Universal Access in the Information Society
DOI 10.1007/s10209-005-0119-0

Long paper

The emotional hearing aid: an assistive tool for children with Asperger syndrome

R. el Kaliouby (✉) · P. Robinson (✉)

R. Kaliouby · P. Robinson
Computer Laboratory, University of Cambridge, JJ Thomson Avenue, Cambridge, CB3 0FD, UK

✉ R. Kaliouby
E-mail: rana.el-kaliouby@cl.cam.ac.uk

✉ P. Robinson
E-mail: peter.robinson@cl.cam.ac.uk

Abstract People diagnosed along the autistic spectrum often have difficulties interacting with others in natural social environments. The emotional hearing aid is a portable assistive computer-based technology designed to help children with Asperger syndrome read and respond to the facial expressions of people they interact with. The tool implements the two principal elements that constitute one's ability to empathize with others: the ability to identify a person's mental state, a process known as mind-reading or theory of mind, and the ability to react appropriately to it (known as sympathizing). An automated mind-reading system attributes a mental state to a person by observing the behaviour of that person in real-time. Then the reaction advisor suggests to the user of the emotional hearing an appropriate reaction to the recognized mental state. This paper describes progress in the development and validation of the emotional hearing aid on two fronts. First, the implementation of the reaction advisor is described, showing how it takes into account the persistence, intensity and degree of confidence of a mental state inference. Second, the paper presents an experimental evaluation of the automated mind-reading system on six classes of complex mental states. In light of this progress, the paper concludes with a discussion of the challenges that still need to be addressed in developing and validating the emotional hearing aid.

Keywords Autism · Asperger syndrome · Facial expression analysis · Mind-reading · Dynamic Bayesian Networks

Introduction

Autism is a spectrum of neuro-developmental conditions that is characterized by abnormalities in a triad of behavioural domains: social development, communication and repetitive behaviour/obsessive interests [1]. Classic autism lies on one extreme of this spectrum and typically involves associated learning difficulties (below average IQ) and language delay. Asperger syndrome is another condition on the autistic spectrum, where individuals exhibit the same social interaction difficulties and restricted patterns of behaviour and interests seen in classic autism, but have an above average IQ and no general delay in language [6].

The mind-blindness theory of autism [3] argues that there are deficits in the process of empathizing in autism spectrum conditions. The term "empathizing" encompasses two major elements. The first component is mind-reading, or theory-of-mind, which describes one's ability to attribute mental states to others by observing their behaviour (e.g. [3, 6]). The second component is that of sympathizing, or the ability to have an emotional reaction that is appropriate to other people's mental states. Empathizing thus essentially allows us to make sense of the behaviour of other people we are observing, predict what they might do next, and respond appropriately to them. With deficits in the ability to empathize, people diagnosed along the autistic spectrum often have difficulties operating in the highly complex social environment in which we live and are, for the most part, unable to read or understand other people's emotions. Despite these difficulties, children diagnosed with Asperger syndrome seek interaction with other children [2, 18]. These attempts to integrate often fail because of a lack of understanding of nonverbal communication such as facial expressions.

In parallel to the deficit in empathizing, people diagnosed with autistic spectrum conditions show intact or even superior "systemizing" abilities used in the analysis and construction of systems [6]. The fascination and obsession with machines, and in particular with the underlying rules and regularities that govern machines' operation, has been documented in many clinical descriptions of children with autism (e.g. [4, 18]). This interest in machines extends to computer-based technologies, making them particularly suited to therapeutic and assistive tools for autism. The impairment in empathizing abilities, along with the superior systemizing skills of people diagnosed along the autism spectrum, provide the basis for the research presented in this paper.

The emotional hearing aid, which was first introduced in el Kaliouby and Robinson [23], is a portable assistive computer-based technology designed to help children with Asperger syndrome read and react to the facial expressions of people they interact with. The emotional hearing aid implements the two principal elements which constitute one's ability to empathize with others: mind-reading and sympathizing. The first component is an automated mind-reading system that infers the mental states of people by analysing their facial expressions. The system combines top-down predictions of mental state models (Dynamic Bayesian Networks) with bottom-up vision-based processing of the face to recognize both cognitive and affective mental states in real-time video. A technical description of the implementation of the automated mind-reading system can be found in el Kaliouby and Robinson [25, 26]. The second component is the reaction advisor. It suggests appropriate reactions for the user to take based on the inferences made by the mind-reading system. By doing so, the system enables the user to have more appropriate sympathetic reactions towards other people [9].

This paper presents recent progress in the development and validation of the emotional hearing aid on two fronts. First, the paper describes the implementation details of the reaction advisor. Currently, this component is implemented as a rule-based system which takes into account important factors such as the persistence, intensity and confidence of a mental state inference. Second, the paper reports novel results on the validation of the automated mind-reading system on six classes of mental states: agreeing, concentrating, disagreeing, interested, thinking and unsure. These mental states are typically grouped as complex mental states: a broad category of affective and cognitive states of the mind which are not part of the basic emotions (happy, sad, surprised, afraid, angry and disgusted). It is imperative that the emotional hearing aid is able to recognize mental states beyond the basic emotions for several reasons. To start with, complex mental states such as unsure or confusion, concentration and worry occur frequently in spontaneous settings [35]. In addition, the ability to infer a person's cognitive state is crucial since such states are important predictors of a person's behaviour (e.g. [32, 33]). Also, the ability to recognize people's affective states enables one to empathize and engage more effectively with other people (e.g. [30]). To the best of the authors' knowledge, the automated mind-reading system is the first attempt to automatically recognize complex mental states from head and facial expressions.

The paper starts with a survey of three research areas in which this work falls. Section 2.1 presents a survey of existing therapeutic tools for autism and draws attention to the lack of assistive tools for people diagnosed with autism spectrum conditions. Motivated by the need for this type of technology, the emotional hearing aid draws inspiration from the "emotional indexing" method,

an approach for teaching children with autism how to read and respond to emotions. This teaching method is introduced in Sect. 2.2. A summary of automated facial analysis systems is presented in Sect. 2.3. An overview of the tool, including typical use-case scenarios and the overall architecture of the emotional hearing aid, is discussed in Sect. 3. Section 4 discusses the automated mind-reading system in more detail, while Sect. 5 describes progress in implementing the reaction advisor. Section 6 presents an experimental evaluation of the automated mind-reading system. In light of these results, Sect. 7 concludes the paper with a discussion of the challenges that still need to be addressed in developing and validating the emotional hearing aid.

Related work

The design and implementation of the emotional hearing aid draws on state-of-the-art research in three different areas:

1. From an application point of view, the emotional hearing aid is closest to computer-based therapeutic tools for autism.
2. From a conceptual point of view, the design of the emotional hearing aid draws on theories of autism and utilizes approaches currently in practice for teaching people with autism conditions how to understand social and emotional interactions.
3. From a technical point of view, the implementation of the emotional hearing aid lies closest to automated facial analysis systems.

Therapeutic tools for autism

An increasing number of studies show that computer-aided learning and therapy are well accepted by individuals with Autistic Spectrum Disorders [29]. Consequently, computing technology is increasingly being used in therapeutic contexts of autism. The “ Mind-reading DVD” [5] is an interactive guide to learning about emotions, which provides children with a library of over 400 videos and games to test their progress on reading those emotions. “ Kidtalk” [12] is a therapist-moderated online chatting environment, where children work through common social situations, such as going to the movies, by chatting online. The virtual sand box [20] and the virtual environment developed by Strickland [36] enable children to interact in a virtual setting modelled around real-life social scenarios.

In addition to computer-based technologies, several therapeutic tools are based on the use of robots. The AURORA project [13, 39] utilizes an autonomous robot as an interactive toy that can

engage children in a therapeutically relevant environment. The aim is to encourage pro-active social behaviours towards the robot, elicit robot-child eye contact and teach the child the basics of turn-taking and interaction games. Different embodiments of this toy have been investigated including a doll and a four-wheeled vehicle-like toy. The affective social quotient project [10] consists of short digital videos that embody one of several basic emotions and a set of physical dolls linked by infrared to the system. The system knows which dolls correspond to which clips, so that the child can explore emotional situations by picking up dolls with certain emotions, or the system can prompt the child to pick up dolls that go with certain clips. Finally, Kozima and Yano [27] investigate the possibilities of using humanoid robots in therapy.

The above technologies are mostly remedial tools aimed at providing a learning environment to teach children the fundamentals of social behaviour. They do not provide assistance to individuals with autism beyond that gained through teaching. In addition, as they do not operate in a natural human-human interaction environment, they risk failing to generalize [22].

Contrary to most existing work, the portable assistive device described in this paper is designed to assist people diagnosed with autism in real life situations. In a sense, this tool is analogous to a hearing aid, which allows people with hearing problems to communicate with the rest of the world.

Emotional indexing in autism

A number of approaches to teaching emotion understanding to children with autism exist. Such methods may differ in the amount of structure involved: highly structured methods use carefully planned teaching material deployed in a relatively controlled environment. The methods also differ in the setting in which the teaching takes place. This ranges from being hypothetical to being natural and whether or not it is interactive.

One approach, especially suitable for use with children, involves emotional indexing of the child's surrounding environment [17]. Typically, the child's carer indexes the emotional content of situations as they arise and suggests possible actions that can be taken by the child. For example "Oh, Mary got hurt. She is crying. Can you tell Mary, 'I am sorry'?" This approach to teaching emotions has reportedly improved the social competence of some children [17, 22]. In contrary to most other teaching approaches, social indexing works in the child's natural interaction environment reinforcing appropriate social behaviour in a spontaneous setting.

Unfortunately, this method is not always available for the child, as it requires the physical presence of the carer, which in some cases (e.g. school) might be impractical. Also, unlike highly

structured approaches, with this method it is almost impossible to recreate events once they have occurred. For example, using the above example, it is hard to get Mary to fall off her bike again in order to recreate her facial expressions and tears as she falls.

Automated facial analysis systems

The automated inference of complex mental states from facial expressions is a challenging problem because of the complexity inherent in the automated recognition of facial expressions in video, and the stochastic nature of the mapping between facial expressions and mental states. From a technical point of view, the implementation of the emotional hearing aid lies closest to automated facial analysis systems. Automated facial analysis systems are concerned with the problem of identifying facial expressions and head gestures in still frames or in video, and recognizing the meaning underlying that expression.

In 1978, Ekman and Friesan [15] published the facial action coding system (FACS). FACS provides a description of facial signals in terms of facial actions based on muscular activity in the human face. The coding system enables the measurement and scoring of facial activity in an objective, reliable and quantitative way and has since become the leading method in measuring facial behaviour.

With the advancements in real time algorithms for machine vision and machine learning, there has been significant progress in automated facial analysis systems. A detailed survey of existing approaches and implementations of automated facial analysis is outside the scope of this paper, but the reader is referred to Pantic and Rothkrantz [31]. The survey shows that most existing systems either recognize the facial actions which constitute any facial expression or focus on recognizing the six basic emotions.

The implementation of the emotional hearing aid combines state-of-the-art computer vision (feature point tracking, motion, shape and colour image analysis) and machine learning tools [Hidden Markov Models and Dynamic Bayesian Networks (DBNs)] to automate the inference of a wide range of mental states from facial expressions in real-time video.

Overview of the emotional hearing aid

The emotional hearing aid aims to provide real time assistance with reading facial expressions of other people, and advice on reacting to it in a child's natural social environment.

Portable assistive device

The emotional hearing aid is a portable assistive device, which consists of a personal digital assistant (PDA), an earpiece speaker and a wearable camcorder. The current prototype has been implemented on a Pentium 4 3.4 GHz, with 2 GB RAM machine, and uses a full-sized commercial digital camcorder. As the size (and price) of cameras are dropping fast, it will be possible to assemble a wearable version of the tool. For instance, Deja View, Inc. has recently introduced to the consumer market an inch-long digital camcorder which is small enough to clip onto a pair of eyeglasses and record video onto flash memory.

Figure 1 illustrates how the emotional hearing aid provides assistance in a typical interaction scenario between a child with Asperger syndrome (character A in the figure) and another person (shown as B). Video sequences of B are sent to the PDA. The PDA is responsible for analysing the incoming video, and any available context cues for mental state information. It also indexes this event for further retrieval, and uses it, along with a repertoire of situations, to suggest a course of action. This advice is sent back to the wearer in real time, but continuous feedback is avoided to minimize the number of distractions. Also depending on the level of engagement, the output can be visual or audio, and varies in the degree of detail presented.

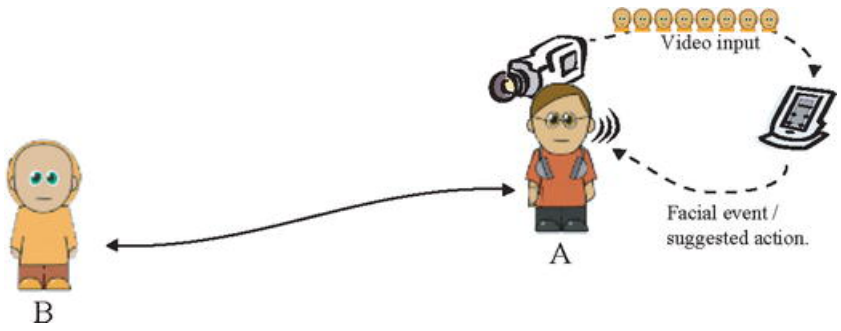

Fig. 1 Child A (diagnosed with Asperger syndrome) is using the emotional hearing aid in an interaction with person B. Video sequences of B and situational context cues are sent to the PDA for analysis and suggested reactions. Depending on the mode of interaction, the output can be visual or audio, and can vary in the degree of detail presented

---

The emotional hearing aid is also designed to work in tandem wi th the child's carer as shown in Fig. 2. The carer can define the emotional and mental states that the tool can infer, can review and update the archived events and can engage in discussions about past events with the child.

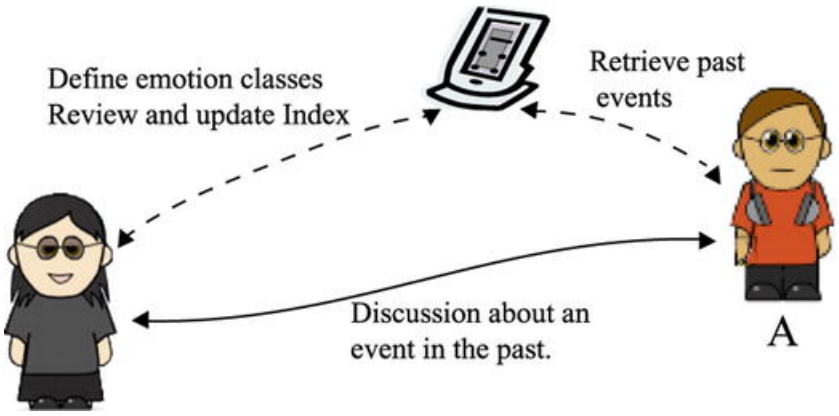

Fig. 2 Another interaction scenario where A (the child) and C (the carer) can query the index for past events. In addition, C is able to define the emotion classes and update the index through the interface

Architecture

The main modules of the emotional hearing aid (shown in Fig. 3) are modelled around the emotional indexing approach. The idea is that the tool is responsible for indexing the emotions of the people the child interacts with, by analyzing the facial expressions of surrounding people, inferring their mental states and communicating that back to the child.

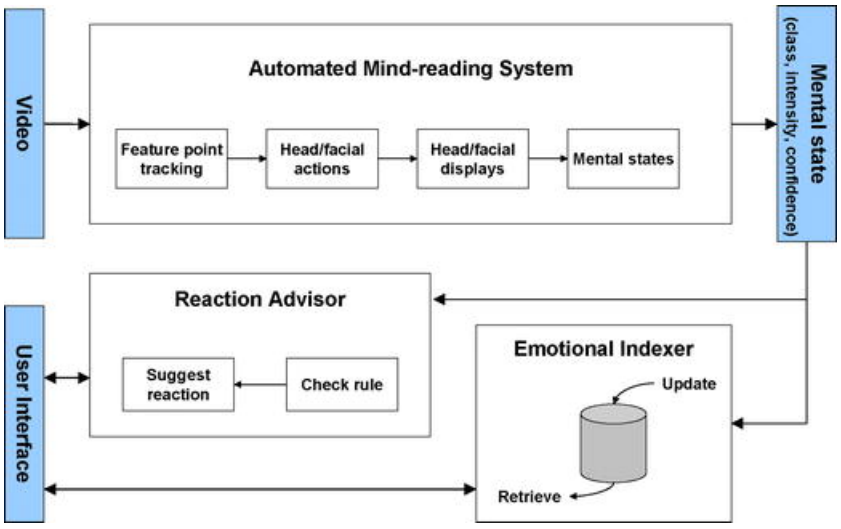

Fig. 3 The main modules defining the emotional hearing aid. For every incoming video frame, the automated mind-reading system identifies head and facial actions, which are combined temporally to form displays. The displays constitute the observation vector for the inference of mental states. Representative frames of the mental state, its label, intensity, valence (whether is a positive or negative state) and accompanying context cues are input to the emotional indexer and reaction advisor. The former appends that event to an index, while the latter utilizes the information to suggest an appropriate reaction. The interface manages the communication between the user and the other modules

The automated mind-reading system identifies a facial event in real time, extracts the dynamic head and facial actions, and infers the underlying mental state conveyed by the video segment. Representative frames of that event along with the inferred mental state label are sent to the emotional indexer to be archived.

The emotional indexer module is responsible for keeping an archive of past events: every event is stored as a tuple in the index containing the representative frames of the mental state, a label, any additional parameters and context cues available. The indexed events are made available (through the interface layer) to the child and carer for discussion, learning and reviewing purposes. This extends the emotional indexing approach to allow events to be replayed. The archive is also made accessible to the automated mind-reading system and the reaction advisor modules to improve inference and suggestions.

The reaction advisor appraises the current video input, analysing it within any context cues that are available, to suggest appropriate courses of actions to take. Timing issues such as latency and frequency of reactions are key factors in the design of this module. A rule-based version of this module has been implemented.

The suggested reactions are communicated to the user via an interface layer. The interface layer manages communication between the child (or carer) and the other modules of the emotional hearing aid. To start with, the interface informs the wearer whenever a facial event occurs, and returns the possible courses of actions suggested by the advisor module. The interface also decides on the modality and format of the output depending on the active profile and level of engagement of the child with the current social scenario. A simple interface layer has been implemented.

A summary-mode is adopted when the user is actively taking part in an interaction, and only needs assistance with the suggested course of action. In this case, the output is unobtrusive to avoid interrupting the interaction, and can be visual (text or graphical) or audio (ambient sounds). In the detailed output mode, all available information pertaining to the event is visually presented to the user. In query mode, the interface allows the index to be queried by both the child and carer using a number of different parameters such as emotion labels, intensity, valence and date of event. In the update mode, the carer is able to retrieve event entries and update them. Finally, the carer can also define the emotion classes that the system supports, activating and de-activating classes as needed.

The following sections discuss the automated mind-reading system and reaction advisor in more detail.

The automated mind-reading system

When mind-reading, people utilize mentally represented generalisations that map observations of particular facial configurations to mental state labels [19, 34]. The automated mind-reading system first builds or "learns" mappings between mental state classes and patterns of facial behaviour as observed in video sequences during a training stage. These mappings are then used during classification to infer the probability of the facial behaviour in an incoming video sequence being "caused" by each of the states. The system combines top-down predictions of mental state models with bottom-up vision-based processing of the face to recognize complex mental states in real time video. A detailed description of the automated mind-reading system can be found in el Kaliouby and Robinson [25, 26].

A procedural description of how inference is carried out in the automated mind-reading system in shown in Fig. 4. When a previously unseen video is presented to the system, processing proceeds bottom-up, but utilizes the mental state models determined off-line from training data. A video is abstracted spatially and temporally into three levels. The output of the system is a number of inference instances, where each instance depicts the probability of each of the mental states. To support real time communication for which the emotional hearing aid is intended, the mind-reading system is designed to meet three important criteria:

1. Full automation/no manual pre-processing
2. Real time execution
3. Categorisation of mental states early enough after their onset to ensure that the resulting knowledge is relevant and actionable.

In terms of full automation, the system builds on FaceTracker [16], a fully automated feature point tracker which does not require any calibration or pre-processing of the images. Also, the system is implemented as a sliding window of size 30 frames, which progresses 5 frames at a time. Accordingly, the system can run for prolonged durations, and there is no need to pre-segment the video into separate expressions or mental states.

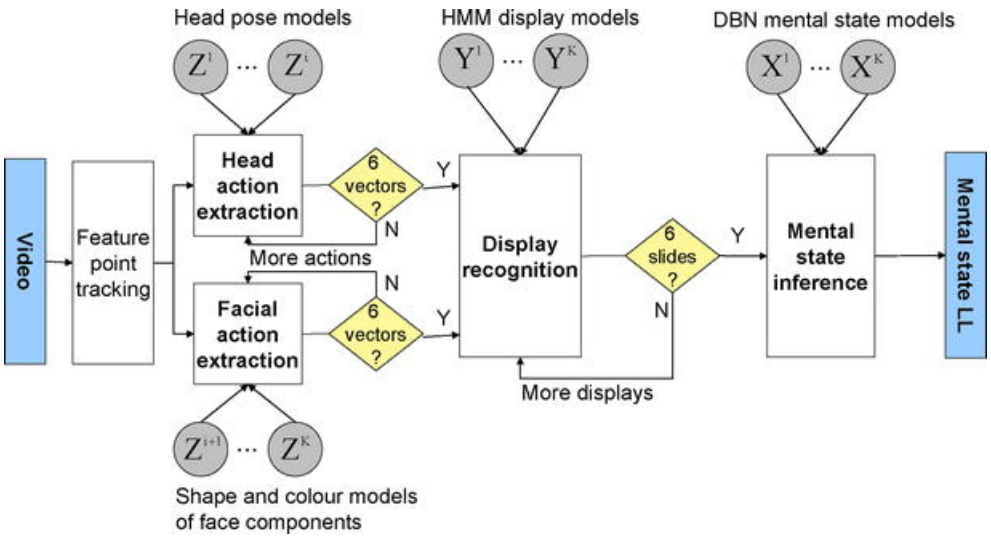

Fig. 4 Procedural description of how inference is carried out in the automated mind-reading system. Vision-based processing of video input of the face is combined with top–down models of actions, displays and mental states, to infer the likelihood of each of the mental states generating the observed facial behaviour. The *grey nodes* represent models that are learnt off line, *blue boxes* are the input and output of the system

---

Operating in real time is a crucial requirement for an automated mind-reading system that is expected to respond appropriately to a person's mental state. It is pointless for the emotional hearing aid to infer that a person is confused and suggest a reaction to the user of the emotional hearing aid to that confused person when the person is no longer experiencing this mental state. In computer-based systems, real time performance is described in terms of the latency of the system, i.e., the time elapsed between a user-driven event and the system responding to it [28, 38]. In the context of mind-reading, latency has been defined as the time elapsed between the onset of a mental state and the system recognizing it. Research has shown that in terms of their utility in decision-making, inferences based on facial expressions are of greatest adaptive value when made at the beginning of an expression [14].

The sliding window implementation has been designed to minimize the latency of the system and ensure that facial expressions are processed as soon as they are made available to the system. With a sliding window size of 30 frames, which progresses 5 frames at a time, mental state inferences are output every 166 ms at 30 fps. At a latency of 166 ms, the system's output is comparable to the time it takes for humans to process emotions in facial expression stimuli. For instance, Batty and Taylor [7] report a latency of 140–200 for the emotional processing of facial expression stimuli of the six basic emotions.

Other key aspects include being unobtrusive and dealing with substantial rigid head motion.

At the first level, head and facial actions based on FACS [15] are extracted from the raw video input. As shown in Fig. 5, video input is abstracted spatially into head rotation along each of the three rotation axes (pitch, yaw and roll) and facial components (lips, mouth and eyebrows). For instance the head pitch is described by a head-up or head-down actions.

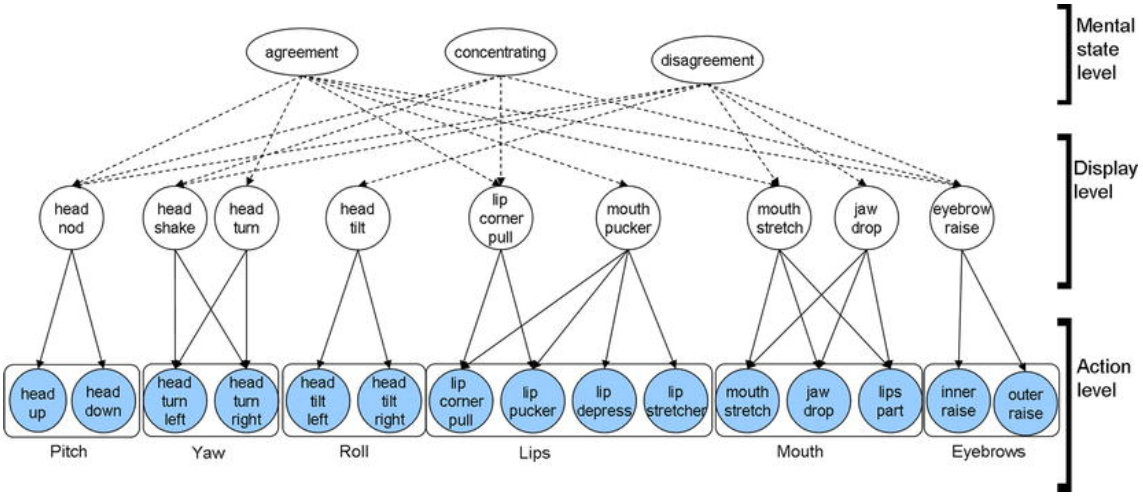

Fig. 5 Video input is abstracted spatially into head rotation along each of the three rotation axes (pitch, yaw and roll) and facial components (lips, mouth and eyebrows). Each spatial abstraction is described by a number of actions (for instance the head pitch is described by a head-up or head-down action). The actions are in turn abstracted into displays (e.g. head nod). The inference of mental states involves several spatial abstractions. The relationship between displays and mental states (shown only for three mental states) is learnt adaptively from data through a feature selection mechanism. Displays present in a mental state's feature set are connected via a dashed arrow

---

The recognition of head and facial actions is based on a feature-point tracking approach. The 22 feature points are typically located on the eyes and eyebrows for the upper face, the lips and nose for the lower face. FaceTracker [16], which is part of Nevenvision's facial feature tracking SDK, is used for feature point tracking. The tracker deals with a wide range of facial characteristics such as varying skin colour. It is robust to a substantial rigid head motion. By tracking these feature points across an image sequence and analysing their displacements over multiple frames, one can establish a characteristic motion pattern for various head and facial action units. For example, the motion patterns of expression-invariant feature points such as the nose tip and inner and outer eye corners over successive frames is used to extract head rotation parameters along each of the three rotation axes. The polar distances between each of the two mouth corners and a stable anchor point are used to identify a lip corner pull, lip stretch and lip pucker. Colour analysis of the mouth determines the presence of aperture and teeth inside the mouth. The ratio of aperture to teeth pixels

determines whether the lips are parted, the jaws dropped, or mouth stretched. Figure 6 shows the results of aperture and teeth detection in selected frames of a video showing the mental state comprehending from the Mind-reading DVD [5]. This sequence of mouth actions is classified as a mouth open since there is a higher ratio of aperture to teeth pixels as shown in the masks.

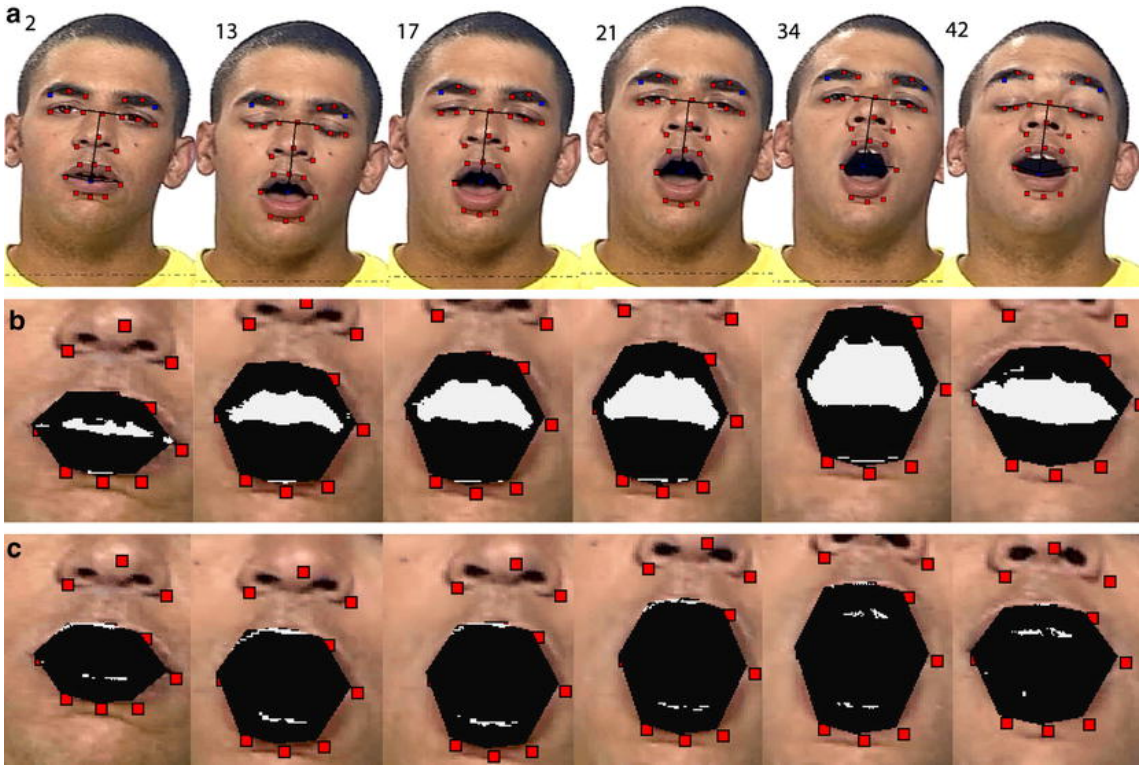

Fig. 6 Tracking the mouth aperture and teeth in a mouth-open display in a video showing the mental state comprehending from the Mind-reading DVD. Note that there is no neutral expression in the transition between the different action units. **a** Selected frames (between 1 and 50) of the localised mouth polygon. These frames are classified as a mouth open. **b** Aperture mask (aperture pixels are highlighted in *white*). **c** Teeth mask (teeth pixels are highlighted in *white*)

Displays make up the second level of the automated mind-reading system, and are defined as head or facial events which have meaning potential in communicative contexts [8]. The input to this recognition problem is a vector of consecutive head and facial actions (extracted from a video which could be of arbitrary length). Each vector spans approximately 1 s of video (at 30 fps), or six consecutive head/facial actions. The sequence of actions is presented to a corresponding hidden Markov model (HMM) classifier. Note that as a sliding window implementation is adopted, the HMM classifiers are invoked every action or 166 ms, i.e. every five frames at 30 fps. Thus, there is no delay between a user performing some display and the system recognizing it.

At the topmost level of the system, the goal is to estimate the most likely mental state model which has given rise to the observed head and facial displays. The mental state models are DBNs that represent high-level mental states given observed displays. Each mental state class is modelled as a separate DBN where the hidden mental state of each DBN is a binary variable (can be either true or false). By having a DBN classifier per class, it is possible that the output of more than one classifier is true. Hence, mental states that are not mutually exclusive or may co-occur, such as thinking and unsure, can be represented by the system. For more details about estimating the parameters of the mental state models from training video sequences and the real time inference of mental states, the reader is referred to [25].

The reaction advisor

The reaction advisor implements the second element of empathizing: having an emotional reaction that is appropriate to the other person's mental state. We discuss two key issues that are fundamental to the design of this module. The first issue is associated with the timing of a suggestion: when and how often should a reaction be suggested, and how soon after a change of emotional state is detected should an action be suggested. The second issue discusses the possible types of reactions and forms of feedback to consider.

Timing considerations in reactions

The reaction advisor is currently implemented as a rule-based system that acknowledges the currently inferred mental state in the form of feedback and suggests simple reactions to it. The reactions are invoked according to the following criteria:

- Persistence: the number of inference instances of a particular mental state has to meet a persistence-threshold to warrant a reaction. This differs from one mental state to another. For example, it is often the case that the state of concentrating lasts longer than a disagreeing reaction.
- Intensity: the intensity of a mental state invokes different reactions. For instance glad, a mild expression of happiness would result in a different reaction from an enthusiastic mental state, even though both belong to the happy group of emotions.
- Confidence: the confidence of a mental state inference has to meet a particular level to invoke a reaction (this is used in combination with the persistence threshold)

– Time elapsed since last inference: there is a minimum threshold which is imposed between recommendations, otherwise it would cognitively overload (and indeed frustrate) the child if an action is suggested with every mental state inference.

Functions of reactions

Deciding on the function of a suggested reaction is also important. Reactions to facial expression can be for the purposes of feedback, empathy [37] or communication with others (e.g. signal turn-taking). The intensity of a reaction that is suggested by the tool also needs to be determined. The intensity of a reaction to a person depends on a combination of factors including the other person's mental state, whether it's a positive or negative state, the intensity of that state, the current situation and the degree of approachability of that other person.

Figure 7 shows a series of screenshots from the reaction advisor. Each screen communicates to the user some temporal information such as the current frame number, the last time a suggestion was made, and the time elapsed since then. The interface also displays the current mental state inference and a recommended action, both textually and graphically.

Fig. 7 A sequence of screenshots from the reaction advisor. Each screen communicates to the user the current frame number, the last time a suggestion was made and the time elapsed since then, the current inference and a recommended action to take (both textual and graphical)

Like mental state inference, the appraisal and reaction to mental states are stochastic by definition. To take into account this stochastic nature, the automated mind-reading system is implemented within a probabilistic framework, namely the dynamic Bayesian network. Our current implementation of the reaction advisor is, however, rule-based and does not take into account the probabilistic nature of this process. Partially observed Markov decision processes are particularly suited to representing different courses of actions within a probabilistic framework (e.g. [21]) and their use in the emotional hearing aid warrants further investigation.

Evaluation of the automated mind-reading system

The automated mind-reading system has been evaluated in terms of its accuracy of classification on a corpus of videos. The accuracy of the automated mind-reading system was considered for the six following complex mental states classes: agreeing, concentrating, disagreeing, interested, thinking and unsure. The classes and the videos were from the "Mind-reading DVD" [5]. Note however that the system can be trained to support any other mental state class provided that training videos of these mental states are available The results of the experiment have implications on the reliability of the automated mind-reading system and predict the system's performance in a spontaneous environment outside of laboratory settings.

The Mind-reading DVD

The emotions library of the Mind-reading DVD has a total of 2,472 videos representing 412 mental states. The mental states are classified taxonomically into 24 mutually exclusive emotion groups such as sure, thinking and unsure, such that each of the concepts is assigned to one and only one group. The 24 groups in this taxonomy were chosen such that the semantic distinctiveness of the different emotional concepts within each group is preserved. In other words, each group encompasses finer shades of the mental state. For example, baffled, confused and puzzled are different shades of the unsure group. The mental states brooding, fantasizing and calculating belong to the meta-group thinking. Out of the 24 groups, the performed study considered the ones that are not in the basic emotion set, and which, as a result have not been addressed by the computer science research community. In particular, the study focuses on the automated recognition of the following six groups—and the emotion concepts they encompass: agreeing, concentrating, disagreeing, interested, thinking and unsure . A total of 164 videos were picked from these groups to test the system.

Video durations vary between 5–8 s, recorded at a frame rate of 30/s. There are no restrictions on the head or body movement of actors in the video. The process of labelling involved a panel of 10 judges who were asked ‘could this be the emotion name?‘ When eight out of ten agree, a statistically significant majority, the video is included in the library of emotion videos. To the authors’ knowledge, the Mind-reading DVD is the only available, labelled resource with such a rich collection of mental states and emotions, even if the videos are of staged, rather than naturally-evoked, mental states.

Note that while most existing automated facial analysis systems operate on short clips (1–2 s), the developed mind reading system places no constraints on the duration of the input, because it keeps a rotating buffer of the input. Hence, there is no need to pre-segment videos into separate facial expressions.

Results

The overall accuracy of the system was evaluated by testing the inference results of 164 videos representing the six mental state classes. The videos span 25,645 frames, or approximately 855 s. Using a leave-one-out methodology, 164 runs were carried out, where for each run the system was trained on all but one video, and then tested with that video. The classification rule that is used to deem whether a classification result is correct is defined as follows: compare the overall probability of each of the mental states over the course of a video. If the video’s label matches that of the most likely mental state or the overall probability of the mental state exceeds 0.6, then it is a correct classification.

Figure 8 shows the output at the different levels of the mind-reading system for a 5.5-s long video portraying the mental state discouraging, which belongs to the mental state group disagreeing. Throughout the video, a number of asynchronous displays which vary in duration are recognized: a head shake, a head turn, a lip corner pull and an open mouth and teeth. The displays affect the mental state outputs as shown in the figure. Since the overall probability of disagreeing is 0.75, this is an example of a correct classification.

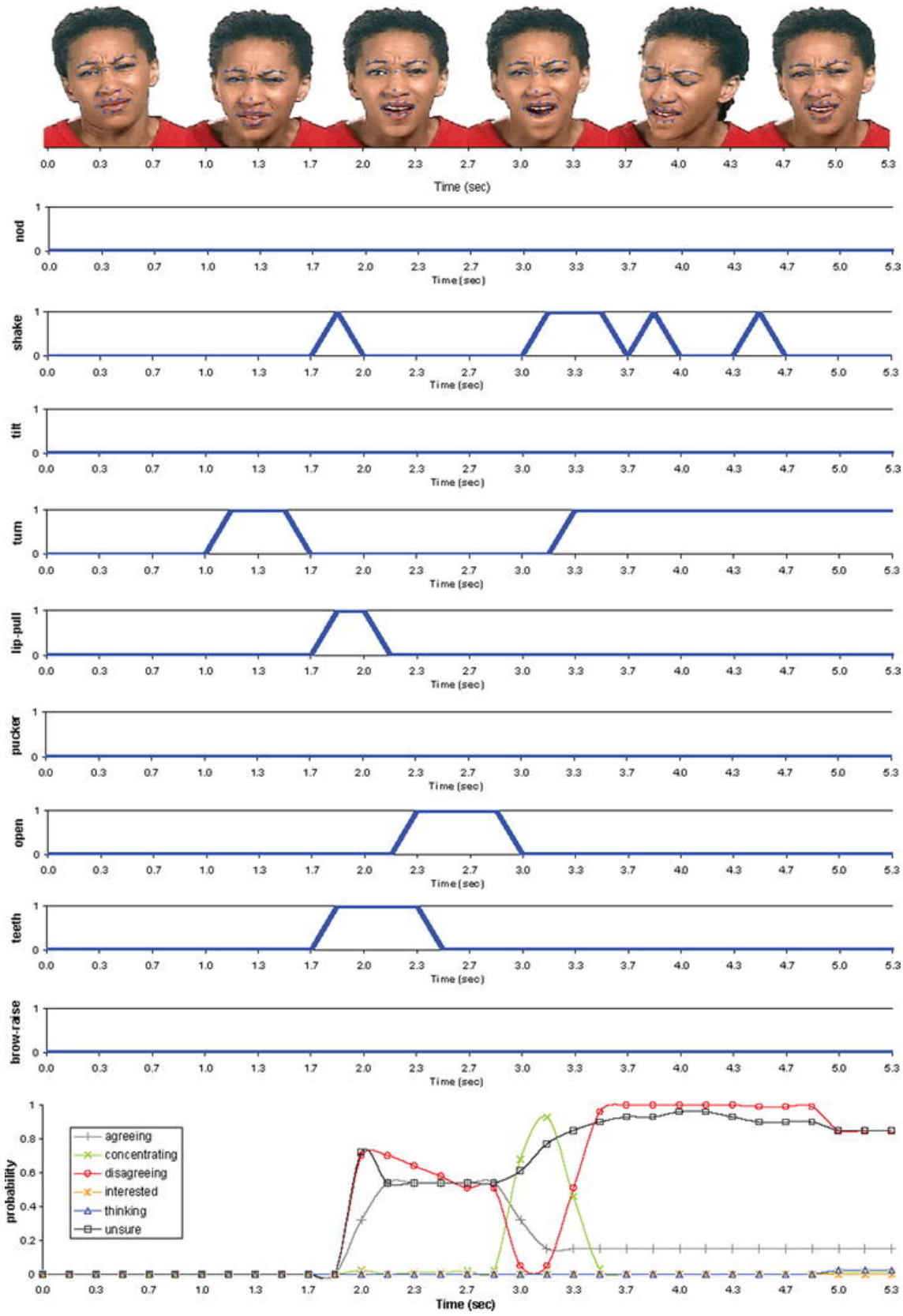

Fig. 8 The output at the different levels of the mind-reading system for a 5.5-s long video portraying the mental state discouraging, which belongs to the mental state group disagreeing. The *top row* shows selected frames from the video sampled every 1 s. The following *nine rows* depict the output of the head and facial display recognition, which constitute the input to the DBN networks. The *final row* shows the output of

each mental state classifier. Since the overall probability of disagreeing is 0.75, this is an example of a correct classification

---

The recognition results are described using a confusion matrix in Table 1. The rows of the matrix describe the classification results of each mental state class. The number of times that a certain mental state was recognized is stated in columns. The last column states the true positive (TP) or classification rate for each class. It is given by the ratio of videos correctly classified as mental state **X** to the total number of videos truth-labelled as **X**. The classification rate is highest for concentrating (88.9%) and lowest for thinking (67.7%). Note that since this is effectively a 6-way forced choice procedure, chance responding is 16.7%. Hence, the recognition results for all the mental state classes are significantly above chance level.

[Table 1 will appear here. See end of document.]

The bottom row in Table 1 describes the false positive (FP) rate for each class. It is computed as the ratio of videos falsely classified as **X** to the total number of videos truth-labelled as anything but **X**. The FP rate is highest for concentrating (10.9%) and lowest for disagreeing (0.7%). The results are also summarized as a 3D bar chart in Fig. 9. The horizontal axis represents the classification results of each mental state class. The percentage of recognition of a certain mental state is represented along the *z*-axis.

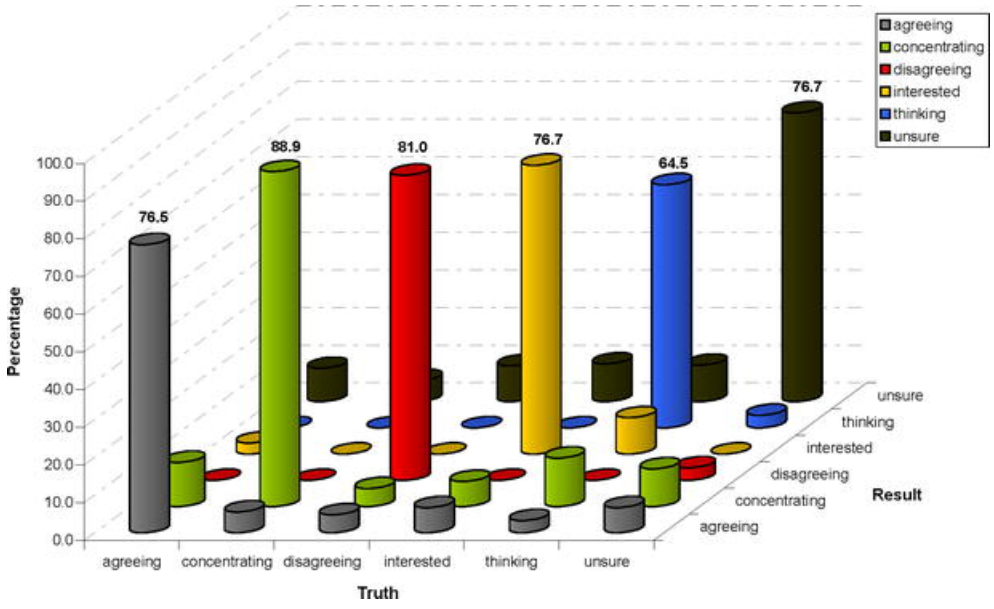

Fig. 9 Confusion matrix of the recognition results in, shown as a 3D bar chart. The *x-axis* (truth) depicts the input data for each of the mental states. The *z-axis* (result) represents the classification result for the input. The *y-axis* (percentage) shows the percentage of classifications per class

---

For a mean FP rate of 4.7%, the overall accuracy of the system is 77.4%. These results have been compared with a previous experiment in which the performance of a group of people in recognizing complex mental states in a similar set of videos from the Mind-reading DVD had been tested (see el Kaliouby et al. [24] for more details). In that experiment, human recognition rate reached an upper bound of 71.0%. Thus, the accuracy of the automated mind-reading system in classifying complex mental states from videos of the Mind-reading DVD compares favourably to that of humans.

We are currently evaluating the generalization performance of the automated mind-reading system. The idea is to train the system on videos from the Mind-reading DVD, and test its performance on a previously unseen corpus with different recording conditions and subjects than those used in training the system. The generalization performance of a system is an important indicator to how well the system does outside of lab settings.

Discussion

Although the overall accuracy of the system compares favourably to that of human performance on a similar recognition task, an accuracy of 77.4% is generally not sufficient for use with a real-time application such as the emotional hearing aid. After all, if users are going to depend on the tool for advice on how to react to other people, the reliability of the system in terms of recognition rate needs to be higher, while the FP rate needs to be lower.

The upper ceiling of 70–80% reported for the automated mind-reading system and with humans too, suggests that other information sources, such as other modalities and contextual cues should be integrated to boost the reliability of complex mental state recognition. Humans make considerable use of the contexts in which facial expressions occur to assist interpretation [11, 14], including situational context. Howlin et al. [22] define situation-based emotions as those that involve inferring a person's emotional state from a particular sequence of events.

It is possible to extend the function of the emotional hearing aid to save information about the context in which an interaction occurs. This information can be integrated in the inference process along with the video input to boost the reliability of the results. A simple implementation of location-context would entail defining several profiles which reflect the various situations the child can be in such as "in school" or "in playground". The contexts would have to be explicitly selected, but as the tool gets more sophisticated, more detailed profile information would be deduced automatically by the system.

In theory, the users of the emotional hearing aid would benefit from the integration of context and other modalities in the system since many people diagnosed with Asperger syndrome have problems integrating mental state concepts from facial expression into wider contexts (e.g. previous encounters with a person or an environment). The integration of context and other modalities within the emotional hearing aid, and a study of its effect on recognition accuracy, is a research direction, worth pursuing.

Conclusions and future work

The emotional hearing aid is a portable assistive computer intended to help children diagnosed with Asperger syndrome read, understand and react to facial expressions in a socially-appropriate way. The design draws inspiration from the "emotional indexing" approach to teaching emotions to children with autism in order to automate the process of empathizing. In developing the emotional hearing aid, an automated mind-reading system that infers complex mental states from facial expressions in real-time video, and a reaction advisor that suggest appropriate reactions for the user to take in real time have been implemented.

This paper has reported progress on two fronts with respect to the implementation and validation of the components of the emotional hearing aid. First, the experimental evaluation of the automated mind-reading system on the following classes of complex mental states: agreeing, concentrating, disagreeing, interested, thinking and unsure. For a mean FP rate of 4.7% an overall recognition accuracy of 77.4% was reported. Second, a rule-based implementation of the reaction advisor has been presented, which takes into account the persistence, confidence and intensity of an inference when suggesting an appropriate reaction.

Admittedly, the emotional hearing aid is an ambitious project and it is no where near being available for use outside of lab settings. Developing and verifying the tool presents a number of challenges which span different research areas.

First, the reliable, real-time, automated inference of a wide range of mental states, including the complex ones, from facial expressions in video continues to challenge the state-of-the-art methods in machine vision and machine learning. This paper has presented novel results in the recognition of six classes of complex mental states which include cognitive states such as thinking as well as affective ones such as interested. The results, though promising, need to be improved before the mind-reading system can be used with an application like the emotional hearing aid.

Second, the rules that govern how people read other people's mind from nonverbal cues and how they react accordingly (mind-reading and sympathizing) continue to challenge researchers in the behavioural sciences. From an engineering point of view, this means that there is no "rule-book" to follow when automating these processes. Instead, statistical machine learning and data-driven approaches have to be combined with the limited domain knowledge that is available to encode the automated system's functions. In the current implementation of the automated mind-reading system, DBNs are used to represent the stochastic nature inherent in facial behaviour and the facial signals of complex mental states. Future directions include the re-implementation of the reaction advisor using partially observed Markov decision processes, so that the utility of the actions is also learnt from data rather than hard-coded as in the current rule-based implementation.

Finally, technical challenges aside, the emotional hearing aid raises a number of questions from a usability and accessibility point of view. Will the auditory or visual output of the system cognitively overload the users of the system? Will the system be able to encode the common sense knowledge that complements mind-reading and other social processes such that the suggested reactions of the system are indeed useful? Will the emotional hearing aid really help people with Asperger syndrome engage with more social interactions in the way it is intended? Are there any side effects of this technology? For example, will the users of the aid learn to depend on it and stop acquiring new knowledge about emotional and social understanding? It is hard to predict the answers to these questions without conducting proper user studies which address each of these concerns. As a more complete prototype of the tool becomes available, the tool will be deployed in a number of user studies to gain feedback on usability.

It is the authors' belief that the emotional hearing aid has the potential to offer children with Asperger syndrome more opportunities to engage in natural social interactions, beyond the hypothetical scenarios used in a teaching environment. The tool is designed to provide assistance even when the child's carer is not available and ensures that events are accessible even after their occurrence for discussion and learning purposes.

Acknowledgements The authors would like to thank Alex Birkby for implementing the reaction advisor in the context of his Computer Science Diploma dissertation at the University of Cambridge. We would also like to thank Professor Simon Baron-Cohen and Ofer Golan, at the Autism Research Centre, University of Cambridge for inspiring discussions about the automated mind-reading system and for making the Mind-reading DVD available to our research, and the anonymous

reviewers for their valuable input. This research was funded by the Computer Laboratory's Wiseman Fund, the Overseas Research Student Award, the Cambridge Overseas Trust, and Newnham College Studentship Research Award.

References

1. APA (1994) DSM-IV diagnostic and statistical manual of mental disorders, 4th edn. American Psychiatric Association, Washington DC
2. Attwood T (1998) Asperger's syndrome: A Guide for Parents and Professionals. Jessica Kingsley, Philadelphia
3. Baron-Cohen S (1995) Mindblindness: An Essay on Autism and Theory of Mind. MIT, Cambridge
4. Baron-Cohen S, Wheelwright S (1999) Obsessions in children with autism or Asperger syndrome: a content analysis in terms of core domains of cognition. Br J Psychiatry 175:484–490
5. Baron-Cohen S, Golan O, Wheelwright S, Hill J (2004) Mind-reading: the interactive guide to emotions. Jessica Kingsley, London (http://www.jkp.com/mindreading)
6. Baron-Cohen S, Wheelwright S, Lawson J, Griffin R, Hill J (2004) The exact mind: empathising and systemising in autism spectrum conditions. In: Goswami U (ed) Handbook of cognitive development. Blackwell, Oxford
7. Batty M, Taylor M (2003) Early processing of the six basic facial emotional expressions. Cogn Brain Res 17:613–620
8. Birdwhistell R (1970) Kinesics and Context. University of Pennsylvania Press, PA
9. Birkby A (2004) The emotional hearing aid, Computer Science Diploma Dissertation. Computer Laboratory at the University of Cambridge
10. Blocher K (1999) Affective social quotient (ASQ): teaching emotion recognition with interactive media and wireless expressive toys. S.M. Thesis, MIT, Cambridge
11. Bruce V, Young A (1998) In the eye of the beholder: the science of face perception. Oxford University Press, New York
12. Cheng L, Kimberly G, Orlich F (2003) KidTalk: online therapy for Asperger's syndrome. Technical Report, Social Computing Group, Microsoft Research
13. Dautenhahn K, Billard A (2002) Games children with autism can play with Robota, a Humanoid Robotic Doll. In: Keates S, Clarkson PJ, Langdon PJ, Robinson P (eds) Proceedings of the 1st Cambridge workshop on Universal Access and Assistive Technology [CWUAAT]. Universal Access and Assistive Technology. Springer, London, pp 179–190
14. Edwards, K (1998) The face of time: temporal cues in facial expression of emotion. Psychol Sci 9:270–276
15. Ekman P, Friesen WV (1978) Facial action coding system: a technique for the measurement of facial movement. Consulting Psychologists Press, Palo Alto
16. Facetracker (2002) Neven Vision's Facial Feature Tracking SDK (http://www.nevenvision.com/products.html)
17. Fling E (2000) Eating an artichoke: a mother's perspective on Asperger syndrome. Jessica Kingsley Publishers Ltd, Philadelphia
18. Frith U (eds) (1991) Autism and Apserger syndrome, Cambridge University Press, London
19. Goldman A, Sripada C.S (2005) Simulationist models of face-based emotion recognition. Cognition (in press)
20. Hirose M, Kijima R, Shirakawa K, Nihei K (1997) Development of a virtual sand box: an application of virtual environment for psychological treatment. In: Riva G (ed) Virtual reality in neuro-psycho-physiology: cognitive, clinical and methodological issues in assessment and treatment. IOS Press, Amsterdam

21. Hoey J, Little J (2004) Decision theoretic modeling of human facial displays. In: Proceedings of European conference on computer vision
22. Howlin P, Baron-Cohen S, Hadwin J (1999) Teaching children with autism to mind-read: a practical guide for teachers and parents. Wiley, New York
23. el Kaliouby R, Robinson P (2003) The emotional hearing aid: an assistive tool for autism. in the proceedings of the 10th international conference on human–computer interaction (HCII): universal access in HCI, vol 4, pp 68–72
24. el Kaliouby R, Robinson P, Keates S (2003) Temporal context and the recognition of emotion from facial expression. In: Proceedings of the 10th international conference on human–computer interaction (HCII): human–computer interaction, theory and practice, vol 2, pp 631–635
25. el Kaliouby R, Robinson P (2004) Real-time inference of complex mental states from facial expressions and head gestures. In: IEEE international workshop on real time computer vision for human–computer interaction at CVPR
26. el Kaliouby R, Robinson P (2004) Mind-reading machines: automated inference of cognitive mental states from video. In: Proceedings of IEEE conference on systems, man and cybernetics, The Hague
27. Kozima H, Yano H (2001) Designing a robot for contingency-detection game. Working Notes Workshop Robotic& Virtual Interactive Systems in Autism Therapy. University of Hertfordshire, Technical Report No 364
28. MacKenzie IS (1995) Virtual environments and advanced interface design. In: Input devices and interaction techniques for advanced computing. Oxford University Press, Oxford, pp 437–470
29. Moore D, McGrath P, Thorpe J (2000) Computer-aided learning for people with autism-a framework for research and development. Innov Educ Training Int 37(3):218–228
30. O' Connell S (1998) Mindreading: how we learn to love and lie. Arrow Books, London
31. Pantic M, Rothkrantz L (2000) Automatic analysis of facial expressions: The state of the art. IEEE Trans Pattern Anal Mach Intell 22:1424–1445
32. Premack D, Woodruff G (1978) Does the chimpanzee have a theory of mind? Behav Brain Sci 4:515–526
33. Phillips AT, Wellman HM (2002) Infant's ability to connect gaze and expression to intentional actions. Cognition 85:53–78
34. Posamentier M, Abdi H (2003) Processing faces and facial expressions. Neuropsychol Rev 13(3):113–144
35. Rozin D, Cohen AB (2003) High frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of naturally occurring facial expressions of American. Emotion 3(1):68–75
36. Strickland D (1996) A virtual reality application with autistic children, presence. Teleoper Virtual Environ 5(3):319–329
37. Surakka E, Hietanen JK (1998) Facial and emotional reactions to duchenne and nonduchenne smiles. Int J Psychophysiol 29(1):23–33
38. Turk M, Kolsch M (2004) Emerging Topics in Computer Vision, chapter Perceptual Interfaces. Prentice Hall, Englewood Cliffs
39. Werry I, Dautenhahn K, Ogden B, Harwin W (2001) Can social interaction skills be taught by a social agent? the role of a robotic mediator in autism therapy. Proceedings CT2001, the fourth international conference on cognitive technology. Lecture Notes in Computer Science, sub series Lecture Notes in Artificial Intelligence, Springer, Berlin Heidelberg New York

Table 1 Confusion matrix of recognition results of the 164 videos

Class	Agreeing	Concentrating	Disagreeing	Interested	Thinking	Unsure	TP (%)
Agreeing	26	4	0	1	0	3	76.5
Concentrating	1	16	0	0	0	1	88.9
Disagreeing	1	1	17	0	0	2	81.0
Interested	2	2	0	23	0	3	76.7
Thinking	1	4	0	3	20	3	64.5
Unsure	2	3	1	0	1	23	76.7
FP (%)	5.4	9.6	0.7	3.0	0.8	9.0	77.4

The rows of the matrix represent the classification results of each mental state class. The number of times that a certain mental state was recognized is stated in columns. The last column states the true positive (TP) or classification rate for each class. The last row states the false positive (FP) rate for each class