Multimodal Interfaces and Sensory Fusion in VR for Social Interactions

Esubalew Bekele[1,*], Joshua W. Wade[1], Dayi Bian[1], Lian Zhang[1], Zhi Zheng[1], Amy Swanson[3], Medha Sarkar[2], Zachary Warren[3,4], and Nilanjan Sarkar[5,1,*]

[1] Electrical Engineering and Computer Science Department
[2] Computer Science Department, Middle Tennessee State University, Murfreesboro, TN, USA
[3] Pediatrics and Psychiatry Department
[4] Treatment and Research in Autism Spectrum Disorder (TRIAD)
[5] Mechanical Engineering Department, Vanderbilt University, Nashville, TN, USA
{esubalew.bekele,nilanjan.sarkar}@vanderbilt.edu

Abstract. Difficulties in social interaction, verbal and non-verbal communications as well as repetitive and atypical patterns of behavior, are typical characteristics of Autism spectrum disorders (ASD). Advances in computer and robotic technology are enabling assistive technologies for intervention in psychiatric disorders such as autism spectrum disorders (ASD) and schizophrenia (SZ). A number of research studies indicate that many children with ASD prefer technology and this preference can be explored to develop systems that may alleviate several challenges of traditional treatment and intervention. The current work presents development of an adaptive virtual reality-based social interaction platform for children with ASD. It is hypothesized that endowing a technological system that can detect the feeling and mental state of the child and adapt its interaction accordingly is of great importance in assisting and individualizing traditional intervention approaches. The proposed system employs sensors such as eye trackers and physiological signal monitors and models the context relevant psychological state of the child from combination of these sensors. Preliminary affect recognition results indicate that psychological states could be determined from peripheral physiological signals and together with other modalities including gaze and performance of the participant, it is viable to adapt and individualize VR-based intervention paradigms.

Keywords: Social interaction, virtual reality, autism intervention, multimodal system, adaptive interaction, eye tracking, physiological processing, sensor fusion.

1 Introduction

Recent advances in human machine interaction enabled the use of computer technology [1], robot-mediated systems [2,3], and virtual reality (VR) based systems [4,5] for use in social interaction for autism spectrum disorders (ASD) intervention. ASD is characterized by a spectrum of developmental disorders that are associated with

[*] Corresponding authors.

R. Shumaker and S. Lackey (Eds.): VAMR 2014, Part I, LNCS 8525, pp. 14–24, 2014.

social, communicative and language deficits [6], generally poor social skills [7], deficits in facial and vocal affect recognition, social judgment, problem solving and social functioning skills [8] and deficits in the ability to use appropriate language in a social context [9]. Hence a deficit in social interaction is core deficit of ASD. Although, these common social and communicational deficits are observed in most children with ASD, the manifestation of these deficits is quite different from one individual to another [10]. These individual differences call for approaches to individualize the therapy as opposed to one-therapy-fits-all strategies.

Traditional intervention requiring intensive behavioral sessions results in excessive life time costs and inaccessibility of the therapy for the larger population [11]. Recent assistive technologies have shown the potential to at least lessen the burden of human therapists, increase effectiveness of the traditional intervention, and provide objective measures. Literature suggests that children with ASD are highly motivated by computer-based intervention tasks [12]. Predictability, objectivity, lack of judgmental behavior, consistency of clearly defined task and the ability to direct focus of attention due to reduced distractions from unnecessary sensor stimuli are among the benefits of technology-enabled therapy [9].

Virtual reality (VR) [13,5] have been proposed for ASD intervention. VR platforms are shown to have the capacity to improve social skills, cognition and overall social functioning in autism [14]. Explicit modalities such as audio visual for natural multimodal interaction [15] and peripheral physiological signals [16,5] and eye tracking [17] to identify the psychological states of the user and hence adapt the interaction accordingly is crucial in social interactions in general and VR in particular [18,19]. Despite this potential to automatically detect and adapt to the social interaction in VR systems, most existing VR systems as applied to ASD therapy focus on performance and explicit user feedback as primary means of interaction with the participant [20]. Therefore, adaptive interaction is limited in these systems. Adaptive social interaction using implicit cues from sensors such as peripheral physiological signals [14] and eye tracking [21] are of particular importance. For such a system to simulate some semblance of naturalistic social interaction, several components are required including conversational dialog management, body language (gesture), facial emotional expressions and eye contact in addition to the implicit user state understanding components. Conversational dialog is an important part of social interaction. Recently spoken conversational modules have been incorporated into VR systems to achieve more natural interaction instead of menu driven dialog management. Instead of large vocabulary, domain independent natural language understanding, limited vocabulary question-response dialog management, which is focused on the specific domain, has been shown to be effective [22,23]. Such multimodal interactions help in individualization of the therapy and in cases of inaccessibility of trained therapists, it may serve as a self-contained therapeutic system.

This paper describes details of an innovative adaptive VR-based multimodal social interaction platform. The platform integrates peripheral psychophysiological signal monitoring for affective state modeling, eye tracking and gaze metrics for engagement modeling and spoken question-answer-based dialog management for a more naturalistic interaction.

The remainder of the paper is organized as follows. Section 2 details individual components of the system. Section 3 presents preliminary physiology-based affective state modeling results. Finally, Section 4 concludes the discussion by highlighting future direction and extensions of the system.

1.1 VR System Details

The social task presentation VR system is composed of four major components: (1) an adaptive social task presentation VR module, (2) a spoken conversation management module (Q/A-based natural language processing, NLP module), (3) a synchronous physiological signal monitoring and physiological affect recognition module, and (4) a synchronous eye tracking and engagement detection module. Each component runs independently in parallel, while sharing data via light-weight network sockets message passing in a highly distributed architecture. The VR task presentation engine is built on top of the popular game engine Unity (www.unity3d.com) by Unity Technologies. The peripheral psychophysiological monitoring application was built using the software development kit (SDK) of the wireless BioNomadix physiological signals acquisition device by Biopac Inc. (www.biopac.com). The eye tracker application was built using the Tobii X120 remote desktop eye tracker SDK by Tobii Technologies (www.tobii.com).

1.2 The VR Social Task Engine

The VR environment is mainly built on and rendered in Unity game engine. However, various 3D software such as online animation and rigging service, Mixamo (www.mixamo.com), and Autodesk Maya were employed for character customization, rigging and animation. The venue for the social interaction task is a virtual school cafeteria (Fig. 1). The cafeteria was built using a combination of Google Sketchup and Autodesk Maya. A pack of 12 fully rigged virtual characters (10 teenagers and 2 adults) with 20 facial bones for emotional expressions and several body bones for various gestural animations were used as templates to instantiate most of the characters in the environment. Details of the VR development can be found in [24].

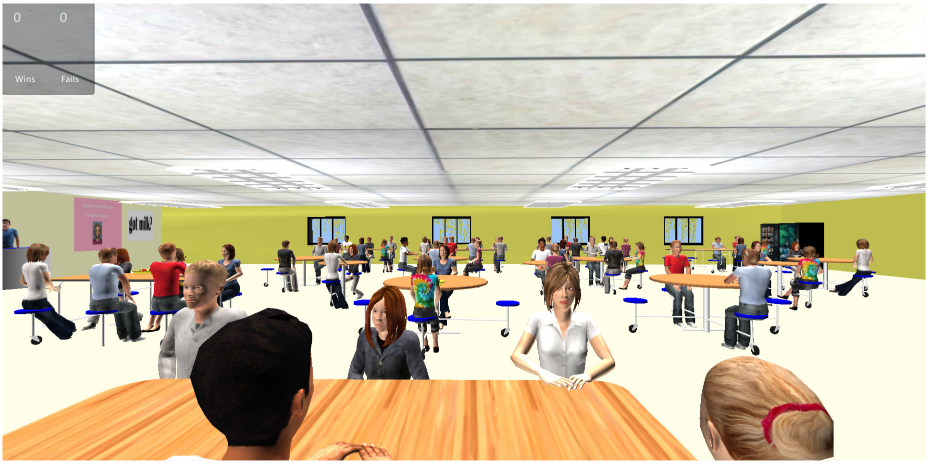

Fig. 1. The VR cafeteria environment for social task training. Dining area (top) and food dispensary area (bottom). The two areas are constructed in separate rooms.

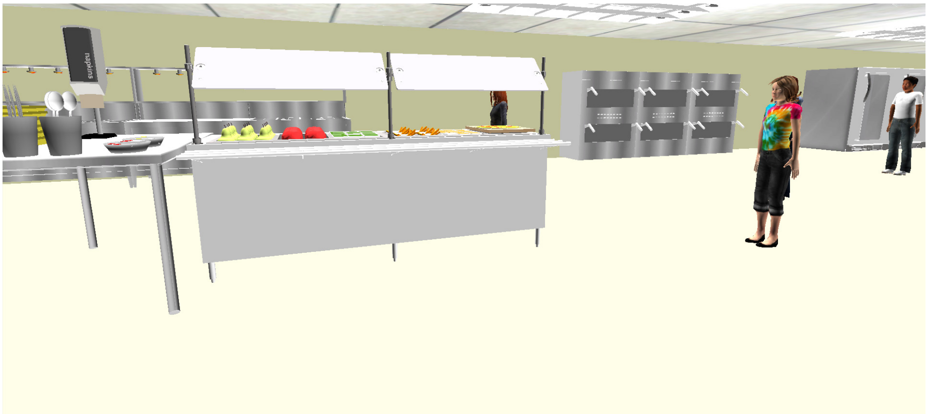

Fig. 1. (*Continued.*)

1.3 Spoken Dialog Management

The verbal conversation component of the VR system creates context for social interaction and emotion recognition in a social setting and is managed by a spoken dialog management module. The dialog manager was developed using the Microsoft speech recognizer from the speech API (SAPI) with domain specific grammar and semantics. The conversation module is based on question-answer dialog and it contained conversational threads for easy (level 1, L1), medium (level 2, L2) and hard (level 3, L3) social tasks with each level having 4 conversational task blocks called missions that the participant is expected to accomplish. Each mission has further components called turns representing back and forth between the participant and the system. L1 missions have one turn, L2 missions have two turns and L3 missions have 3 turns (Fig 2). Each turn was represented by a tree of dialog with nodes representing each option and a particular branch in the tree representing the dialog alternative paths from the initial question to the final correct answer. Failure and success is measured in each conversational turn and there is a hierarchical scoring mechanism that keeps track of performance in conversation turn level as well as mission level. Options in each turn are presented to the participant using a list of items and the participant speaks out their choice through a microphone. Kinect is employed for this purpose as its microphones have superior sound directional localization and background noise cancellation features.

Overall performance, i.e., success/failure (S/F) is used to switch across missions (levels) as shown in Fig. 2 in the "performance only" version of the system. In the adaptive system, physiological affect recognition as well as eye tracking-based engagement detection are combined to adapt the level of difficulty of the interaction in addition to the overall performance of the participant.

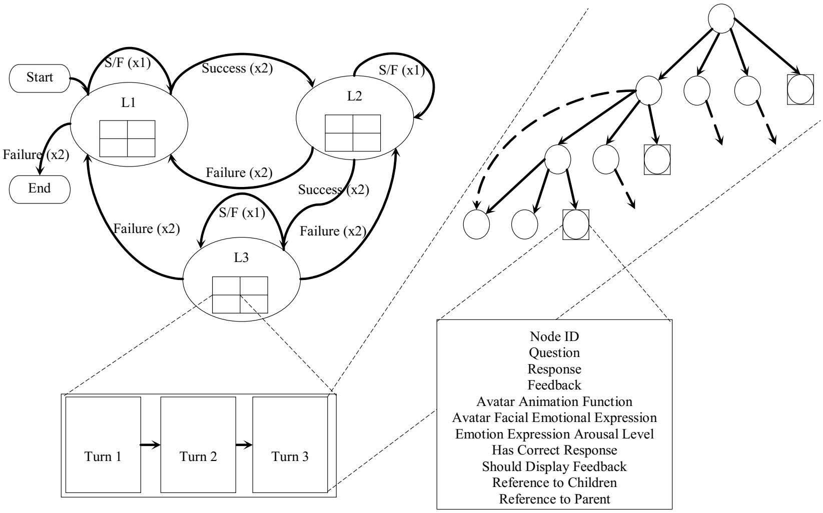

Fig. 2. Finite state diagram showing a level switching logic

1.4 Physiology-Based Affective State Modeling

The physiological monitoring application collects 8 channels of physiological data and was developed using the Biopac software development kit (SDK) and BioNomadix wireless physiological acquisition modules with a sampling rate of 1000 Hz. The physiological signals that were monitored were: electrocardiogram (ECG), pulse plethesymogram (PPG), skin temperature (SKT), galvanic skin response (GSR), 3 electromyograms (EMG), and respiration (RSP). Various features extracted out of these signals are used for supervised training of a machine learning algorithm for later affective state classification in the actual interaction. Training data was collected separately using a study designed to elicit target affective states such as liking, engagement, boredom, and stress. We developed and used computer-based pong and anagram solving cognitive games with trials carefully designed to elicit the states. Details of the cognitive tasks are presented in Table 1.

Table 1. Number of trials and trial durations for the games

Games	Sub-sessions	Number of Trials	Trial Duration
Anagram	sub-session 1	3	90 s
	sub-session 2	3	180 s
	sub-session 3	6	180 s
Pong	sub-session 1	3	120 s
	sub-session 2	3	120 s
	sub-session 3	9	120 s

The data were passed through various successive signal processing stages. First, the data was passed through very large signal uncorrelated outlier rejection block. Then, every channel was subsampled to lower frequency to keep most of the signal content while reducing computational burden. Signals such as EMG were subsampled at higher frequency whereas slow moving signals such as SKT, GSR, and RSP were down sampled at much lower frequency. The data were then filtered to remove high frequency uncorrelated noise, motion artifacts, very low frequency trending and DC bias, power line noise, and inter-channel interference (e.g.: ECG artifact on EMG). Finally, features were extracted from the channels for supervised training. For this preliminary comparative study, we choose four channels (which are prominent in capturing the autonomic system response), i.e., ECG, PPG, GSR, and SKT, and out of them 16 features were extracted.

Generally, people recognize emotions in speeches with an average 60% and from facial expressions with 70-98% [25]. Emotion recognition using peripheral physiological signals and machine learning techniques such as artificial neural networks (ANN) and support vector machines (SVM) under controlled experiments achieved a comparable recognition rates [26-29]. We comparatively studied the performance of SVM and ANN with three separate learning methods each. The most popular learning algorithm to solve the error minimization problem in ANN is the back propagation (BP) algorithm [30]. However due to its slow convergence and other issues such as convergence to local minima, a variety of methods have been proposed to improve time-space and error of performance BP. These methods range from adhoc methods with adaptive learning rate and momentum (GDX) [31] to using numerical approximations including Newton's secant method by Broyden, Fletcher, Goldfarb, and Shanno (BFGS) [32] and non-linear least squares called Levenberg-Marquardt (LM) [30]. To optimize the error margins of SVM, the performance of the quadratic programming (QP) [33], sequential minimal optimization (SMO) method that decomposes the larger QP problem in to a series of smaller QP problems [34], and the least square (LS) solver [33] were explored in this study.

1.5 Eye Gaze Based Engagement Modeling

The main eye tracker application computed eye physiological indices (PI) such as pupil diameter (PD) and blink rate (BR) and behavioral indices (BI) [21] such as fixation duration (FD) from raw gaze data. For each data point, gaze coordinates (X, Y), PD, BR, and FD were computed and logged together with the whole raw data, trial markers and timestamps in addition to being used as features for the rule-based engagement detection mechanism. The fixation duration computation was based on the velocity threshold identification (I-VT) algorithm [35].

A rule-based system for engagement detection is developed to infer engagement using the behavioral as well as physiological indices from the tracking data as features. The rules use adaptive thresholds and these thresholds are standardized using baseline data recorded before interaction.

1.6 Multimodal Decision Fusion

At this stage a decision tree based decision fusion for the multimodal interfaces is developed. The decision tree combines the outputs of the physiological affective state model, the engagement model, and performance of the participant as variables to come up with overall system difficulty level adjustment. This module will be used in the main pilot study of the overall system to illustrate its performance. In the results section, we present only the physiological-based affect modeling study that was briefly described in Section 2.3.

1.7 Experimental Procedures

With the designed distributed emotion modelling system based on computer games (Section 2.3), a total of 10 children with ASD participated in the study. Each participant went through all the Pong and Anagram sub-sessions as shown in Table 1. In each trial we monitored all the physiological signals described in section 2.3 and extracted 16 features out of each trial to obtain a total of 192 data points. These include 147 positive valence (liking and engagement/enjoyment) and 45 negative valence points (frustration/Boredom and anxiety). A trained therapist rated each trial on a Likert scale of 1-9 for each of the 4 perceived psychological states and the result was normalized and classified into positive and negative valence classes. All the data was trained to the three training methods of MLP and the three learning methods of SVM described in section 2.3.

2 Results

As described in Section 2, this system development is an ongoing effort which is tested for usability incrementally. The first phase was developing the virtual environment, the characters and endowing facial emotional expressions to the characters. This stage was evaluated with 10 children with ASD and 10 typically developing children in a separate study [19]. After that, this current study, develops more capabilities such as various animations, the cafeteria environment, the speech-based dialog management, affective modeling with supervised training methods and eye tracking based engagement modeling. The system development of the social task VR environment was presented in [24]. This paper presents the current status by adding results of the preliminary physiological-based affect modeling component.

2.1 Preliminary Physiological Modeling Results

We have conducted a separate study to collect training physiological data for affect modeling as described in Section 2.3.

Model Fitting. We performed model selection to get the best parameters for each learning algorithm. For multilayer perceptron (MLP) ANN, we fixed the number of epochs at 10,000, the error requirement at 0, minimum gradient at 1e-5, and the number of validation checks to 1,000 across all the model selection process. For SVM, we selected radial basis function (rbf) as the kernel and the variance of the rbf as 1.0.

We used minimum validation error as criteria to choose the best model (Fig. 3). However, whenever the testing error is not closer or at a local minimum when the validation is at global minima, we chose the next minimum validation error point. Table 2 shows the best model parameters.

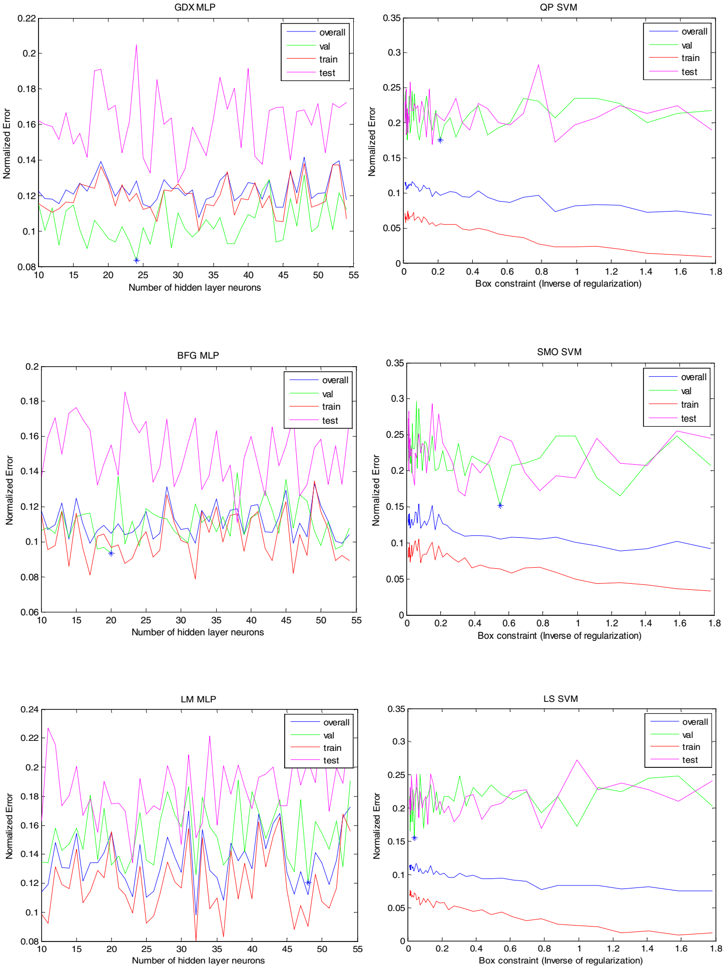

Fig. 3. Normalized Error vs. number of hidden layer neurons (MLP ANN) and vs. inverse of regularization parameter for SVM. Best validation parameters indicated by dark blue point on the green validation line.

Table 2. Selected Best Model Parameters

Classifier	Training Algorithm	Best Model Parameters
	GDX	29.00
	BFG	20.00
MLP ANN	LM	48.00
	QP	0.21
	SMO	0.55
SVM	LS	0.04

Performance Comparisons. Using the best parameters found in the model selection stage, we performed performance of all the six learning algorithms and two classifiers described in section 2.3.

Table 3. Performance Comparisons

Classifier	Training Algorithm	Accuracy	AUC	F_1-score
MLP ANN	GDX	89.09%	85.16%	92.85%
	BFG	94.80%	93.51%	96.58%
	LM	91.70%	89.24%	94.54%
SVM	QP	91.67%	93.02%	94.33%
	SMO	90.62%	88.47%	93.79%
	LS	90.63%	93.11%	93.53%

Table 3 shows the accuracy, area under the curve (AUC) of the receiver operating characteristics curve (ROC), and F1-score. The AUC is basically the average of sensitivity and specificity, whereas F1-score is the harmonic mean of precision with equal weights. The results indicated that both SVM and ANN were able to classify with high accuracy with the BFG algorithm achieving the highest performance for this particular physiological dataset.

3 Conclusion and Future Direction

The main contribution of this work is to present the development of a realistic multimodal VR-based social interaction platform that can be used for ASD intervention. The uniqueness of this platform relies on its ability to gather objective eye gaze and physiology data while a participant is engaged in a closed-loop VR-based adaptive social interaction. This paper presents the preliminary physiological modeling results that seem to indicate the viability of such multimodal social interaction environment as an intervention platform for ASD therapy. Future extensions of this system would add a more advanced multimodal sensory fusion and major pilot study to evaluate the whole system as an intervention platform specifically for its efficacy in ASD intervention.

Acknowledgement. This work was supported in part by National Science Foundation Grant 0967170 and National Institute of Health Grant 1R01MH091102-01A1.

References

1. Goodwin, M.S.: Enhancing and Accelerating the Pace of Autism Research and Treatment. Focus on Autism and Other Developmental Disabilities 23(2), 125–128 (2008)
2. Feil-Seifer, D., Matarić, M.: Toward socially assistive robotics for augmenting interventions for children with autism spectrum disorders. Paper presented at the Experimental Robotics 54 (2009)
3. Bekele, E., Lahiri, U., Davidson, J., Warren, Z., Sarkar, N.: Robot-Mediated Joint Attention Tasks for Children at Risk with ASD: A Step towards Robot-Assisted Intervention. In: International Meeting for Autism Research (IMFAR), San Deigo, CA (2011)

4. Andreasen, N.C.: Scale for the assessment of positive symptoms. University of Iowa, Iowa City (1984)
5. Welch, K., Lahiri, U., Liu, C., Weller, R., Sarkar, N., Warren, Z.: An Affect-Sensitive Social Interaction Paradigm Utilizing Virtual Reality Environments for Autism Intervention. Paper presented at the Human-Computer Interaction, Ambient, Ubiquitous and Intelligent Interaction (2009)
6. Lord, C., Volkmar, F., Lombroso, P.J.: Genetics of childhood disorders: XLII. Autism, part 1: Diagnosis and assessment in autistic spectrum disorders. Journal of the American Academy of Child and Adolescent Psychiatry 41(9), 1134 (2002)
7. Diagnostic and Statistical Manual of Mental Disorders: Quick reference to the diagnostic criteria from DSM-IV-TR. American Psychiatric Association, Amer Psychiatric Pub Incorporated, Washington, DC (2000)
8. Demopoulos, C., Hopkins, J., Davis, A.: A Comparison of Social Cognitive Profiles in children with Autism Spectrum Disorders and Attention-Deficit/Hyperactivity Disorder: A Matter of Quantitative but not Qualitative Difference? Journal of Autism and Developmental Disorders, 1–14 (2012)
9. Gal, E., Bauminger, N., Goren-Bar, D., Pianesi, F., Stock, O., Zancanaro, M., Weiss, P.L.: Enhancing social communication of children with high-functioning autism through a co-located interface. AI & Society 24(1), 75–84 (2009)
10. Ploog, B.O., Scharf, A., Nelson, D., Brooks, P.J.: Use of Computer-Assisted Technologies (CAT) to Enhance Social, Communicative, and Language Development in Children with Autism Spectrum Disorders. Journal of Autism and Developmental Disorders, 1–22 (2012)
11. Ganz, M.L.: The lifetime distribution of the incremental societal costs of autism. Archives of Pediatrics and Adolescent Medicine 161(4), 343–349 (2007)
12. Bernard-Opitz, V., Sriram, N., Nakhoda-Sapuan, S.: Enhancing social problem solving in children with autism and normal children through computer-assisted instruction. Journal of Autism and Developmental Disorders 31(4), 377–384 (2001)
13. Parsons, S., Mitchell, P., Leonard, A.: The use and understanding of virtual environments by adolescents with autistic spectrum disorders. Journal of Autism and Developmental Disorders 34(4), 449–466 (2004)
14. Kandalaft, M.R., Didehbani, N., Krawczyk, D.C., Allen, T.T., Chapman, S.B.: Virtual Reality Social Cognition Training for Young Adults with High-Functioning Autism. Journal of Autism and Developmental Disorders, 1-11 (2012)
15. Lang, P.J., Bradley, M.M., Cuthbert, B.N.: International affective picture system (IAPS): Technical manual and affective ratings. The Center for Research in Psychophysiology, University of Florida, Gainesville, FL (1999)
16. Herbener, E.S., Song, W., Khine, T.T., Sweeney, J.A.: What aspects of emotional functioning are impaired in schizophrenia? Schizophrenia Research 98(1), 239–246 (2008)
17. Lahiri, U., Bekele, E., Dohrmann, E., Warren, Z., Sarkar, N.: Design of a Virtual Reality based Adaptive Response Technology for Children with Autism. IEEE Transactions on Neural Systems and Rehabilitation Engineering: A Publication of the IEEE Engineering in Medicine and Biology Society PP (early access), (99), p. 1 (2012)
18. Zeng, Z., Pantic, M., Roisman, G.I., Huang, T.S.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(1), 39–58 (2009)
19. Bekele, E., Zheng, Z., Swanson, A., Crittendon, J., Warren, Z., Sarkar, N.: Understanding How Adolescents with Autism Respond to Facial Expressions in Virtual Reality Environments. IEEE Transactions on Visualization and Computer Graphics 19(4), 711–720 (2013)

20. Parsons, T.D., Rizzo, A.A., Rogers, S., York, P.: Virtual reality in paediatric rehabilitation: A review. Developmental Neurorehabilitation 12(4), 224–238 (2009)
21. Lahiri, U., Warren, Z., Sarkar, N.: Design of a Gaze-Sensitive Virtual Social Interactive System for Children With Autism. IEEE Transactions on Neural Systems and Rehabilitation Engineering (99), 1–1 (2012)
22. Kenny, P., Parsons, T., Gratch, J., Leuski, A., Rizzo, A.: Virtual patients for clinical therapist skills training. Paper presented at the Intelligent Virtual Agents (2007)
23. Leuski, A., Patel, R., Traum, D., Kennedy, B.: Building effective question answering characters. In: Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue, pp. 18–27. Association for Computational Linguistics (2009)
24. Bekele, E., et al.: A step towards adaptive multimodal virtual social interaction platform for children with autism. In: Stephanidis, C., Antona, M. (eds.) UAHCI/HCII 2013, Part II. LNCS, vol. 8010, pp. 464–473. Springer, Heidelberg (2013)
25. Picard, R.W.: Affective computing. MIT Press, Cambridge (1997)
26. Picard, R.W., Vyzas, E., Healey, J.: Toward machine emotional intelligence: Analysis of affective physiological state. IEEE Transactions on Pattern Analysis and Machine Intelligence 23(10), 1175–1191 (2001)
27. Liu, C., Conn, K., Sarkar, N., Stone, W.: Online affect detection and robot behavior adaptation for intervention of children with autism. IEEE Transactions on Robotics 24(4), 883–896 (2008)
28. Rani, P., Liu, C., Sarkar, N., Vanman, E.: An empirical study of machine learning techniques for affect recognition in human–robot interaction. Pattern Analysis & Applications 9(1), 58–69 (2006)
29. Kim, J., Ande, E.: Emotion recognition based on physiological changes in music listening. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(12), 2067–2083 (2008)
30. Hagan, M.T., Menhaj, M.B.: Training feedforward networks with the Marquardt algorithm. IEEE Transactions on Neural Networks 5(6), 989–993 (1994)
31. Riedmiller, M., Braun, H.: A direct adaptive method for faster backpropagation learning: The RPROP algorithm. In: IEEE, vol. 581, pp. 586–591 (1993)
32. Battiti, R.: First-and second-order methods for learning: between steepest descent and Newton's method. Neural Computation 4(2), 141–166 (1992)
33. Suykens, J.A.K., Vandewalle, J.: Least squares support vector machine classifiers. Neural Processing Letters 9(3), 293–300 (1999)
34. Platt, J.C.: 12 Fast Training of Support Vector Machines using Sequential Minimal Optimization (1998)
35. Salvucci, D.D., Goldberg, J.H.: Identifying fixations and saccades in eye-tracking protocols. Paper presented at the Proceedings of the 2000 Symposium on Eye Tracking Research & Applications (2000)