SmileMaze: A Tutoring System in Real-Time Facial Expression Perception and Production in Children with Autism Spectrum Disorder

Jeff Cockburn[1], Marni Bartlett[2], James Tanaka[1],
Javier Movellan[2] , Matt Pierce[1] and Robert Schultz[3]

[1]Department of Psychology, University of Victoria, Victoria, British Columbia V8W 3P5 Canada

[2] Institute for Neural Computation, University of California, San Diego, La Jolla, CA 92093-0445, USA

[3] Children's Hospital of Philadelphia, Philadelphia, PA 19104 USA

Abstract

Children with Autism Spectrum Disorders (ASD) are impaired in their ability to produce and perceive dynamic facial expressions [1]. The goal of SmileMaze is to improve the expression production skills of children with ASD in a dynamic and engaging game format. The Computer Expression Recognition Toolbox (CERT) is the heart of the SmileMaze game. CERT automatically detects frontal faces in a standard web-cam video stream and codes each frame in real-time with respect to 37 continuous action dimensions [2]. In the following we discuss how the inclusion of real-time expression recognition can not only improve the efficacy of an existing intervention program, Let's Face It!, but it allows us to investigate critical questions that could not be explored otherwise.

1. Introduction

The field of computer vision has made significant progress in the past decade, notably within the domain of automated facial expression recognition. The field has now matured to a point where its technologies are being applied to important issues in behavioral science. Cutting edge computer vision technologies can now be leveraged in the investigation of issues such as the facial expression recognition and production deficits common to children with *autism spectrum disorder* (ASD). Not only can these technologies assist in quantifying these deficits, but they can also be used as part of interventions aimed at reducing deficit severity.

The Machine Perception Lab at University of California, San Diego, has developed the *Computer Expression Recognition Toolbox* (CERT), which is capable of measuring basic facial expressions in real-time. At the University of Victoria in British Columbia, the *Let's Face It!* (LFI!) system was developed as a training program that has been shown to improve the face processing abilities of children with ASD. By combining the expertise behind these two technologies in disparate disciplines we have created a novel face expertise training prototype *SmileMaze*. SmileMaze integrates the use of facial expression production into an intervention program aimed at improving the facial expression recognition and production skills of children with ASD. In the following text we will describe CERT, LFI!, and their union, SmileMaze, a novel expression recognition and production training prototype. We also wish to pay special attention to the scientific opportunities beyond the technologies themselves. In particular, we will discuss how an interdisciplinary approach combining cutting edge science from both computer and behavioral sciences has provided an opportunity to investigate high impact issues that have previously been intractable.

2. CERT: The Computer Expression Recognition Toolbox

Recent advances in computer vision open new avenues for computer assisted intervention programs that target critical skills for social interaction, including the timing, morphology and dynamics of facial expressions. The Machine Perception Laboratory at UCSD has developed the *Computer Expression Recognition Toolbox* (CERT), which analyzes facial expressions in real-time. CERT is based on 15 years experience in automated facial expression recognition [3] and achieves unmatched performance in real-time at video frame rates [4]. The system automatically detects frontal faces in a video stream and codes each frame with respect to 37 continuous

dimensions, including basic expressions of anger, disgust, fear, joy, sadness, surprise, as well as 30 facial action units (AU's) from the Facial Action Coding System.

The technical approach is a texture-based discriminative method. Such approaches have proven highly robust and fast for face detection and tracking [5]. Face detection and detection of internal facial features is first performed on each frame using boosting techniques in a generative framework [6]. Enhancements to Viola and Jones include employing Gentleboost instead of AdaBoost, smart feature search, and a novel cascade training procedure, combined in a generative framework. Automatically located faces are rescaled to 96x96 pixels, with a typical distance of roughly 48 pixels between the centers of the eyes. Faces are then aligned using a fast least squares fit on the detected features, and passed through a bank of Gabor filters with 8 orientations and 9 spatial frequencies (2:32 pixels per cycle at 1/2 octave steps). Output magnitudes are then normalized and passed to the facial action classifiers.

Facial action detectors were developed by training separate support vector machines to detect the presence or absence of each facial action. The training set consisted of over 8000 images from both posed and spontaneous expressions, which were coded for facial actions from the Facial Action Coding System. The datasets used were the Cohn-Kanade DFAT-504 dataset [7]; The Ekman, Hager dataset of directed facial actions [8]; A subset of 50 videos from 20 subjects from the MMI database [9]; and three spontaneous expression datasets collected by Mark Frank (D005, D006, D007) [10]. Performances on a benchmark datasets (Cohn-Kanade) show state of the art performance for both recognition of basic emotions (98% correct detection for 1 vs. all, and 93% correct for 7 alternative forced choice), and for recognizing facial actions from the Facial Action Coding System (mean .93 area under the ROC over 8 facial actions, equivalent to percent correct on a 2-alernative forced choice).

In previous experiments, CERT was used to extract new information from spontanious expressions [11]. These experiments addressed automated discrimination of posed from genuine expressions of pain, and automated detection of driver drowsiness. The analysis revealed information about facial behavior during these conditions that were previously unknown, including the coupling of movements. Automated classifiers were able to differentiate real from fake pain significantly better than naïve human subjects, and to detect driver drowsiness above 98% accuracy. Another experiment showed that facial expression was able to predict perceived difficulty of a video lecture and preferred presentation speed [12]. Statistical pattern recognition on large quantities of video

data can reveal emergent behavioral patterns that previously would have required hundreds of coding hours by human experts, and would be unattainable by the non-expert. Moreover, automated facial expression analysis enables investigation into facial expression dynamics that were previously intractable by human coding because of the time required to code intensity changes.

3. LFI!: The Let's Face It! program

While most people are social experts in their ability to decode facial information, an accumulating body of evidence indicates that individuals with *autism spectrum disorder* (ASD) lack many of the rudimentary skills necessary for successful face communication. ASD is clinically diagnosed as impaired socialization and communicative abilities in the presence of restricted patterns of behavior and interests [13].

3.1. Facial recognition deficits in ASD

Children with ASD frequently fail to respond differentially to faces over non-face objects, are impaired in their ability to recognize facial identity and expression, and are unable to interpret the social meaning of facial cues. For children with ASD, facial identity recognition is specifically impaired in the midst of a normally functioning visual system [14]. Also, children with ASD demonstrate marked impairment in their ability to correctly recognize and label facial expressions [15].

Recognizing faces, identification of expression, and recognition of identity are fundamental face processing abilities. However, the pragmatics of everyday face processing demand that people go beyond the surface information of a face in an effort to understand the underlying message of its sender. For example, in the real world, we read a person's eye gaze to decipher what they might be thinking, or we evaluate a person's expression to deduce what they might be feeling. Not surprisingly, children with ASD also show deficits in eye contact [16], joint attention [17], and using facial cues in a social context [18].

3.2. Let's Face It!: A computer-based intervention for developing face expertise

Let's Face It! (LFI!) is a computer-based curriculum designed to teach basic face processing skills to children with ASD [19]. For ASD populations, there are several advantages to a computer-based approach. Children with ASD may actually benefit more from computer-based instruction than traditional methods [20]. Computer- versus teacher-based approaches in object naming skills have also been compared [21]. It was found that children in the computer-based instruction learned significantly

more new words and showed greater motivation for learning activity than children in the traditional teacher-based approach. Also, the features, such as music, variable-tone intensity, character vocalizations, and dynamic animations, are particularly motivating and reinforcing for persons with ASD and can easily be incorporated into computer-based instruction [22]. Finally, a computer-based curriculum offers a way to provide cost-effective instruction to ASD children in either a home or school setting.

LFI! targets skills involved in the recognition of identity, interpretation of facial expressions and attention to eye gaze through a set of diagnostic assessments as well as a set of training exercises. The assessments provide a diagnostic tool for clinicians, teachers and parents to identify areas of deficit. The exercises provide a training environment through which children learn to improve their face processing skills using a number of engaging games. A single exercise can be used to train a wide range of face processing skills, while each exercises presents training material in a unique way.

Preliminary findings from a randomized clinical trial indicate that children who played LFI! for 20 hours over a twelve-week intervention period showed reliable, (t (59) = 2.61, p=.006; Cohen's d = .69) gains in their ability to recognize the expression and identity of a face using holistic strategies. These results show that "face expertise", like other forms of perceptual expertise, can be enhanced through direct and systematic instruction. Although these preliminary results are promising, a limitation of the LFI! program is that it only uses static training stimuli and does not incorporate the subjects' own dynamic facial productions. In light of evidence suggesting that individuals with autism have impaired or atypical facial expression production abilities [23] the shortcomings of LFI! could be addressed by incorporating dynamic interactions.

4. Training facial expression expertise

While LFI! provides a comfortable and engaging training environment for children with ASD, it only addresses recognition, not production of expressive faces. Relative to neurotypical individuals, individuals with autism are less likely to spontaneously mimic the facial expressions of others [24] and their voluntary posed expressions are more impoverished than those generated by typically developing individuals [25]. Several studies have already shown that a human interventionist can effectively train individuals with an ASD on facial expressions, including some generalized responding [26], providing even greater impetus for our goal of using software for this training. Moreover, training in facial

expression production may improve recognition, as motor production and mirroring may be integral to the development of recognition skills. As such, entangling facial expression perception and production training may prove more fruitful than either training paradigm would alone.

4.1. SmileMaze

We have incorporated the real-time face recognition capabilities of CERT into the LFI! treatment program in a prototype training exercise called SmileMaze. The goal of the exercise is to successfully navigate a maze while collecting as many candies as possible. The player controls a blue pacman-like game piece using the keyboard for navigation (up, down, left, right) and uses facial expressions to move their game piece past obstacles at various points within the maze. As shown in Figure **1**a, the player's path is blocked by a yellow smile gremlin. In order to remove the gremlin and continue along the maze path, the player must produce and maintain a smile for a fixed duration of time.
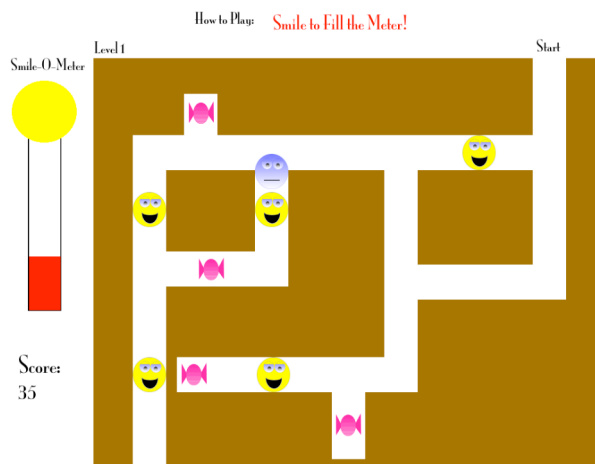
a.
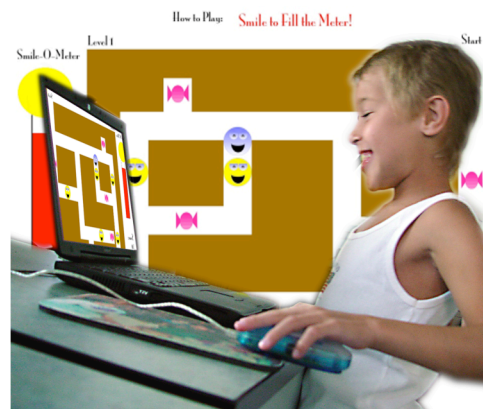
b.

Figure 1. a. Screenshot of the SmileMaze game b. Sample interaction

Video input is captured using a standard laptop video camera and continually analyzed by CERT. If CERT detects a smile, the Smile-O-Meter (Figure 1a, red bar left side) begins to fill. As long as CERT continues to detect a smiling face, the Smile-O-Meter will continue to fill. However, the moment a non-smile is detected the Smile-O-Meter will cease to fill until a smile is once again detected. Once the Smile-O-Meter has been filled, the obstacle is removed and the player may pass. Feedback regarding the facial expression being detected is provided via the game piece, which will change to show the expression being detected.

Informal field-testing indicates that children with ASD, neurotypical children and adults enjoy playing the SmileMaze exercise. Observations suggest that the game initially elicits voluntary productions of smile behaviors. However, users find the game to be naturally entertaining and amusing thereby evoking spontaneous smiling expressions during game play. SmileMaze demonstrates the connection between voluntary and involuntary expression actions in a gaming format where voluntary productions can lead to involuntary productions and changes in affective state.

4.2. Training in a real-world environment

CERT has been shown to perform with high accuracy in a wide variety of real-world conditions; however, the nature of a training environment implies that the system will need to cope with a substantial degree of atypical expressive input. Indeed, pilot testing demonstrated that players enjoy trying to trick the system, posing with odd expressions. While this makes things more difficult for the system, we wish to encourage this sort of exploratory and entertaining behavior. In order to provide a predicable and intuitive user interaction we have incorporated a number of design techniques into SmileMaze to ensure that CERT can cope with atypical interactions. Not only is a stable and robust system desirable from a training standpoint, it is also of paramount importance for a natural and comfortable user interaction.

CERT was designed to work with full-frontal face images. To account for this, we designed SmileMaze such that players naturally orient themselves ensuring that the video camera will capture a full-frontal face image. This was achieved by using a camera mounted at the top center of the computer screen as opposed to a stand-alone camera beside the monitor. This allows players to interact with the system using the video camera without explicitly directing their behavior to the camera. Indeed, pilot testing showed that players seldom explicitly look for, or at the camera when producing facial expressions; rather, their focus is directed at the computer monitor. This provides a tight coupling between user input and system feedback, resulting in a natural and intuitive interaction.

As mentioned previously, pilot testing showed that players enjoyed trying to trick the system with unnatural facial expressions. To assist CERT in accurately labeling facial expressions we always provide a target expression for players to produce. This limits the scope of the possible expressions CERT is required to detect at any given point in time. By providing a target expression, CERT is only required to detect a smiling face, while any other facial expression is deemed to be a failure to produce the appropriate expression. A binary decision (is a smile present or not) reduces the decision space, resulting in very robust expression detection.

5. Future work

Here we do not discuss the intended future work in automated face and expression recognition; rather, we would like to focus on some of the behavioral questions that we can begin to explore by leveraging technologies on the cutting edge of automated real-time face recognition.

5.1. An extended training program

We developed SmileMaze as a proof of concept prototype that could be used to explore the dynamics of training expressive production alongside perception. We noted that participants were more engaged in their tasks and that they found the training exercises more fun if they could actively produce expressions as opposed to passive viewing. Formal measures of the benefits of including production into the LFI! training have not yet been collected as we only have a single expression production-based exercise. However, with the addition of more production-based exercises we intend on quantifying the benefits of a perception/production based training program.

In extending the variety of production-based exercises we are also able to address a number of open scientific questions. One such question relates to stimulus familiarity. Using CERT, we are now able to capture and quantify training stimuli from the participant's environment. Parents, teachers, siblings and friends can all have images captured via web-cam. Captured images can be quantified and labeled by CERT, and integrated into the training stimuli set. While this may provide a more engaging environment for the participant, it may benefit in other ways as well. Participants may be able to bootstrap learning from familiar faces onto learning novel faces, facilitating a generalization of the skills they have learned. Also, using familiar faces from the participant's environment may help in translating the skills learned in training exercises to the real world. Learning facial

expressions using images of mom and dad can be directly applied and experience in the home environment.

A second question we can explore in expanding the diversity of production-based exercises relates to the generality of expressions. We are now able to develop an "Emotion Mirror" application (Figure 2) in which players control the expressions of a computer-generated avatar and/or images and short video clips of real faces. This supports a highly important skill, namely expression invariance. Here, participants can explore the same expression on difference faces. This aids in training a generalized understanding of facial expressions. It also knits expressive production and perception as it is the participant's own face that drives the expressions shown on the avatar and/or image.
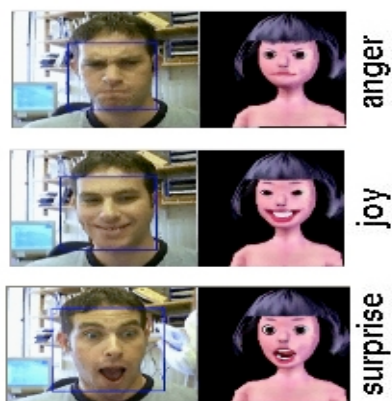

Figure 2. Emotion Mirror: An avatar responds to facial expressions of the subject in real-time.

5.2. An extended expression assessment battery

One of the primary contributions of LFI! was the development of a face perception assessment battery. With the addition of CERT, we are now able to augment the LFI! perception battery with measures of expression production as well. In preliminary research we tested 15 typically developing children ages 4-10 on a set of facial expression production tasks, including imitation from photograph or video, posed expressions with a vignette (e.g., no one gave Maria presents on her birthday. She is sad. Show me sad), and spontaneous expressions (e.g. children are given a clear locked box with a toy inside and the wrong key), as well as smiles recorded during SmileMaze. Preliminary analysis was performed using CERT to characterize the distribution of facial expression productions of normally developing children. This distribution can then be used to measure the expressive production of individuals in a range of tasks and scenarios.

5.3. Translating trained skills into the world

The main goal of an intervention like LFI! is to train and develop skills that translate into improvements in living standards. The goal of LFI! is not just to have participants show improvements on assessment batteries, but to show improvements in their real-world skills. Recognizing when a friend or business customer is unhappy and understanding what that means is crucial to social interaction.

With the inclusion of expression production into the LFI! training and assessment battery we are now able to probe the integration of expressions and emotions at a level not previously possible. Physiological measures such as heart rate and skin conductance have been shown to be sensitive to affective state [27]. It has also been shown that the production of facial muscle movements associated with an emotion produce automatic nervous system responses associated with those emotions [28].

Given that CERT allows us to train and assess the production of facial expressions, we are now able to measure changes in the physiological affects of expression production as in indicator of the integration of expressions and their meanings. While this may not provide conclusive evidence of development beyond production and perception into a cognitive understanding it would provide a strong contribution towards a convergence of evidence not otherwise possible.

6. Conclusions

The CERT system has been shown to encode face activation units and label basic facial expressions in real-time with a high degree of accuracy under a variety of environmental conditions. Also, the LFI! intervention program has been shown to affectively diagnose face processing deficits and improve those skills through a face training curriculum. We have combined these two technologies into a facial expression production training exercise, SmileMaze. While still in a prototype phase, pilot tests strongly suggest that including expressive production into the LFI! program is a great benefit to its users. Further, and of interest to the scientific community, the inclusion of automatic expression recognition allows a number of high-impact and previously intractable issues to be explored.

Acknowledgements

author(s) and do not necessarily reflect the views of the National Science Foundation.

References

[1] R. Adolphs, L. Sears, and J. Piven, "Abnormal Processing of Social Information from Faces in Autism," *J. Cogn. Neurosci.*, vol. 13, pp. 232-240, 2001.

[2] P. Ekman and W. Friesen, *The Facial Action Coding System: A technique for the measurement of facial movement*. Palo Alto: Consulting Psychologists Press, 1976.

[3] D. Gianluca, B. Marian Stewart, C. H. Joseph, E. Paul, and J. S. Terrence, "Classifying Facial Actions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, pp. 974-989, 1999.

[4] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind, and J. Movellan, "Dynamics of facial expression extracted automatically from video," *Image and Vision Computing*, vol. 24, pp. 615-625, 2006.

[5] P. Viola and M. J. Jones, "Robust Real-Time Face Detection," *International Journal of Computer Vision*, vol. 57, pp. 137-154, 2004.

[6] I. Fasel, B. Fortenberry, and J. Movellan, "A generative framework for real time object detection and classification," *Computer Vision and Image Understanding*, vol. 98, pp. 182-210, 2005.

[7] T. Kanade, J. F. Cohn, and T. Yingli, "Comprehensive database for facial expression analysis," presented at Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on, 2000.

[8] M. S. Bartlett, J. Hager, P. Ekman, and T. Sejnowski, "Measuring facial expressions by computer image analysis," *Psychophysiology*, vol. 36, pp. 253-263, 1999.

[9] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," presented at Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on, 2005.

[10] M. S. Bartlett, G. Littlewort, C. Lainscsek, I. Fasel, and J. Movellan, "Machine learning methods for fully automatic recognition of facial expressions and facial actions," presented at Systems, Man and Cybernetics, 2004 IEEE International Conference on, 2004.

[11] M. S. Bartlett, G. Littlewort, E. Vural, K. Lee, M. Cetin, A. Ercil, and J. Movellan, "Data mining spontaneous facial behavior with automatic expression coding," *Lecture Notes in Computer Science*, vol. 5042, pp. 121, 2008.

[12] J. Whitehill, M. S. Bartlett, and J. Movellan, "Automated teacher feedback using facial expression recognition," in *CVPR workshop*, 2008.

[13] A. American Psychiatric, D.-I. American Psychiatric Association. Task Force on, and PsychiatryOnline.com, *Diagnostic and statistical manual of mental disorders DSM-IV-TR*. Washington, DC: American Psychiatric Association, 2000.

[14] A. Klin, S. S. Sparrow, A. de Bildt, D. V. Cicchetti, D. J. Cohen, and F. R. Volkmar, "A Normed Study of Face Recognition in Autism and Related Disorders," *Journal of Autism and Developmental Disorders*, vol. 29, pp. 499-508, 1999.

[15] P. Hobson, J. Ouston, and A. Lee, "What's in a face? The case of autism," *British Journal of Psychology*, vol. 79, pp. 441-453, 1988.

[16] R. M. Joseph and H. Tager-Flusberg, "An Investigation of Attention and Affect in Children with Autism and Down Syndrome," *Journal of Autism and Developmental Disorders*, vol. 27, pp. 385-396, 1997.

[17] A. L. Lewy and G. Dawson, "Social stimulation and joint attention in young autistic children," *Journal of Abnormal Child Psychology*, vol. 20, pp. 555-566, 1992.

[18] D. Fein, D. Lueci, M. Braverman, and L. Waterhouse, "Comprehension of Affect in Context in Children with Pervasive Developmental Disorders," *Journal of Child Psychology and Psychiatry*, vol. 33, pp. 1157-1162, 1992.

[19] J. W. Tanaka, S. Lincoln, and L. Hegg, "A framework for the study and treatment of face processing deficits in autism," in *The development of face processing*, H. Leder and G. Swartzer, Eds. Berlin: Hogrefe, 2003, pp. 101–119.

[20] M. Heimann, K. Nelson, T. Tjus, and C. Gillberg, "Increasing reading and communication skills in children with autism through an interactive multimedia computer program," *Journal of Autism and Developmental Disorders*, vol. 25, pp. 459-480, 1995.

[21] M. Moore and S. Calvert, "Brief Report: Vocabulary Acquisition for Children with Autism: Teacher or Computer Instruction," *Journal of Autism and Developmental Disorders*, vol. 30, pp. 359-362, 2000.

[22] M. Ferrari and S. Harris, "The limits and motivating potential of sensory stimuli as reinforcers for autistic children," *Journal of Applied Behavioral Analysis*, vol. 14, pp. 339-343, 1981.

[23] D. N. McIntosh, A. Reichmann-Decker, P. Winkielman, and J. L. Wilbarger, "When the social mirror breaks: deficits in automatic, but not voluntary, mimicry of emotional facial expressions in autism," *Developmental Science*, vol. 9, pp. 295-302, 2006.

[24] D. N. McIntosh, A. Reichmann-Decker, P. Winkielman, and J. L. Wilbarger, "When the social mirror breaks: deficits in automatic, but not voluntary, mimicry of emotional facial expressions in autism," *Dev Sci*, vol. 9, pp. 295-302, 2006.

[25] S. J. Rogers, S. L. Hepburn, T. Stackhouse, and E. Wehner, "Imitation performance in toddlers with autism and those with other developmental disorders," *J Child Psychol Psychiatry*, vol. 44, pp. 763-81, 2003.

[26] J. DeQuinzio, D. Townsend, P. Sturmey, and C. Poulson, "Generalized imitation of facial models by children with autism," *Journal of Applied Behavior Analysis*, vol. 40, pp. 755-9, 2007.

[27] J. T. Cacioppo, L. G. Tassinary, and G. G. Berntson, *Handbook of psychophysiology*. Cambridge, UK; New York, NY, USA: Cambridge University Press, 2000.

[28] P. Ekman, R. W. Levenson, and W. V. Friesen, "Autonomic nervous system activity distinguishes among emotions," *Science*, vol. 221, pp. 1208-1210, 1983.