Supporting Children’s Social Communication Skills through Interactive Narratives with Virtual Characters

Mary Ellen Foster* Katerina Avramides† Sara Bernardini† Jingying Chen§ Christopher Frauenberger‡ Oliver Lemon* Kaska Porayska-Pomsta†
* School of Mathematical and Computer Sciences, Heriot-Watt University, Edinburgh, UK
† Institute of Education, London Knowledge Lab, London, UK
§ School of Informatics, University of Edinburgh, Edinburgh, UK
‡ School of Informatics, University of Sussex, Brighton, UK
{M.E.Foster, O.Lemon}@hw.ac.uk jchen3@inf.ed.ac.uk C.Frauenberger@sussex.ac.uk
{K.Avramides, S.Bernardini, K.Porayska-Pomsta}@ioe.ac.uk

ABSTRACT

The development of social communication skills in children relies on multimodal aspects of communication such as gaze, facial expression, and gesture. We introduce a multimodal learning environment for social skills which uses computer vision to estimate the children’s gaze direction, processes gestures from a large multitouch screen, estimates in real time the affective state of the users, and generates interactive narratives with embodied virtual characters. We also describe how the structure underlying this system is currently being extended into a general framework for the development of interactive multimodal systems.

Categories and Subject Descriptors: H.5.1 Multimedia Information Systems: Artificial, augmented, and virtual realities; D.2.11 Software Architectures: Patterns

General Terms: Design

Keywords: Technology-enhanced learning

1. INTRODUCTION

The development of social communication skills in children relies on multimodal aspects of communication such as gaze, facial expression, and gesture: being able to interact successfully in social situations requires the participants to be able to understand and produce a wide range of such multimodal social signals.

We introduce the ECHOES technology-enhanced learning environment, in which both typically developing children and children with Autism Spectrum Disorders (aged 5–7 years) can explore their social interaction and collaboration skills [17], with the aim of improving those skills. The environment allows children to interact with a rich multimodal world supporting a wide range of activities. These activities are underpinned by the goals and principles of the Social Communication, Emotional Regulation and Transactional Support (SCERTS) model of assessment and intervention for children with Autism Spectrum Disorder [18]. The SCERTS model is unique in that it is based on the major theories of child development, while incorporating clinical and educational practice. The model targets the cognitive and behavioural prerequisites for successful social interaction and communication.

The ECHOES learning activities take place in a ‘sensory garden’ (Figure 1), which has been designed to be a suitably engaging environment for children. An intelligent virtual character participates with the child in learning activities through which the child can experiment with social communication using a combination of verbal and non-verbal strategies.

ECHOES monitors the child’s actions using a range of sensors including computer vision and multi-touch gestures on a large screen (Figure 2). The multimodal technologies facilitate a number of communicative abilities and their integration: gaze tracking; multimodal fusion; social signal processing; user modelling; planning for embodied, affect-enabled characters; and real-time rendering of a virtual world.
2. ARCHITECTURE AND COMPONENTS

Like most interactive multimodal systems, the ECHOES learning environment is made up of a large number of independent components, which communicate with one another to detect and process user actions and select appropriate responses. Figure 3 illustrates the overall architecture, showing the set of high-level components and the inter-module communication links. The components communicate with each other using the open-source Ice object middleware [12], which provides platform- and language-independent communication among the modules and supports direct module-to-module communication as well as publish-subscribe messaging.

At a high level, the ECHOES system operates as follows. Messages from all of the low-level input-processing components (e.g., head-pose estimation and touch-screen input) are published continuously, along with updates to the graphical state of the world. The multimodal fusion module receives the incoming messages and creates combined messages representing composite multimodal events based on the low-level events. The composite events are sent to the intelligent engine, which selects actions for the virtual character and specifies changes to the state of the world based on the current learning objectives and the child’s behaviour; the child-model component of the intelligent engine also maintains a constantly-updated estimate of the child’s goals and affective state. Finally, the requested character actions and world updates are sent to the rendering engine, which modifies its display and behaviour as necessary. In the remainder of this section, we give detailed descriptions of the four main technical components of the system.

2.1 Visual Processing

Appropriate gaze behaviour is a central part of social communication, and is one of the clearest diagnostic tests for Autism Spectrum Disorders (ASD) [1]. While typical adults attend preferentially to the eye region of the face, people with ASD do not tend to inspect the face region any more than other face parts [15]; it has also recently been shown that autistic children are particularly bad at inferring the locus of attention of a partner’s face [3].

Hence, gaze tracking is important to ensure that ECHOES knows what a child is paying attention to—in particular, whether they are attending to the face, eyes, or mouth regions of a virtual character, or not looking at the face at all. The gaze estimation is determined by two factors: the orientation of the head (head pose) and the orientation of the eyes (eye gaze). Head pose determines the global direction of the gaze, while eye gaze determines the local direction of the gaze. Global gaze and local gaze together determine the final gaze of the person. In our system, a feature-based head pose estimation approach [4] is used. This approach detects a human face using a boosting algorithm and a set of Haar-like features [20], locates the eyes based on Haar-like features, detects two inner eye corners using their intensity probability distribution and edge information, finds two mouth corners using its intensity probability distribution, estimates two nostrils based on their intensity and geometric constraints, and tracks the detected facial points using optical flow based tracking. Examples of the detected feature points are given in Figure 4.

Once the facial features have been detected, the head pose is estimated from tracked points and a facial feature model (i.e., the 3D locations of the six feature points) using the RANSAC [8] and POSIT [6] algorithms. The system is able to detect tracking failures using constraints derived from a facial feature model and to recover from them by searching for one or more features using the feature detection algorithms. Currently, local gaze estimation is under development based on the use of a stereo camera.

2.2 Multimodal Fusion

The multimodal fusion component processes and combines messages from all of the low-level sensors, along with information from the rendering engine about changes to the contents of the display, and sends the recognised multimodal events to the rest of the system for processing. The primary goal of this component is to create higher-level multimodal events by combining messages from the individual channels.

High-level multimodal communication events fall into several classes. One type of event represents related actions on more than one channel—for example, looking at a flower and touching it within a short time window. Other combined events may represent a sequence of actions on one or more channels, such as touching the screen repeatedly, or looking at an object shortly after the character looks at it. A third type of high-level event represents a discrete state change drawn from a continuous data stream: for example, while the vision system continually publishes the users’ gaze coordinates at a high frame rate, the fusion module only informs the rest of the system when the gazed-at object or screen region changes or when a gaze fixation lasts longer than a configurable threshold.
2.3 Intelligent Engine

The intelligent engine is the component of ECHOES responsible for providing the users with meaningful learning experiences in which they are active participants. The users interact with a virtual environment populated by one or more virtual characters. The narrative for each use of the system is not authored in advance, but unfolds on the basis of the specific personalities of the characters involved, the real-time multimodal communication actions performed by the user, and their estimated emotional and cognitive states during the interaction. Since ECHOES is a pedagogical environment, a ‘drama manager’ component monitors the unfolding of the story and may intervene in order to encourage the achievement of the learning objectives and maintain the child’s engagement.

The intelligent engine consists of three main components. First, (semi-)autonomous agents control the decision-making processes of the embodied virtual characters. Each agent is characterised by: (1) a set of goals; (2) a set of strategies to achieve these goals; and (3) an affective system regulating the agent’s emotional tendencies. The real-time interaction between the child and the autonomous agents gives rise to emergent narratives. The architecture of the autonomous agents is based on the FaTiMA system [7], which was designed to control the behaviour of emotionally intelligent virtual characters. FaTiMA incorporates a reactive level, which drives the emotional reactions of the agents, and a deliberative level, which supports their goal-oriented behaviour.

Next, the drama manager is responsible for keeping the embodied agents and the user on track to achieve a particular interaction experience and a set of pedagogical goals. The drama manager intervenes in the interaction between the agents and the child whenever the agent’s execution of an action interferes with the achievement of the learning objectives.

Finally, the child model assesses in real-time the goals and cognitive and affective states experienced by the child during interaction with ECHOES. The child model uses this information to infer the child’s progress in relation to the specific learning goals in order to adapt the interaction to the child’s individual needs. The real-time assessment of the child is based on: (1) static information about the child such as age, gender, and general preferences; (2) information about their previous interactions with the system; and (3) real-time information coming from the multimodal communication stream, as processed by the fusion component. In part, our approach to this topic is similar to that employed by Kapoor et al. [14]: we have gathered a multimodal corpus of children interacting with ECHOES and have annotated the corpus with the appropriate affective states. The resulting corpus is currently being used as a resource for training supervised-learning models designed to estimate a child’s affective state on-line based on the multimodal sensor data, which will shortly be evaluated.

2.4 Rendering Engine

The rendering engine is the component of ECHOES that is responsible for both audio-visual output and low-level interactive behaviour. It produces the graphical scene on the multi-touch screen, it generates sound events (including synthesised and pre-recorded speech), and adjusts the scene according to events or commands received from the fusion module and the intelligent engine. It also constantly publishes any updates to the graphical state to the fusion system for use in creating composite multimodal events.

The module combines several software packages to achieve the variety of tasks. The core is programmed in Python to ensure portability between target platforms and interoperability with the Ice middleware and the other components. The PyOpenGL package is used to render the graphics through OpenGL. Simpler objects and most of the environment are rendered directly, while for more complex animated characters, the Piavca package is used [11]. Piavca is a platform independent programming interface for virtual characters and avatars. It provides high-level control over characters, their movements, facial expressions and other behaviours. The package is originally written in C++, but provides Python bindings which made the integration into ECHOES seamless.

To be able to generate sound output that can be manipulated in real-time we also chose to integrate the SuperCollider application [22] into the system. SuperCollider is a real-time sound synthesis environment that provides a high-level scripting language to create and manipulate a great variety of sounds. The Python binding to SuperCollider by ixi audio facilitates the communication between the core Rendering Engine and SuperCollider.

3. TOWARDS A GENERAL MULTIMODAL SYSTEMS FRAMEWORK

Like most other interactive multimodal systems, the ECHOES learning environment is made up of a number of distributed, heterogeneous software components, developed by different research groups and drawing on diverse research paradigms, as described in the preceding section. These components must all communicate with one another to support the learning activities.

Supporting such inter-module communication is a complex technical task, and one where a general-purpose framework is useful: not only does it simplify the process of developing such a system, by providing a communication layer and general representations for a range of multimodal data, but it also allows components developed independently to be easily integrated and compared with each other and to be reused across multiple systems.

We are currently extending the ECHOES infrastructure described in the preceding section to be a general-purpose framework suitable for use in a range of multimodal systems. As part of this process, we are defining standard interfaces representing a wide range of multimodal input and output channels, as well as representations for higher-level combinations of multimodal data such as those produced by the fusion component (Section 2.2). The framework will allow for traditional rule-based interaction management, and will also permit components that make use of machine-learning techniques [10] to be trained and incorporated.

Using the Ice middleware as the basis for this framework provides several advantages: it is an open-source package with an active development and support community, it is used in several industrial applications, it supports a wide set of operating systems and programming languages, it permits both publish-subscribe messaging and direct module-to-module communication, and it allows for the use of structured, strongly typed messages.

Previous general-purpose frameworks for multimodal interactive systems include the Open Agent Architecture (OAA) and MULTIPLATFORM. OAA [5] is a domain-independent framework for constructing multi-agent systems which has been used as an integration platform for more than 35 applications, combining technologies such as image processing, speech recognition, text extraction, virtual reality, and robot systems. MULTIPLATFORM [13] supports distributed, heterogeneous components and is designed to support components at all levels of a multimodal system, and has been used in systems including SmartKom [21] and COMIC [9].

The current framework has several advantages over these previous proposals. First, Ice allows complex interfaces to be defined

1http://www.ixi-audio.net/content/body_backyard_python.html
2http://www.ai.sri.com/oaa/applications.html
at a language-independent level, and to be sent and received using the native types of the implementation language; this contrasts with both OAA and MULTIPLATFORM, where it is the programmer’s responsibility to convert messages to and from the correct format for transmission. Using such structured messages also improves the robustness of the system, especially during development: modules that use incorrect types will either fail to compile or will provide informative error messages at run time. Ice also supports both publish-subscribe and direct communication, while MULTIPLATFORM supports only the former and OAA only the latter. Finally, unlike the existing frameworks, Ice is still under active development and has a broad support community; it also supports a wider range of platforms and programming language than either of the existing frameworks.

4. SUMMARY AND FUTURE WORK

We have presented the ECHOES technology-enhanced learning environment, which integrates gaze estimation, multi-touch input, and intelligent autonomous characters into an interactive virtual world designed to support children’s exploration of social interaction and communication skills.

Previous systems have addressed similar goals: for example, Tartaro and Cassell [19] developed a life-sized virtual peer able to participate in collaborative narrative together with children with ASD, and found that the use of the virtual peer resulted in higher levels of contingent discourse among the target population; Milne et al. [16] have created a virtual tutor that instructs children with ASD on social skills including detecting and responding appropriately to facial expressions and dealing with bullying, with initial positive evaluation results: while the FearNot! system [2] used virtual drama among artificial characters as part of a system for anti-bullying education for typically developing children, also using FA-tiMA [7] to choose the actions of the characters.

The goals of ECHOES are similar but distinct: we aim to support emergent narratives, rather than providing explicit tutorial interactions as in [16], and are specifically focusing on aspects of shared gaze and joint attention, as these are developmental precursors for many higher-level social communication tasks and have also been clearly demonstrated to be particularly problematic for children with ASD.

The framework underlying the ECHOES environment is being extended to serve as a general-purpose integration platform for multimodal interactive systems. We have explained how the proposed framework provides an advantage to developers of multimodal interactive systems over both building communication infrastructure from scratch, and how it improves on previous frameworks such as OAA and MULTIPLATFORM.

We have recently carried out an evaluation in a local specialist school of an initial learning activity designed to test findings from the literature about joint attention in children with ASD, with promising results. We are also working towards a final large-scale intervention study where the final ECHOES system will be tested in a number of schools. This final evaluation will take place in the context of the SCERTS framework [18], an educational model for children with ASD that embeds assessments and interventions designed to support social communication, emotional regulation, and transactional support into a child’s daily routine. The impact of the learning environment will be assessed through a range of measures including pre- and post-tests of various sorts, along with analysis of the recorded interactions.

5. ACKNOWLEDGEMENTS

The ECHOES project is funded by the UK ESRC/EPSC TLRP-TEL Programme (http://www.tlrp.org). More information about ECHOES can be found at http://www.echoes2.org. The authors thank all of our ECHOES colleagues for helpful discussions and collaboration.

6. REFERENCES