Experimental setup

We train 4-class RF models with classes *neutral*, *happiness*, *anger* and *sadness* on JEMImE-Paris, JEMImE-Nice and JEMImE-All databases. Trees are trained by generating 20 distances features, 20 angles and 80 randomly samples HOG for each split node, with 25 thresholds per candidate feature. We grow 500 trees with a maximum depth of 16 without early stopping. RFs are evaluated using the Out-Of-Bag (OOB) error estimate [21]. More specifically, bootstraps for individual trees of both static and pairwise classifiers are generated at the subject level. Thus, during evaluation, each tree is applied only on subjects that were not used for its training. The OOB error estimate is an unbiased estimate of the true generalization error [21] which is faster to compute than leave-one-subject-out or k-fold cross-evaluation estimates. Also, it has been shown to be generally more pessimistic than traditional error estimates [22], further empathizing the quality of the proposed contributions. We use the unweighted accuracy (trace of the confusion matrix) as the evaluation metric.

B. FE recognition on JEMImE databases

In Table III we compare accuracies obtained by training classification models on JEMImE-Paris database, and testing on JEMImE-Nice, and vice-versa. Note however that the two databases were collected using a similar protocol and with the same sensors, so this benchmark does not exactly mimic a cross-database scenario. However, it provides some insight on the generalization capacities of predictive models in slightly different contexts - luminosity, as well as cultural discrepancies.

Table III
TEST ON JEMIME (% ACCURACY)

Train-test	JEMImE-Paris	JEMImE-Nice	JEMImE-All
JEMImE-Paris	78.4	74.6	75.6
JEMImE-Nice	79.6	82.2	81.7
JEMImE-All	81.3	82.1	81.9

Models trained on JEMImE-Paris does not generalize very well on JEMImE-NICE database, and therefore does pretty bad on the concatenated dataset JEMImE-All. Interestingly, we still observe a drop in performance when training on JEMImE-Nice and testing on JEMImE-Paris, so this can not be only attributed to the lower number of examples in JEMImE-Paris database.

Table IV presents the per-FE classification scores on JEMImE-All database, along with the average accuracy among the FE classes. As one can see, the classifiers have different biases, as the model trained on JEMImE-Paris output better accuracies for *neutral* and *happiness* classes, with very poor performance for *sadness*. Indeed, *sadness* is the more subtle FE and we believe the low number of examples does not allow to efficiently capture the variability to describe this class. This is confirmed by the accuracies outputted by the models trained on JEMImE-Nice and JEMImE-All that allows more satisfying accuracies for *anger* and *sadness*. Table V shows the confusion matrix obtained for the best overall model, trained on JEMImE-All. Due to the sheer subtlety and variability of the FEs, *anger* is frequently misclassified as *neutral* and *sadness* is confounded with either *anger* or *neutral*.

Table IV
CLASSIFICATION OF FACIAL EXPRESSIONS ON JEMIME-ALL (% ACCURACY)

Train	Neutral	Happiness	Anger	Sadness	Avg.
JEMImE-Paris	87.3	92.6	78.6	43.9	75.6
JEMImE-Nice	84.2	89.1	85.7	67.9	81.7
JEMImE-All	86.4	91.2	83.9	65.9	81.9

Table V
CONFUSION MATRIX FOR TRAINING/TESTING CLASSIFIERS ON JEMIME-ALL DATABASE

	Neutral	Happiness	Anger	Sadness
Neutral	86.4	2.8	6.9	3.8
Happiness	4.4	91.2	2.1	2.3
Anger	5.5	4.5	83.9	3.0
Sadness	12.6	8.9	12.6	65.9

As such, the accuracies are already satisfying for discriminating childrens' FEs. Interesting directions for further improvement of the FE recognition module will include multimodal fusion of 2D and depth information, inclusion of FE temporality and head pose handling [24], as well as the handling of occasional occlusions [25]. Finally, we might want to directly predict the FE quality on a continuous scale, by applying regression models.

V. Designing the game

Based on the FE recognition module described in Section III, we have designed a video game environment to teach children how to produce adequate FE according to a social context. As in JeSTIMULE [15], the game is divided into 2 main phases that will be described in this section: the training (Section V-A) and playing phases (Section V-B), of gradual difficulty.

A. The training phase

During this first phase, children are trained to produce emotions by two ways: either they mimic an avatar displaying a specific emotion (Fig. 2), or they have to produce an emotion on request (Fig. 3). This visual support can be accompanied by background images with an emotional

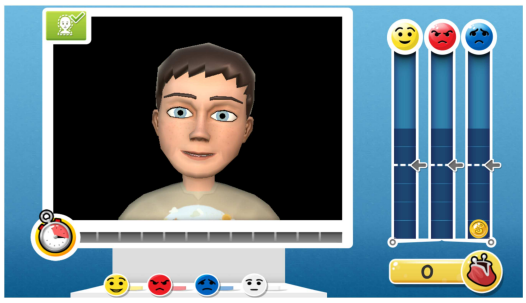

Figure 2. illustration of the imitation task

Figure 3. illustration of the FE-on request task for FE *happiness*

Figure 4. Example of background image used to elicit FE *sadness*

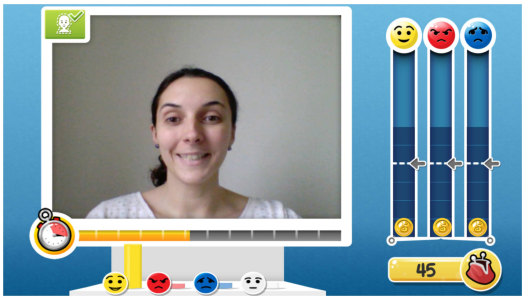

Figure 5. Example of feedback for FE *happiness*. Notice that the yellow gauge below the video widget is very high, indicating that FE *happiness* is well portrayed.

content (e.g. a boy with a flat soccer ball which shall induce the FE *sadness*, see Fig. 4).

Children then have to produce the FE themselves. As visual feedback, children see both their own face and a set of colored gauges displaying how well each FE is recognized by the algorithm in real time. Children can then adapt their production in order to maximize the score of the FE they have to display (Which is related to the output probabilities outputted by the RF predictors described in Section III - see Fig. 5). Each time the FE is correctly displayed by children in the allotted time (*i.e.* if they can hold the score related to the requested FE above a threshold for a certain time frame), they win a virtual coin. The whole level is validated if the child is able to correctly produce a certain proportion of each FE.

B. The playing phase

During this second phase, the child controls an avatar in a virtual world (Fig. 6). In this world, the child is facing social scenarios. He will have to put into practice what he has learned in the first phase and produce the expected FE for each given social scenario. For example, in the situation depicted in Fig. 6, top-right corner, virtual characters playing soccer are asking if the child wants the ball. If he agrees, one of the following two scenarios randomly happens. Either one of the character gives the ball to the child as in Fig. 6, bottom-left corner (in that case the child has to produce the FE *happiness*) or the character tells him that the ball is not for him (Fig. 6, bottom-right). In this latter scenario, either *anger* or *sadness* are accepted as appropriate FEs. The child wins a reward if the adequate FE is correctly produced (the threshold used to decide if a FE is correctly produced is customized according to the progress of the child).

Discussion and Conclusion

In this paper, we introduce JEMImE, a serious game to allow children with ASD to learn how to produce FEs adequately, in response to a specific social context. The game is based on a FE recognition module that consists of a machine learning model trained on a large corpus of children portraying the 4 FEs. We conduct experimental validation to measure the accuracy of FE predictive models, and integrate these into multiple scenarios of gradual difficulty to allow the children to smoothly learn how to produce the FEs. These multiple playing phases are wrapped into colored and beautiful graphics to create the incentive to play the game.

Future developments involve improving the FE recognition module, which include using multimodal information (depth information in addition to the RGB video stream), increasing robustness to head pose and partial occlusions, and using spatio-temporal information to improve the accuracy of the predictive models. We will also try to model the FE quality information on a continuous scale by using regression models. Last but not least, we will conduct studies to evaluate the impact of the game on children with ASD.

Figure 6. Illustrations of the playing phase. Top-left corner: the virtual world in which the child can interact. Top-right, bottom-left and bottom-right corners illustrate an instance of scenario designed to elicit emotions.

ACKNOWLEDGMENT

This work has been supported by the French National Agency (ANR) in the frame of its Research JCJC program (FacIL, project number ANR-17-CE33-0002) and its Research CONTINT program (JEMImE, project number ANR-13-CORD-0004).

REFERENCES

[1] S. H. Spence, "Social skills training with children and young people: Theory, evidence and practice," *Child and adolescent mental health*, vol. 8, no. 2, pp. 84–96, 2003.

[2] C. E. Izard, "Emotional intelligence or adaptive emotions?" *Emotion*, vol. 1, no. 3, pp. 249–57, 2001.

[3] L. F. Barrett, B. Mesquita, and M. Gendron, "Context in emotion perception," *Current Directions in Psychological Science*, vol. 20, no. 5, pp. 286–290, 2011.

[4] S. Boucenna, A. Narzisi, E. Tilmont, F. Muratori, G. Pioggia, D. Cohen, and M. Chetouani, "Interactive technologies for autistic children: A review," *Cognitive Computation*, vol. 6, no. 4, pp. 722–740, 2014.

[5] V. Bernard-Opitz, N. Sriram, and S. Nakhoda-Sapuan, "Enhancing social problem solving in children with autism and normal children through computer-assisted instruction," *Journal of autism and developmental disorders*, vol. 31, no. 4, pp. 377–384, 2001.

[6] V. Knight, B. R. McKissick, and A. Saunders, "A review of technology-based interventions to teach academic skills to students with autism spectrum disorder," *Journal of autism and developmental disorders*, vol. 43, no. 11, pp. 2628–2648, 2013.

[7] N. Josman, H. M. Ben-Chaim, S. Friedrich, and P. L. Weiss, "Effectiveness of virtual reality for teaching street-crossing skills to children and adolescents with autism," *International Journal on Disability and Human Development*, vol. 7, no. 1, pp. 49–56, 2008.

[8] O. Grynszpan, P. L. Weiss, F. Perez-Diaz, and E. Gal, "Innovative technology-based interventions for autism spectrum disorders: a meta-analysis," *Autism*, vol. 18, no. 4, pp. 346–361, 2014.

[9] C. Grossard, O. Grynspan, S. Serret, A.-L. Jouen, K. Bailly, and D. Cohen, "Serious games to teach social interactions and emotions to individuals with autism spectrum disorders (asd)," *Computers & Education*, vol. 113, pp. 195–211, 2017.

[10] J. H. Park, B. Abirached, and Y. Zhang, "A framework for designing assistive technologies for teaching children with asds emotions," in *CHI'12 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2012, pp. 2423–2428.

[11] T. Fernandes, S. Alves, J. Miranda, C. Queirós, and V. Orvalho, "Lifeisgame: A facial character animation system to help recognize facial expressions," in *International Conference on ENTERprise Information Systems*. Springer, 2011, pp. 423–432.

[12] C. T. Tan, N. Harrold, and D. Rosser, "Can you copyme?: an expression mimicking serious game," in *SIGGRAPH Asia 2013 symposium on mobile graphics and interactive applications*. ACM, 2013, p. 73.

[13] J. Cockburn, M. Bartlett, J. Tanaka, J. Movellan, M. Pierce, and R. Schultz, "Smilemaze: A tutoring system in real-time facial expression perception and production in children with autism spectrum disorder," in *ECAG 2008 workshop facial and bodily expressions for control and adaptation of games*. Amsterdam, 2008, p. 3.

[14] R. Brewer, F. Biotti, C. Catmur, C. Press, F. Happé, R. Cook, and G. Bird, "Can neurotypical individuals read autistic facial expressions? atypical production of emotional facial expressions in autism spectrum disorders," *Autism Research*, vol. 9, no. 2, pp. 262–271, 2016.

[15] S. Serret, S. Hun, G. Iakimova, J. Lozada, M. Anastassova, A. Santos, S. Vesperini, and F. Askenazy, "Facing the challenge of teaching emotions to individuals with low- and high-functioning autism using a new serious game: a pilot study," *Molecular Autism*, vol. 5, no. 1, p. 37, Jul 2014.

[16] C. Grossard, S. Hun, S. Serret, O. Grynszpan, P. Foulon, A. Dapogny, K. Bailly, L. Chaby, and D. Cohen, "Rducation de lexpression motionnelle chez lenfant avec trouble du spectre autistiquegrce aux supports numriques: le projet jemime," *Neuropsychiatrie de l'Enfance et de l'Adolescence*, vol. 65, no. 1, pp. 21 – 32, 2017.

[17] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Conference on Computer Vision and Pattern Recognition*, vol. 1, 2001, pp. 511–518.

[18] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *International Conference on Computer Vision and Pattern Recognition*, 2013, pp. 532–539.

[19] A. Dapogny, K. Bailly, and S. Dubuisson, "Pairwise conditional random forests for facial expression recognition," *International Conference on Computer Vision*, pp. 1–9, 2015.

[20] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *British Machine Vision Conference*, 2009.

[21] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[22] T. Bylander, "Estimating generalization error on two-class datasets using out-of-bag estimates," *Machine Learning*, vol. 48, no. 1-3, pp. 287–297, 2002.

[23] C. Chen, A. Liaw, and L. Breiman, "Using random forest to learn imbalanced data," *University of California, Berkeley*, 2004.

[24] A. Dapogny, K. Bailly, and S. Dubuisson, "Dynamic pose-robust facial expression recognition by multi-view pairwise conditional random forests," *IEEE Transactions on Affective Computing*, 2017.

[25] ——, "Confidence-weighted local expression predictions for occlusion handling in expression recognition and action unit detection," *International Journal of Computer Vision*, 2017.