Context-Aware Automated Analysis and Annotation of Social Human-Agent Interactions

TOBIAS BAUR, Augsburg University
GREGOR MEHLMANN, Augsburg University
IONUT DAMIAN, Augsburg University
PATRICK GEBHARD, DFKI GmbH
FLORIAN LINGENFELSER, Augsburg University
JOHANNES WAGNER, Augsburg University
BIRGIT LUGRIN, Augsburg University
ELISABETH ANDRÉ, Augsburg University

The outcome of interpersonal interactions depends not only on the contents that we communicate verbally, but also on nonverbal social signals. As a lack of social skills is a common problem for a significant number of people, serious games and other training environments have recently become the focus of research. In this work we present NovA (**Non**verbal behavior **A**nalyzer), a system that analyzes and facilitates the interpretation of social signals automatically in a bi-directional interaction with a conversational agent. It records data of interactions, detects relevant social cues, and creates descriptive statistics for the recorded data with respect to the agents behavior and the context of the situation. This enhances the possibilities for researchers to automatically label corpora of human-agent interactions and to give users feedback on strengths and weaknesses of their social behavior.

Categories and Subject Descriptors: H.1.2 [**User/Machine Systems**]: Human information processing

General Terms: Social Signal Processing, Serious Games, Virtual Agents, Affective Computing

Additional Key Words and Phrases: Social Cue Recognition, Virtual Job Interviews, Serious Games, Automated Behavior Analysis, Interaction Design

ACM Reference Format:
Tobias Baur, Gregor Mehlmann, Ionut Damian, Patrick Gebhard, Florian Lingenfelser, Johannes Wagner, Birgit Lugrin and Elisabeth André, 2015. Context-Aware Automated Analysis and Annotation of Social Human-Agent Interactions. *ACM Trans. Interact. Intell. Syst.* 5, 2, Article 11 (July 2015), 31 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

This work has been partially funded by the European Commission within FP7-ICT-2011-7 (Project TARDIS, grant agreement no. 288578) and has received funding from the European Unions Horizon 2020 research and innovation programme (Project ARIA-VALUSPA, grant agreement no. 645378).
We thank the teachers Bernhard Pietzowski and Richard Endrass from the Parkschule Stadtbergen for helping to organize the evaluation study and the pupils for their participation. We also thank Julia Brombach and Claudia Lange from the Career Service of the Augsburg University for volunteering as practitioners. Furthermore, we thank the people from Charamel GmbH for their continuous support and for providing us with the virtual character Gloria.
Author's addresses: T. Baur, G. Mehlmann, I. Damian, F. Lingenfelser, J. Wagner, B. Lugrin and E. André, Human Centered Multimedia, Augsburg University, Universitätsstr. 6a, D-86159 Augsburg, Germany. P. Gebhard, DFKI Stuhlsatzenhausweg 3, D-66123 Saarbrücken, Germany

DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

In interpersonal communication nonverbal social signals play an important role as they convey a large part of information that may have an even deeper influence on the outcome of a conversation than the word meanings themselves. In stressful situations, such as job interviews, humans often employ nonverbal behavior that has a negative impact on how they are perceived. However, particularly in such scenarios it is especially important to give the impression of confidence and attentiveness to convince others of one's strengths and competence. In recent years, a variety of research projects tackled the problem of tutoring users to improve social behaviors. With the help of serious games or similar training environments such systems aim to simulate interpersonal communication with conversational agents or robots. Compared to real human trainers, this also comes along with several advantages for training purposes, such as the avoidance of additional stress, lower costs, permanent availability and reproducibility of training sessions. Examples for rather specific use cases are public speeches [Batrinca et al. 2013], social humor situations [Niewiadomski et al. 2013; Mancini et al. 2014], inter-cultural communication [Endraß et al. 2013], negotiation scenarios [Traum et al. 2012] or psychotherapy [Kang et al. 2012].

A first step to help people effectively improve their performance typically consists of finding critical behaviors during a training session. Therefore, it is common practice for experts to take notes on what attracted their attention. In more sophisticated environments data is annotated alongside audiovisual recordings following a predefined annotation scheme, so users can reflect on their behavior while feedback is given by experts. A negative aspect of this approach is that the annotation of such recordings requires several iterations and is often extremely time-consuming. Additionally, annotations often vary a lot between annotators, as subjective perception differs from person to person. Some more advanced training systems partly process human behavior, such as the amount of smiles or the prosody [Hoque et al. 2013], but are still limited in the range of modalities that are automatically analyzed. Depending on the application focus, the main goals from a social coach's point of view are particularly the analysis and control of high-level concepts, such as signaling engagement and attention, the establishment of rapport and trust and the effective and fluent communication through grounding. Such concepts are based on bi- and multi-directional behavior patterns that are generated by the temporal alignment of the interlocutors' actions. Examples for this purpose are directed gaze, declarative pointings, gaze following, turn-taking signals, backchanneling and mirroring. Depending on the context of the situation, such behavior patterns may vary considerably and social signals need to be interpreted accordingly. In state-of-the-art training systems, such phenomena have not been in focus as such systems are at the most concerned only with the user's nonverbal signals.

In this article we present a system, named Nonverbal behavior Analyzer (NovA), for automating the annotation process using real-time social signal processing techniques in combination with interaction modeling concepts. The system was originally developed as part of the TARDIS [Anderson et al. 2013] EU-project. TARDIS aims to help young people improve their social skills during job interviews by providing a training environment with a virtual recruiter. The proposed system also provides concepts and tools for future research in behavior analysis in human-human and especially human-agent interactions. By analyzing both the user's behavioral cues and the agent's interaction cues, the system allows to automatically annotate a social human-agent interaction. Such cues are computed in real time during the interaction by using sensing devices, such as a Microsoft Kinect to recognize behavioral characteristics, and by recording the agent's verbal and non-verbal behavior. Further, scenario specific meta information is logged to improve the analysis of the situation by considering conversational context. Figure 1 gives an overview on the architecture of our system. We see the main contributions of this work in:

- *A wide range of real-time multimodal social cue recognition algorithms for automated annotations.* As we aim to automate the annotation process of human behavior we created a set of general real-time social cue recognizers for multiple modalities. This set is built on a plugin system, that allows recognizers to be turned on or off depending on the user's needs and available hardware.

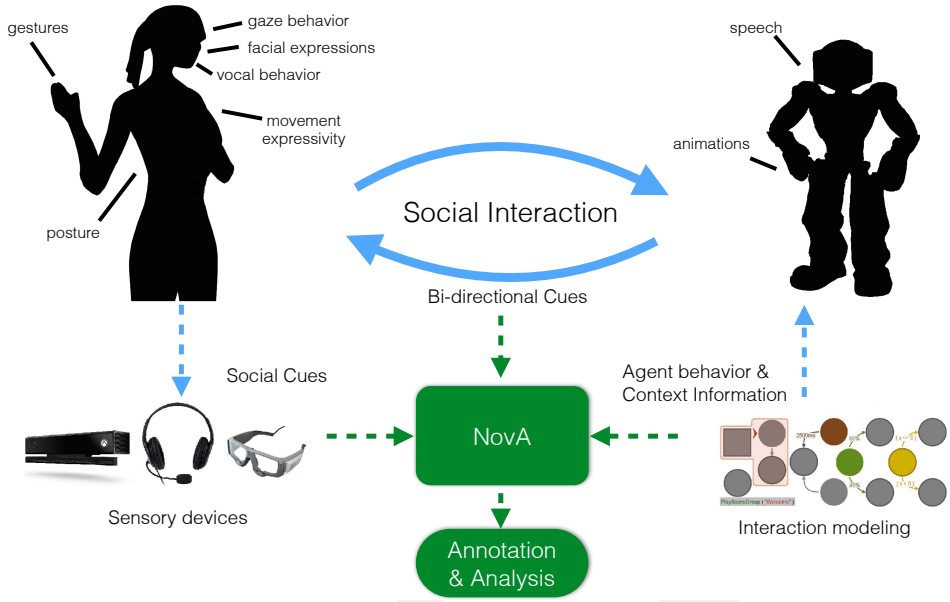

Fig. 1. The general architecture of the NovA system. It recognizes social cues from human behavior with sensors, as well as agent behavior and context information from an interaction modeling component. Additionally bi-directional cues, that are based on behavior of both interlocutors are processed. This information is used by the NovA System to analyze and annotate the interaction automatically.

The recognizers have been developed in cooperation with psychologists and social-coaching trainers. They are also customizable for specific needs in various scenarios. Once adjusted, recognizers will return objective and comparable observations, which is essential for the automated labeling process.

- *Annotation of agent behavior, dialog context as well as bi-directional interaction cues.* A crucial problem is that the simple occurrence of a behavioral cue does not necessarily allow us to make straight-forward assumptions about the users intentions. As an example, a smile can be a sign of happiness (which is assumed by most systems), but could, for example also be an expression of embarrassment (also see [McKeown et al. 2015]). Therefore to interpret behavioral cues correctly, the dialog context has to be taken into account. If the agent is telling a funny joke and the user reacts with a smile, one could assume that the agent amuses the user. If the agent discourages the user, but he or she still smiles, it is more likely that the user is overplaying the situation. In addition, there are tightly coordinated behaviors that are elicited by two or several interlocutors. Such behaviors, further called bi-directional cues, include phenomena, such as grounding, mutual gaze, mirroring or backchanneling, that are relevant to the analysis of dialogue dynamics. In order to be able to correctly interpret behavior signals, it is highly important to annotate not only the user's, but also the agent's behaviors.
- *Context-aware analysis of the interaction by considering meta information, such as current topics or conditions.* Another aspect that has to be considered is the influence of the conversational context on the assessment of behaviors. While some behaviors are more appropriate in one situation, they might not be in another. To give an example: A high amount of hand movements during a job interview might be interpreted as a sign of engagement when the candidate is talking about his or her hobbies. As a consequence, the candidate's behavior would result into a positive assessment by the interviewer. However, a candidate who is fidgeting with his or her hands while responding to a difficult question would rather be evaluated negatively because the candidate's behavior might

be taken a sign of nervousness. In our system, context information is used to influence the model for the automated analysis of the user's behavior.

- *Real-time inferring of social attitudes based on probabilistic models to receive implicit information.* Information on the users' social attitude, such as their level of attentiveness, is typically conveyed by more than one signal behavior at a time. We therefore suggest the use of probabilistic models to infer social attitudes based on multiple inputs, such as (bi-directional) multimodal behavior cues and context information. Creating such continuous outputs helps pointing out the most relevant incidents of a social interaction which is especially worth aspiring to in social coaching environments. This information further is suited to be used as real-time input for an agent to react to the user's current social attitude.
- *Automated analysis and visualization of discrete and continuous annotations and automated statistical analysis tests of the interaction.* We created a graphical user interface for the visualization of detected behaviors. Inspired by classical annotation tools, it presents tier-based labeling. Additionally it also features continuous time-line diagrams, bar and pie charts, heat maps and social attitude outputs to point out characteristics encountered in interpersonal interaction in a fully automated manner. Experts might use such information when counseling users about their behaviors in coaching scenarios.

In the next chapter we review related work in the area of annotation tools, automated behavior analysis and interaction modeling. Next we introduce the TARDIS job interview game, to illustrate a use-case application that makes use of our general architecture and concepts. We then describe our four main components: the social cue recognition module, the interaction modeling module, the social attitude component and the analysis and annotation tool that serves as coaching user interface. Finally we discuss experiences from evaluating our use-case application in a field study at a local school that showed that a combination of the suggested concepts leads to significant improvements in coaching compared to traditional methods.

2. RELATED WORK

Our proposed system combines work on annotation tools with technologies to automatically analyze human behavior as well as interaction modeling techniques. The user interface of NovA has been inspired by existing annotation tools and makes use of multiple tracks to code relevant social features. However, unlike conventional annotation tools, our system performs the segmentation and labeling of the data completely automated. NovA distinguishes from earlier work on automated behavior analysis by considering not only the behavior of the user, but as well as that of the agent and the dynamics that arise during the conversation in a social training scenario. In our work we further investigate interaction and conversation context for the improved analysis and annotation of human-agent conversations.

Progress in the field of analyzing human social behavior has been boosted by a variety of annotation tools that facilitate the labeling of corpora at different levels of granularity following a pre-defined coding scheme. Examples include European distributed corpora project Linguistic ANotator (Elan) [Wittenburg et al. 2006], ANotation of VIdeo and Language (Anvil) [Kipp 2013], and Exmaralda [Schmidt 2004] which offer layer-based tracks to insert time-anchored labeled segments. Another example is FEELtrace [Cowie et al. 2000], a tool that allows an observer to track the emotional content of an audio-visual stimulus over time based on activation-evaluation space. A newer edition of FEELtrace is the General trace (Gtrace) [Cowie et al. 2012] with the ability to let people use their own dimensions and scales. While it is unquestionable that these tools offer much help in describing audio-visual material with a high level of detail, they offer only little automation. However, since creating descriptions for several hours of interaction remains an extremely time consuming task, methods to automate the coding process are highly desirable.

Techniques for the automated analysis of social behaviors patterns were pioneered by Pentland and his group at MIT Media Lab with the development of wearable devices, so-called sociometers, to capture people's verbal and non-verbal signals. They investigated not only the social behaviors of

people engaged in face-to-face conversations [Curhan and Pentland 2007], but also analyzed interaction patterns from larger groups of people using smart-phones with dedicated sensors [Pentland 2007]. To analyze social behaviors, a large variety of verbal and non-verbal cues has been taken into account. [Dong et al. 2007] analyze speech activity and fidgeting, that is the amount of movement in a person's hands and body, to detect functional roles in a group. [Hung and Gatica-Perez 2010] studied audio cues (such as overlapping speech), video cues (such as motion energy), and audio-visual cues (such as the amount of movement during speech) to determine the level of group cohesion in meetings. Methods have been developed to detect social attitudes from various modalities including facial expressions [Sandbach et al. 2012], gestures [Caridakis et al. 2006; Michelet et al. 2012; Mahmoud et al. 2013], speech [Vogt et al. 2008], postures [Kleinsmith and Bianchi-Berthouze 2011] and physiological measurements [Kim and André 2008]. [Nakano and Ishii 2010] developed a method to assess user engagement from eye gaze in user-agent interactions. [Thompson and Bohus 2013] developed a system that supports feature annotation for model building using a variety of machine learning techniques. Also, multimodal approaches to improve emotion recognition accuracy have been reported, mostly by exploiting audiovisual combinations [Camurri et al. 2005; Scherer et al. 2012; Sebe et al. 2006; Gunes et al. 2008]. Results suggest that integrated information from audio and video leads to improved classification reliability compared to a single modality. Even though the role of context has been recognized in the area of social signal processing [Pantic et al. 2005], work that actually exploits context information to improve recognition rates is rare. An example includes the work by [Conati and Maclaren 2009] who describe a tutor agent that makes use of a Bayesian Network for interpreting the learner's behavior. Following bio-sensor and UI inputs, as well as context in form of previously prompted personality traits, the system tries to infer emotions of the learner. The Virtual Human Toolkit [Gratch et al. 2013] provides social signal analysis combined with dialogue act-based generation of nonverbal agent conversational behavior. Context information is used to refine the analysis of social signals. While this is modeled within dialogue acts, our approach uses an explicit interaction model to represent contextual information (e.g. discourse phase or agent states). An example of how context is used for improving social signal processing can be found in [Morency et al. 2007]. By considering specific types of questions they achieved improvements in recognizing head movements in a human-agent interaction. The SEMAINE [Schröder et al. 2012] platform focuses on the analysis of emotions in an interactive communicative setting. Therefore visual and acoustic signals in speech and listening phases during an interaction with a virtual agent are analyzed. In our approach we consider an extended set of modalities for the analysis of user behavior. Further by applying a more fine-grained interaction dialogue model and control we extend the analysis to bi-directional behavior and discourse context information.

3. APPLICATION SCENARIO

In the following, we present our approach by means of an illustrative example application. The EU-funded project TARDIS[1] attempts to support young adults in job interviews by developing a scenario-based serious game with virtual agents acting as recruiters. One large issue Europe faces is the rising number of young people who are not in employment, education or training (NEETs). NEETs often have underdeveloped socio-emotional and interaction skills [Hammer 2000], such as a lack of self-confidence, lack of sense of their own strengths or social anxiety [Pan et al. 2012]. This often leads to problems in various critical situations, such as job interviews, where they need to convince the recruiter of their fit in a company. To address this issue, many European countries have specialized inclusion centers meant to aid young people secure employment through coaching by professional practitioners. Unfortunately, such an approach is expensive and time-consuming. Considering this, technology-enhanced solutions, such as digital games, present themselves as viable and advantageous alternatives to the existing human-to-human coaching practices.

[1]Training young Adult's Regulation of emotions and Development of social Interaction Skills - **http://tardis-project.eu**

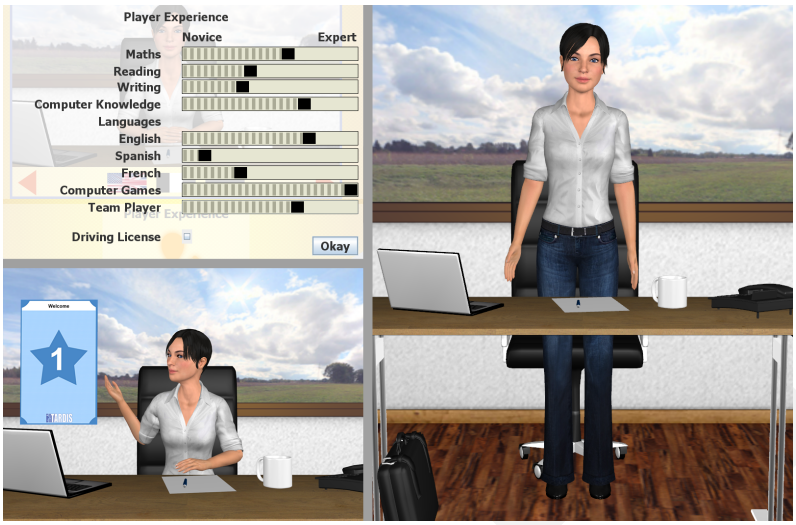

Fig. 2. Player experience, scenario, and welcome game phase.

The *TARDIS Social Cue Training Game* is an approach for supporting young adults in job interviews. It employs gaming techniques and methods to motivate adolescents and young adults to improve their social skills. For the game we make use of the virtual character Gloria as seen in Figure 2, that has been developed by Charamel GmbH. The scenario is set up in a virtual environment that is modeled like a typical office environment (see Figure 2, right side).

The game is structured similarly to a job interview. It features three interview phases, namely *Welcome*, *Company Presentation* and *Strength and Weaknesses*. Prior to the *Welcome* phase, the user is given a short introduction into how the system works. At the start of the game, the user is also asked to provide some information about general skills and background (see Fig. 2, left side, top). This information is used throughout the game to adapt the flow of the interview to the user's profile. For example, if the user stated to be experienced in a specific language, the game will ask questions related to her or his expertise in that area.

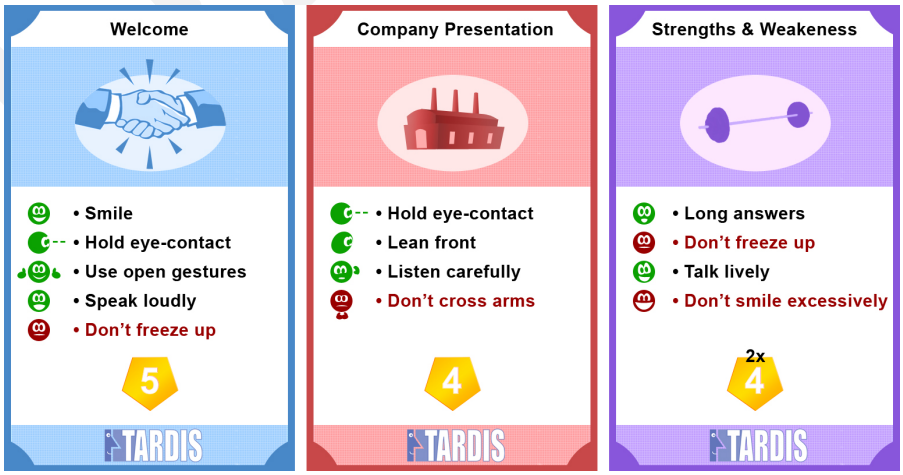

Fig. 3. Social cue action cards for each game phase.

While playing the game, the participant is asked to adapt to specific social task situations, which are related to the game phase. The *Welcome* phase is related to the social task of presenting one self. The subsequent phase of *Company Presentation* is related to the task of listening carefully, and the last phase (*Strength and Weaknesses*) is related to conversation about the user's profile. The user is expected to adapt her or his behavior to each phase. What kind of behavior is appropriate to a specific phase, is described on physical game cards (see Figure 3). Each game card contains several social cues which the user should or should not perform. For example, the *Welcome* card instructs the user to 1) smile, 2) hold eye contact, 3) use open gestures, 4) speak loudly, and 5) don't freeze up. These social cues have been identified by experts, e.g. social workers and job recruiters.

The cards are given to the user prior to the interaction. During the game, the virtual agent informs the user before each phase which game card is relevant in the upcoming phase and only proceeds with the interview once the user acknowledges having read the game card (Figure 2, left side, bottom). Furthermore, the behaviors are also displayed directly on the game screen using graphical symbols. These symbols change in appearance depending on whether the user performed the underlying social cue or not. For example, if the user shows an appropriate amount of smiling during the *Welcome* phase, the smile symbol gets highlighted.

Fig. 4. Reward for smiling during the Welcome phase.

To encourage adequate behaviors, the system also scores the users based on their performance. More precisely, every time a user behaves in compliance with the game card, i.e. performs a requested social cue, she or he receives a point towards the total score. Some of the cues have to be performed (or not performed) for the whole duration of the interview phase (which represents the conversational context), e.g. *do not freeze up*. After a game session a professional job interview coach gives feedback on how to improve with help of the NovA coaching tool that will be described in later sections.

4. SOCIAL CUE RECOGNITION

As a first step for the automated analysis of user behavior we created a set of social cue recognizers. Our system uses the Social Signal Interpretation Framework[2] (SSI) [Wagner et al. 2013] which offers capabilities to record, analyze and recognize human behavior in real-time. In particular, SSI supports the parallel and synchronized processing of data from multiple sensory devices, such as cameras, multi-channel microphones and various physiological sensors. It further supports fusion of multiple channels, machine learning techniques and the synchronization between multiple computers. Figure 5 gives an overview on the functionality of the SSI Framework.

[2]**http://openssi.net**

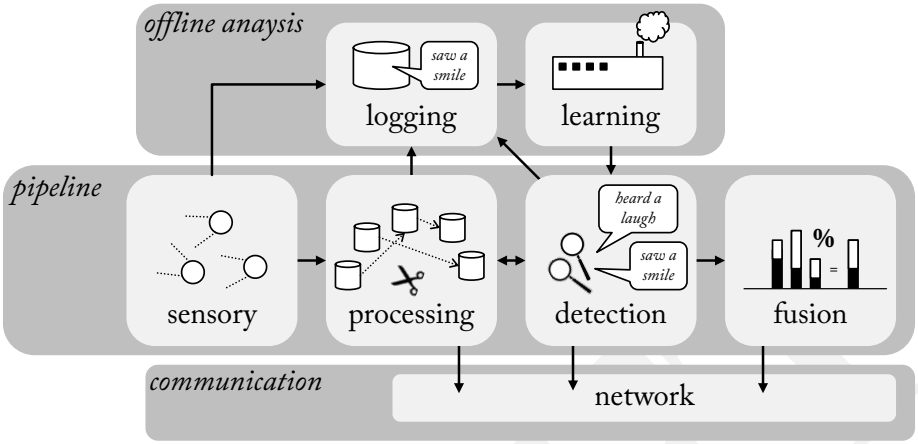

Fig. 5. Overview of the SSI Recording and Online Processing Framework.

For our system we make use of various sensory devices to capture human behavior data. One important sensor for the detection of social signals in human behavior is a depth camera, specifically the Microsoft Kinect which has a number of advantages: It is rather low-cost, it does not require any time-consuming configuration and it is relatively robust against lighting conditions. Furthermore, there are software development kits for skeleton and face tracking available for the Kinect sensor that provide a good starting point for human behavior analysis. SSI also enables the easy integration of more sophisticated sensors, such as a motion capture suit (e.g. Xsens MVN), to achieve more robust tracking. Figure 6 shows the system's interface for the recording of user data. It illustrates the skeleton and face tracking, RGB video, and audio feature capturing.

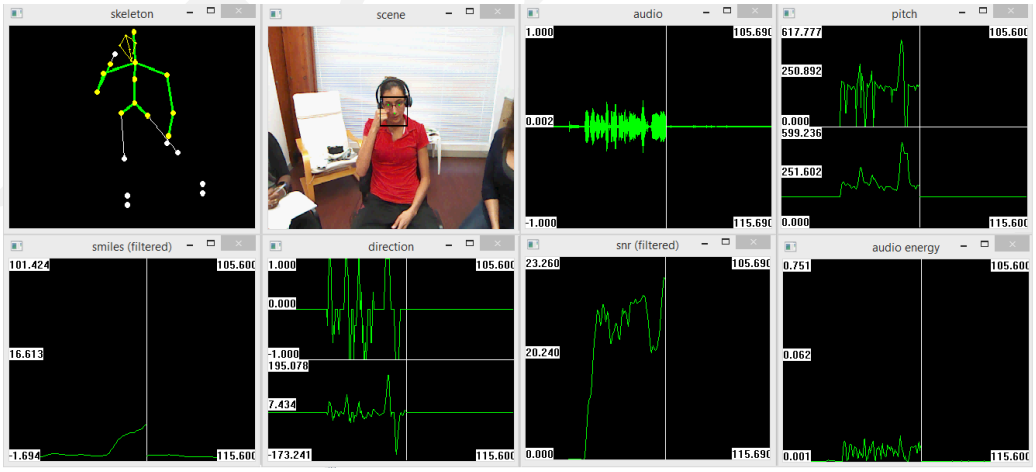

Fig. 6. Recording and processing of Social Signals in real-time. This instance is showing the skeleton, tracked by a Microsoft Kinect, the RGB Image with facial feature detection, including smiles, the audio intensity, pitch and pitch direction, signal-to-noise ratio and the audio energy.

Our system supports continuous, as well as event-based annotations. In the case of continuous annotations, a value referring to a particular attribute, such as the distance between the two hands of a person or the orientation of the head, that is computed continuously over time. For event-based

annotations, we integrated a mechanism that triggers an event each time when the beginning or the end of a social cue is detected. In this process, the event recognizers act on a rather strictly defined level.

These events are additionally saved in an XML-based structure including a synchronized timestamp and the event's duration for later offline analysis. The next subsections describe modules for social cue recognizers in different modalities. As our system is modular, recognizers can be turned off if specific modalities are not of interest for an analysis or required hardware is not available. Further, fine tuning recognizers to specific conditions and needs is possible by adjusting simple options.

4.1. Gesture & Posture Detection

For event-based gesture and posture analysis, we make use of the Full Body Interaction (FUBI) system [Kistler et al. 2012] which has been integrated in our framework. Overall, three categories of body postures and gestures are supported by FUBI where more complex behaviors are detected as ensembles of more elementary behaviors:

1. Static postures: They describe specific relations between the tracked joints and consequently the configuration of a part of the skeleton. An example is the "catapult stance" where both hands are positioned behind the head with the elbows facing outward.

2. Gestures with linear movement: They describe a linear movement with a particular direction and speed of one or more joints. An example includes the lean forward posture shift where the position of the joint corresponding to the torso is moving forward.

3. Combined postures and gestures: They consist of a set of static postures and/or gestures with linear movement that are combined according to specific time constraints. An example includes the start of an opening up procedure [Pease 1988] which consists of a posture with arms and legs crossed followed by a posture with legs uncrossed and feet placed in a neutral position.

To this end, we defined a set of behavioral primitives, which includes the absolute and relative positions of the hands/feet, the distance between hands/feet, the absolute and relative positions of the elbows/feet, touches of the body with the hands and head movements. Based on the behavioral primitives, we created a repertoire of:

1. Typical hand positions: hands together at a particular height of the body, neck touch with left/right hand, head touch with left/right hand, head touch with both hands

2. Characteristic Leg Configurations: standing or sitting with legs apart, closed or crossed

3. Characteristic Arm Configurations: standing or sitting with spread arms at a particular height of the body, arms close to body at a particular height, arms stemmed in hips and arms behind the head with elbows facing outward

4. Common Postures for the Upper Trunk: leaning forward and leaning backward

5. Common Head Movements: looking away, head nods, head shakes, head tilts

FUBI comes with an XML-based posture and gesture specification language which enables a declarative definition of postures and gestures to be recognized in a particular application using the FUBI framework. In our previous work [Baur et al. 2013a] we investigated the reliability of our predefined set of FUBI recognizers, achieving a mean detection rate of 88.64%

4.2. Movement Expressivity

In addition to a mechanism for the detection of postures and gestures, our system provides measurements for their quality in terms of expressivity features. Based on the work by Wallbott [Wallbott 1998] and Caridakis et al. [Caridakis et al. 2006], we decided to compute the expressivity features seen in table 4.2 as indicators of how a person is perceived.

Table I. Expressivity features as calculated by the Social Cue Recognition module.

Audio Feature	Description
Energy/Power	Represents the dynamic properties of a movement (e.g. weak versus strong). It is calculated from the first derivative of the motion vectors in all three dimensions
Fluidity	Differentiates smooth movements from jerky ones. This feature aims to capture the continuity between movements. It is calculated as the sum of the variance of both hands' motion vectors' norms
Spatial extent	Is modeled as the space used for gesturing in front of the recorded person. It is calculated as the maximum Euclidean distance of the position of the two hands
Overall activation	Represents the quantity of the movement (passive versus active). It is calculated as the sum of the motion vectors' norm of both hands
Temporal extent	Represents the duration of a gesture (short vs sustained). The duration of each gesture is computed from the starting and end points synchronized with the recording time in the SSI framework

4.3. Facial Expressions

For detecting head poses and facial expressions we make use of the capabilities of the Microsoft Kinect SDK, as well as the Intraface Face Tracker [Xiong and De la Torre 2013] and Fraunhofer's Shore [Ruf et al. 2011] (See Figure 6). We implemented threshold-based trigger mechanisms for the realization of specified events. To detect smile occurrences we use the facial expression "happy" computed by Shore. Once this facial expression exceeds a certain intensity threshold, our system reports a smile occurrence. Smiles are important social cues as they are able to convey a broad range of emotions besides happiness, such as friendliness, anxiety and others [Kraut and Johnston 1979] [Harrigan and Taing 1997].

4.4. Eye Tracking

To further investigate gaze behavior in Human-Agent interaction we implemented support for Eye-Tracking devices, in form of the Eyetribe[3] and SMI[4] stationary eye trackers and eye-tracking glasses (ETGs), to our online recognition module. The eye tracking glasses are equipped with an additional camera showing the user's point of view, which, combined with algorithms for facial feature detection (see section 4.3) enable us to detect and analyze the interlocutor's face from the user's perspective as well as determine the user's gaze type.

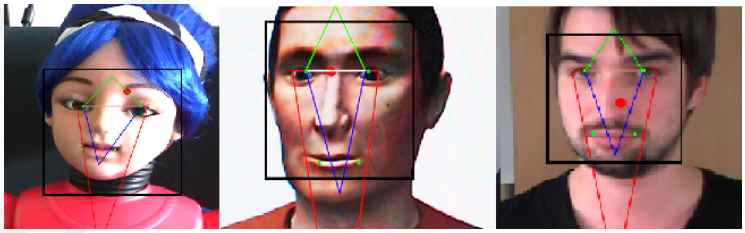

Fig. 7. Detection of Face regions using Eye Tracking glasses within our system for humanoid Robots (with human like faces), virtual characters and humans.

Figure 7 shows an example usage of this approach: Matching gaze coordinates to specific gaze areas. Pease [Pease 1988] discriminates three basic types of gazing in a social interaction: *Power gaze* (green triangle above eyes), *social gaze* (blue triangle between eyes and mouth) and *intimate gaze* (red triangle between eyes and chest).

Business/Power Gaze - When having discussions on a business level, e.g. a job interview or a negotiation, one can imagine there is a triangle on the other person's forehead. By keeping the gaze

[3] **http://theeyetribe.com/**
[4] **http://www.smivision.com/**

directed at this area, a serious atmosphere is created and the other person senses that one means business. If the gaze does not drop below the level of the other person's eyes, one is able to maintain control of the interaction.

Social Gaze - When the gaze drops below the other person's eye level, a social atmosphere develops. Experiments into gazing reveal that during social encounters the eyes also look in a triangular area on the other person's face, in this case between the eyes and the mouth. It is non-aggressive and shows comfort.

Intimate Gaze - The gaze is across the eyes and below the chin to other parts of the person's body. In close encounters it is the triangular area between the eyes and the chest and for distant gazing from the eyes to the crotch. Men and women use this gaze to show interest in each other and those who are interested will return the gaze.

The analysis of the type of gaze behavior when creating mutual gaze with others provides further insights. This information can for example be used in training scenarios where prolonged use of the business gaze is of high importance, for example in job interviews.

4.5. Voice features

Another aspect of nonverbal behavior, is paralanguage. Even without knowledge about what a person is actually saying, the way somebody raises their voice, for example, facilitates a high amount of implicit information. Table II shows the range of audio cues our recognition system is able to compute in real-time. To compute the audio features intensity, loudness, pitch and energy we use OpenSMILE [Eyben et al. 2013]. Other features are calculated using PRAAT [Boersma and Weenink 2005; de Jong and Wempe 2009] algorithms. Both systems have been integrated into the SSI Framework to process all features in real-time. Relevant parts (e.g. only when the user is speaking) are segmented by voice activity detection to calculate features on utterances of speech. Further we integrated the Microsoft Speech Platform to our system to allow key word detection for simple answers and backchanneling, as well as agent and scene control. For example a user can tell the agent when to proceed with the interaction after finishing a specific given task or give simple answers like yes or no to alternate the interaction flow (See Section 5).

Table II. Audio features recognized by the Social Cue Recognition module.

Audio Feature	**Description**
Voice Activity	Presence or absence of voice
Intensity, loudness, energy	Energy-based features of the audio signal
Pitch value	The pitch (F0) of the audio signal
Jitter, shimmer, voice breaks, harmonicity	Quality-of-voice features [Boersma and Weenink 2005] computed from pitch
Length of speech segments	The duration in seconds of the user's speech segments determined by voice activity detection
Speech rate	Rate of user's speech [Boersma and Weenink 2005]

5. INTERACTION MANAGEMENT

The application's interaction management as well as the agent's multimodal dialog behavior are modeled with the *VisualSceneMaker* authoring software [Gebhard et al. 2012; Mehlmann and André 2012]. VisualSceneMaker has been specifically designed for the rapid development and prototyping of interactive performances with artificially intelligent agents, such as conversational embodied characters [Mehlmann et al. 2011] and social or collaborative robots [Mehlmann et al. 2014].

We decided to use VisualSceneMaker because it offers two significant advantages over other approaches for interaction and dialog management that we have taken into consideration.

First, VisualSceneMaker's modeling concepts intuitively support recording the agents' behavior as well as meta information about the dialog state and discourse context at any point of the interaction. Second, VisualSceneMaker's software architecture easily supports the integration of the other main software components of our system to which it can propagate information regarding the agent's behavior in real time. This enables the subsequent analysis and visualization of the users' behavior in alignment with the agents' behavior recognized by the social cue recognition module.

In the following, we explain how VSM has been integrated into the overall software architecture and how data is exchanged between VisualSceneMaker and the other main software components. Subsequently we explain how the modeling concepts of VisualSceneMaker have been taken advantage of for automatically recording the virtual characters' behavior and discourse context during an interactive performance for the visualization and high-level recognition with NovA.

5.1. Software Architecture Overview

Figure 8 depicts the essential parts of our application's software architecture. It shows the different software components and knowledge bases of VisualSceneMaker and their integration with the other software components described in this article.

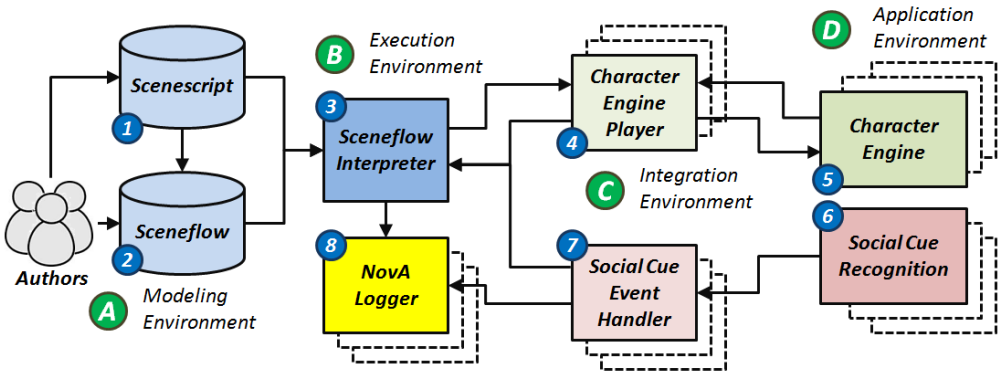

Fig. 8. The software architecture of VisualSceneMaker integrating the main components of our application.

The *Modeling Environment* (Fig. 8 Ⓐ) consists of a visual editor tool that is used by the authors to visually and textually specify the interaction and dialog behavior of the agents. Dialog content is described with a *Scenescript* (Fig. 8 ①) that resembles a movie script consisting of the agents' utterances and stage directions for controlling gestures, postures, gaze and facial expressions. The Scenescript syntax is basically a generic specification format for multi-modal behavior and can be created both, manually by an author and automatically by external generation modules. While manually written Scenescripts are commonly used for the prototyping of interactive performances in an early stage of development, they can easily be replaced by an artificially intelligent dialogue planning component that allows even more emergence and variability while often requiring less effort for the specification of behavior. The possibility to use placeholders in scenes may be exploited to create scenes in a hybrid way between fixed authored scene content and variable content, such as retrieved information from user interactions, sensor input or generated content from knowledge bases. The interaction logic and dialog structure are modeled with a *Sceneflow* (Fig. 8 ②) which is a hierarchical and concurrent state chart variant specifying the logic and temporal order according to which scenes from the Scenescript are played back and commands of the underlying programming language are executed.

The *Execution Environment* (Fig. 8 Ⓑ) is responsible for executing the interaction model that has been specified before (Fig. 8 Ⓐ). For this purpose, it relies on a run-time interpreter software (Fig. 8 ③) that is able to directly call functions of the underlying implementation language to

exchange information with the other software components. The most important example of such an external functionality in our application is the *NovA Logger* component (Fig. 8 (8)). This component provides a variety of logging functions that are explicitly called from within the interaction model by the authors or automatically called by the Sceneflow interpreter at specific points of the execution to let the NovA system know about the context, content and progress of the current interactive performance. This information is then used for labeling context within the NovA user interface. Furthermore, the logger is also used to record events and activities for online processing with the Social Cue Recognition module.

The *Integration Environment* (Fig. 8 (C)) includes the software components that are integrating VisualSceneMaker with the software components in the *Application Environment* (Fig. 8 (D)). The interface to the Social Cue Recognition module (Fig. 8 (6)) is used with an event handler (Fig. 8 (7)) that is responsible for receiving and processing user input events from the recognition modules in the application environment and to translate them for an adequate reaction in the sceneflow interpreter. The communication with the character engine (Fig. 8 (5)) is realized with a *Sceneplayer* (Fig. 8 (4)) which is responsible for translating the generic behavior specifications of the Scenescript into the respective action set of the character engine. It communicates with the character engine via a notification protocol that first requests certain activities of the virtual characters, such as text-to-speech synthesis or animation scheduling and afterwards awaits their execution by the character engine.

5.2. Recording Agent Behavior

Modeling concepts of VisualSceneMaker intuitively support the recording of information about the agents' behavior, the various kinds of dialog context knowledge and the progress of the interaction. For example, each parallel process represents a certain responsibility in the model, such as a cognitive or autonomous behavioral process and may be represented as an individual track in the NovA user interface. Furthermore, the hierarchical decomposition of the model implicitly creates a hierarchy of dialog contexts and the sequential composition usually characterizes the sequence of consecutive dialog phases within a certain context. Consequently, the parallel, hierarchical and sequential structuring of the model usually already implies or prescribes the way how dialog context and meta information has to be supplied to NovA. Thus the structure of the interaction model resembles the annotation scheme and vice-versa. Consequently, based on the modeling concepts of VisualSceneMaker, different mechanisms may be used to provide the NovA system with information about the agents' behavior and context knowledge. These mechanisms work analogous to the mechanisms used for the generation of behavior and the structuring of the interaction model. We now illustrate these different mechanisms in more detail by explaining different parts of a simplified version of the dialogue structure used in our demonstrator application which is shown in Figure 9.

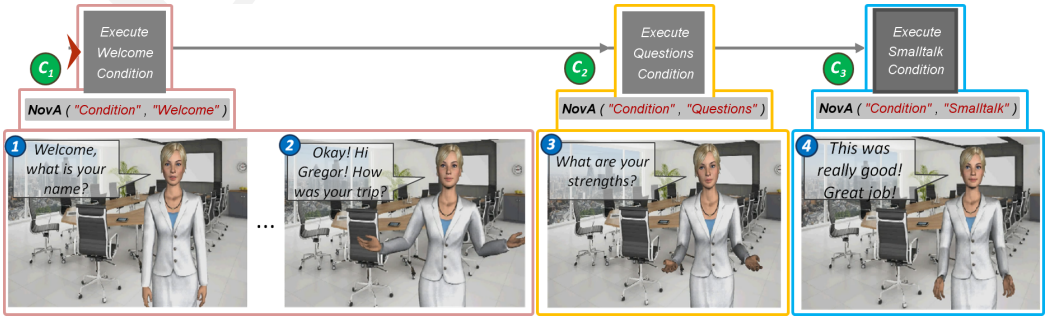

Fig. 9. A simplified version of the interaction design used in our demonstrator application illustrating different ways to supply the NovA software with information about the agents' behavior and the various kinds of dialogue context knowledge.

First, the author of an interactive performance may use a variety of predefined function calls to the NovA Logger from within the Sceneflow and the Scenescript. These commands may be used to log the agents' verbal and nonverbal behavior, cognitive or emotional states and arbitrary meta information about the state of the dialog or expectations for user reactions. These calls to the NovA Logger produce events that are recorded in an event structure for later annotation, as well as for the automatic analysis of the user's reactions to certain actions and his behavior at certain points in time, or during individual dialog phases. In addition, these events are propagated to the same event board as the behavioral cue events created by the social cue recognition module. Such events are further processed for a variety of purposes. For example the temporal alignment of the agents' nonverbal behavior with the user's behavior is automatically analyzed and evaluated in the social attitude detection module described in Section 6. In the virtual job interview application of the TARDIS project described in Section 3, we use the logger's functionality to record the agent's nonverbal behavior as well as the current condition of the dialog during a job interview. The performance of the job candidate during different dialogue phases is automatically analyzed and subsequently presented to and discussed with the user. Figure 9 shows a simplified extract of the interaction model in this application. The dialogue is divided into four consecutive conditions (Fig. 9 C_1-C_4). An event is generated via a call to the logger function at the start and end of each of the conditions. In the `Welcome` phase (Fig. 9 C_1) the agent first infers some general knowledge about the job candidate (Fig. 9 ①②), such as the user's name and age, and afterwards welcomes the candidate. The `Welcome` phase ends and the execution proceeds with the `Questions` phase (Fig. 9 C_2), in which the agent asks a range of questions regarding the capabilities of the job candidate (Fig. 9 ③). Thereby, each single question is again recorded via a call to the NovA Logger to be able to analyze the user's physiological and nonverbal reactions to each individual question subsequent to the interview. The dialogue ends with a `Smalltalk` phase (Fig. 9 C_3), in which the agent commends the candidate's performance, gives some feedback and says goodbye (Fig. 9 ④).

Fig. 10. An author may use logging directives within a Scene to provide NovA with arbitrary meta information.

Beside using function calls to the NovA logger in the Sceneflow, an author may also specify arbitrary logging directives in a Scene. Figure 10 shows an example in which an author uses log directions directly in the textual description of a Scene to explicitly create events whenever the agent provokes a certain emotional reaction in the user. These emotion eliciting events and the user's prompt emotional reactions are later automatically analyzed with NovA to find out if they have achieved their purpose by upsetting the user. In the example on the left side in Fig. 10, three kinds of information will be logged a) What the agent is saying (with start and end time) b) The occurrence of a nonverbal behavior (smile) and c) meta information that this sentence is a compliment. Tags are defined in a dictionary and depend on the character rendering engine.

5.3. Detecting bi-directional cues

Beside the explicit recording of information using specific actions in a Scene or function calls to the NovA logger in the Sceneflow, there is a further implicit mechanism used to provide information to the NovA system. While a Scene is scheduled on the character engine, the scheduling algorithm notifies the start and end event of each gesture animation, facial expression or synthesis of a spoken utterance to the Sceneplayer which then automatically forwards these events to the NovA system. When a user is performing a specific social cue, a possible question for the analysis of this behavior is if this social cue is a response to a stimulus. For example does the user nod his head because she or he is trying to encourage the interlocutor in what they are currently talking about? Rich et al. [Rich et al. 2010] investigated four bi-directional cues, namely backchanneling, mutual and directed gaze and adjacency pairs. We adapted these for our analysis as a starting point, but we nevertheless want to mention that in inter-human interaction there are many more of such bidirectional cues that have to be considered. Examples are mirroring, declarative pointings or turn-talking signals. Combining the agent events with events received from the Social Cue Recognition module (Section 4) enhances the automated analysis and annotation process.

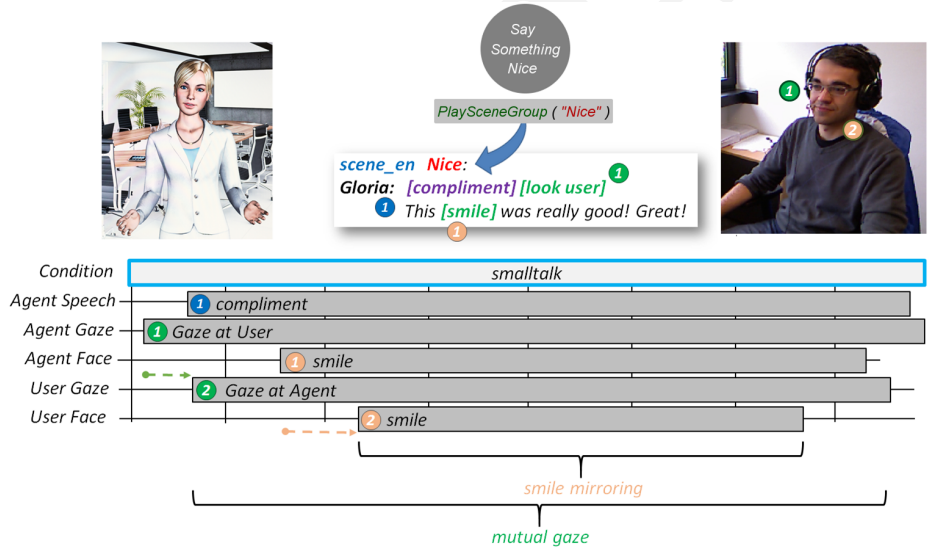

Fig. 11. The Sceneplayer implicitly records the beginning and the end of utterances and nonverbal behaviors. This allows setting them in relation to the user's social cues. Considering temporal alignment allows the creation of bi-directional cues, such as mirroring or mutual gaze.

The example in Figure 11 illustrates how the NovA system combines behavior of the user, and the agent, as well as the context to detect bi-directional cues. Following the example in Figures 9 $\textcircled{C_4}$ and 10 VisualSceneMaker provides the information that the actual discourse context is the condition *smalltalk*. Further, according to the agent's scene script the system is provided with the meta information that the agent's actual utterance is a compliment. The agent is looking at the user's face directly before starting to speak. If the user returns the gaze within a certain timespan (1.6 seconds, according to [Rich et al. 2010]) by looking at the agent's face the system will label a successful mutual gaze. Similarly, the system recognizes the mirroring of a smile. For example, as the agent shows a smile and the user smiles back within a predefined timespan, as visualized in Figure 11, this bi-directional behavior (mirroring) will also be labeled. A high amount of bi-directional cues increases the dynamics of a conversation and is considered to reflect a high engagement, which will be discussed in the next section. Additionally, as meta information about the topic and the utterance

are present, the automated analysis of such social cues is enhanced by including this information. In this case the system would suggest this is a friendly and polite smile, as this is a natural and expected reaction to a compliment. Performing or not performing such expected behaviors also impacts the system's calculation of social attitudes as described in the next section.

6. INFERRING HIGHER-LEVEL SOCIAL ATTITUDES

To analyze the user's social attitude we combine social cues, bi-directional behavior and context information. To determine such higher-level concepts we suggest the use of a probabilistic model, more specifically Dynamic Bayesian Networks (DBN) [Murphy 2002] which are updated online using the social cues detected by the Social Cue Recognition module and the agent cues received from the Interaction modeling module as evidences. We illustrate the functionality using *Engagement* as an example, which is a good indicator of critical situations in social interactions. This approach might also be used to compute other social attitudes (e.g. dominance, self efficacy, rapport) in an analogous way.

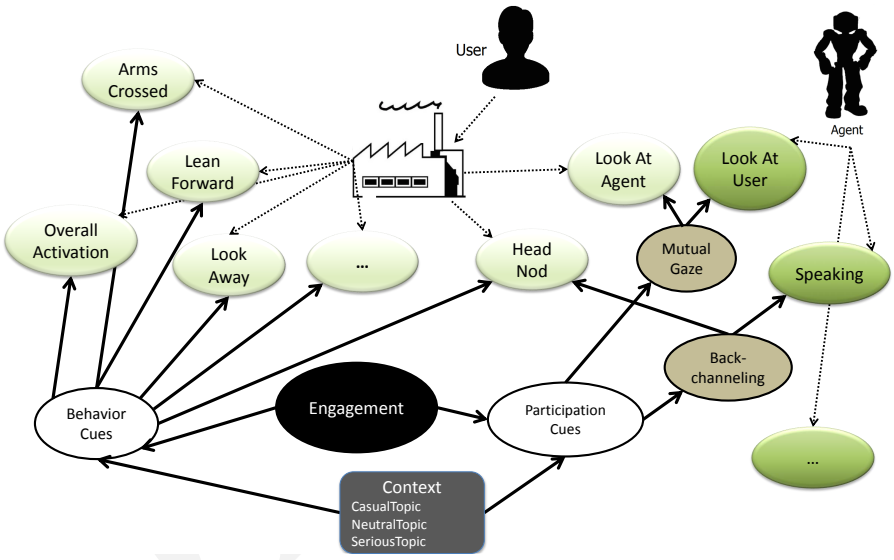

Fig. 12. A simplified Illustration of a Bayesian network to determine Engagement while considering bi-diretional and context information.

According to Sidner and colleagues [Sidner et al. 2004], *Engagement* "is the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake." In a day-to-day sense, engagement is related to showing attention and interest in a conversation. For example, Pease [Pease 1988] provides a number of examples to demonstrate how engagement is portrayed by body postures and movements. Typically, engagement is shown by an orientation of the body and the face towards the interlocutor while an orientation of the body and face away from the interlocutor may be interpreted as a sign of disengagement. Further, creating a high amount of mutual and directed gaze, back-channeling, and adjacency pairs indicates a high amount of engagement [Rich et al. 2010]. There are also specific hand gestures that reveal whether a listener is engaged or not. For example, engaged people typically touch their chin without bracing the head. A slight variation of this gesture would, however, reveal the opposite. Bored people may also touch their chin. But in this case, the hand typically fully braces the head [Pease 1988]. As seen in these examples, nonverbal signals can not be straightforwardly interpreted in every case. Considering the interpretation of nonverbal behavior is often hard to read even for

humans, an automatic recognition of social attitudes from low level social signals is a particularly challenging task. An important aspect of how humans interpret nonverbal signals correctly is the context of the situation. As consequence, to automate the analysis and interpretation of low and high level features context-awareness is of vast importance. Based on the inputs of the interaction management component we make use of the discourse and interaction context to improve the interpretation of nonverbal behavior. A typical network designed to recognize *Engagement* consists of several unconditional, observed nodes that describe the evidences and probabilities monitored by the social cue recognition module and the interaction modeling component and are constantly updated by the system in real-time. These evidence nodes feed into conditional nodes that estimate a higher level statement based on the recognized cues. Observed, as well as conditional nodes lead to the final child node, which models a social attitude. The Bayesian networks used in our system can be modeled with existing tools, such as GeNIe[5]. Figure 12 shows the diagram of a simplified Bayesian network, meant to recognize *Engagement*.

Social cues, as well as the agent's behavior cues and the perceived context information are represented with unconditional parent nodes (Arms Crossed, Overall Activation, Lean Forward, Look Away, Head Nod, .., Agent Speaking, Agent looking at user, .., Context). Bi-directional cues are represented by hybrid nodes that are influenced by agent and user behavior and the alignment of both, as described in section 5.3. For easier illustration, only Mutual Gaze and Backchanneling are presented in the figure. Unconditional and hybrid nodes are updated by the social cue recognition component and the interaction modeling tool in real-time. These parent nodes influence interconnected conditional nodes that represent sub-concepts (*Participation Cues, Behavior Cues*) which finally lead into the outcome *Engagement* node. The network is recalculated every update cycle, e.g. every 500ms. The output of the network delivers a value between 0.0 and 1.0, which represents a continuous assertion about the presence of the modeled social attitude. A BN can also have multiple final states or alternatively, multiple networks can be calculated in parallel.

An important concept behind inferring higher-level social attitudes is to make it easier to point out critical incidents during the interaction. For example, this could help a social coach in an post-interaction analysis find critical situations faster. The outputs of the social attitude detection are annotated alongside other behavioral and interaction cues in NovA's user interface. Illustrating examples will be given in later chapters (see: Section 7.4). As a prospect for future research another imaginable use case for the high level outputs of this component is providing an agent with the possibility to react and adapt to the user's behavior more naturally in real-time.

7. NOVA'S USER INTERFACE

The graphical user interface of NovA [Baur et al. 2013b] has been developed following the requirements of tools for annotating human social interactions. It offers automated annotations on multiple tracks based on a user-defined coding scheme that has been, in our case, adapted to the situation of a human-agent dialogue. It visualizes specific behavior patterns of the interaction and the relation between the user and the agent's behaviors. Figure 13 gives an overview on the graphical user interface. The next subsections will have a more detailed look at the specific modules of the UI.

7.1. Video Panel

The video panel (Figure 13 A) plays back the recordings of the interaction. For the recording of the user, traces for each joint of the skeleton may be shown on demand in the video by manually selecting them. Their movement is visualized in three dimensions whereby the z-value is represented by the corresponding alpha value in the trace. Besides the Kinect video, and additional high-quality camera recordings of the user, NovA is also capable to play back the video and audio recording of the agent, which allows a direct reference between both interlocutors.

Further it supports the playback of Point-of-view videos, e.g. from Eye Tracking glasses (see Figure 14). Users also have the option to create heat maps based on eye tracking data to analyze gaze

[5] **http://genie.sis.pitt.edu/**

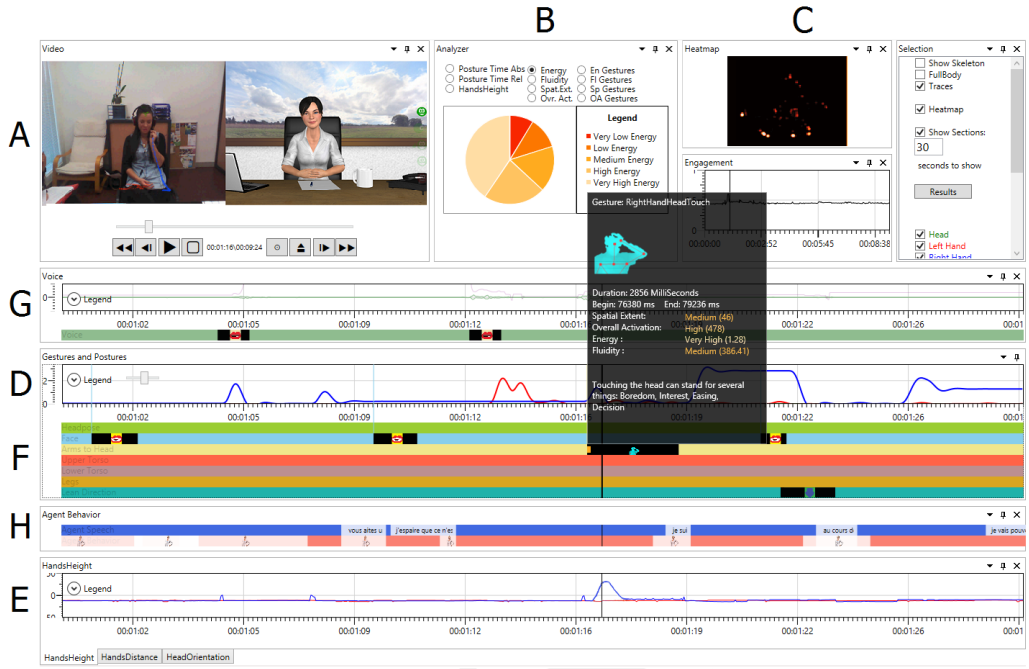

Fig. 13. NovA's graphical user interface. In this instance data for a user and a virtual agent has been loaded. It shows both recordings (A), pie charts for expressiveness features (B), movement heat-maps (C), the waveform graph with voice activity detection events (G), the timeline graph showing automatically created annotations of behavioral cues (F), the hands height graph (E) and the agent behavior graph showing the agent's speech and animation outputs (H).

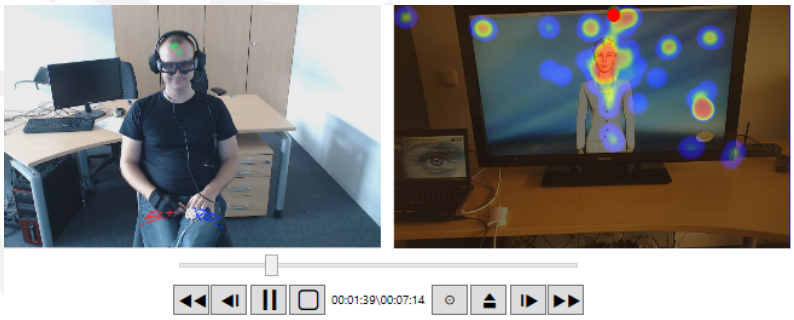

Fig. 14. An instance of NovA's video panel showing two videos: a user wearing eye tracking glasses (left video) when interacting with a virtual agent. Additionally his point-of-view video (right video) is shown with a heat-map overlay based on eye tracking data. The buttons present from left to right: last section, last annotation, play/pause, stop, record, open, next annotation, next section.

behavior. By further adding an additional video, for example, a recording from another perspective, NovA is capable of synchronously replaying up to four videos at a time. When selecting a particular point in any of the graphs, the video position moves to the corresponding point in time so videos and graphs are always synchronized. It is also possible to directly jump to the next/last annotation, as well as next/last section.

7.2. Panels with Descriptive Statistics

The User Interface of NovA is capable of generating a variety of descriptive statistics. Examples include expressiveness features, absolute and relative duration of gestures (e.g. the duration of open in comparison to closed gestures), relative height of hands and audio features. Statistics are visualized for example by pie chart diagrams that represent proportions between the single classes, as seen in Figure 13 B and Figure 15 A. These charts refer, depending on the user's specification, either to a chosen timespan or dialogue condition, or alternatively to the complete recording.

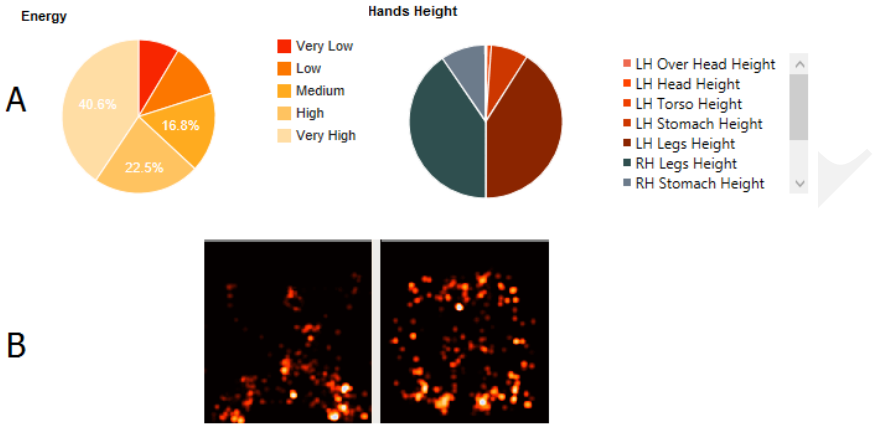

Fig. 15. Examples for pie chart diagrams for energy and the height of the user's hands that visualize user behavior for a selected section. Further the figure shows movement heat maps for two alternative scenes, showing differences in gesticulation. The left user kept his hands basically calm and spread apart while the right one was fidgeting a lot with his hands.

In addition, NovA visualizes motion data, aggregated over time with heat maps (Figure 14 right and Figure 15 B). For example, the light points in Figure 15 B show the most frequent positions of selected joints (head, left/right hand) within a certain time interval.

7.3. Timeline Panel

NovA supports both continuous and event-based annotations. Consequently, NovA's timeline panel includes two kinds of tracks:

1. Tracks that correspond to behavioral characteristics collected frame by frame, such as motion energy (Figure 13 D) or the height of the hands (Figure 13 E) and
2. Tracks that correspond to events, such as the occurrence of particular postures and gestures (Figure 13 F), Speech (Figure 13 G) or Agent Behavior (Figure 13 H) .

7.3.1. Continuous Annotation. In the case of continuous annotations, a value referring to a particular feature of a behavior is computed at each point in time. E.g. the expressiveness parameters introduced in section 4.2 are computed frame-by-frame. Figure 13 D displays the energy for selected joints of the user. The single joints are distinguished by different colors. The user's movements can be easily recognized by the peaks in the curves whereby higher peaks indicate more energetic movements.

Furthermore, the height and distance of particular joints, such as the hand joints, are calculated frame-by-frame. For instance, we make use of a height-of-hands graph (see Figure 16) that visualizes the height of both hands in relation to the torso. Values above zero refer to movements above the torso, such as head touches while values below zero indicate that the person is keeping his hands down. The distance-of-hands graph, which visualizes the distance between the two hands of a person in the x and z dimension, looks similar. Peaks in this graph represent a large distance between

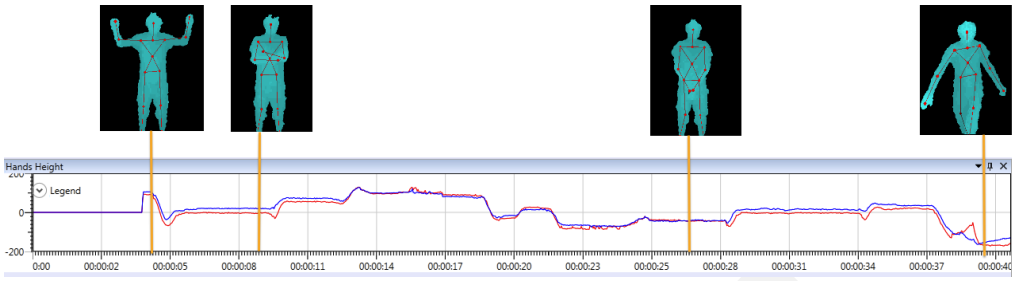

Fig. 16. Continuous annotation of the hands' height in relation to the torso and to each other.

the two hands while points close to zero on the y axis indicate that the two hands are close to each other. The distance-of-hands graph may also be regarded as a measurement for the spatial extent of hand gestures. The height-of-hands and the distance-of-hands graphs also allow us to explore how the two hands of a person are synchronized with each other.

Another example of a frame-by-frame annotation is the representation of the waveform corresponding to the audio signal (Figure 13 G). Phases with high peaks indicate high intensity (e.g. a loud voice) while phases without peaks represent silent phases. Currently, the audio signal is used to determine whether the analyzed person is speaking or listening which is also represented with event-based annotations.

7.3.2. Event-Based Annotation. In the case of event-based annotation, the annotations are triggered by specific events corresponding to particular user or agent behaviors or bi-directional behaviors. According to [McKeown and Sneddon 2014], the *standard method* for annotating data is to have fixed and known time segments associated with a descriptive label. In our system agent events deliver start and end boundaries automatically, as they are triggered by the interaction modeling component. However, to recognize and label the user's behaviors, our event recognizers use strictly defined rules to detect event boundaries: The FUBI framework for gesture recognition tries to to map predefined concatenations of poses, which have to be performed in a certain chronological sequence. These strict definitions do not leave space for interpretation, as requirements for the gesture event are either met or missed. Same goes for eye tracking events, as the inspection of a region of interest within the face is a boolean decision. For facial expression events, we rely on a pre-trained module with thresholds that have shown to perform well in practical applications.

Figure 13 F shows the tracks for the selected user behaviors whereby the single tracks refer to a particular behavior type. To be able to analyze relationships between the behaviors of the user and the agent, two separate sets of tracks have been included. The user set contains annotations for the following behavior types by default: hand-to-head gestures, upper-body gestures, head poses, leg postures and body shifts. The set for the agent contains speech annotations (e.g. spoken text), and behavior annotations, (e.g. played animations). Annotation schemes are fully customizable to the user's needs. Therefore a user of the system can choose the cues that belong to a certain category of behavior types and also add or remove cues. Self-defined FUBI gesture recognizers will automatically be added to the annotation scheme by default. Each annotation includes additional information as shown in Figure 17 that may be displayed on demand. Depending whether the category of the annotation is a gesture/posture, a voice activity or an action by the agent, information windows vary. For gestures, calculated expressiveness parameters, as well as descriptions of possible interpretations are shown (Fig. 17 left). Voice activity annotations include calculated audio features (with differences to mean values) for the specific utterance (Fig. 17 center). Agent behavior annotations contain either the full content of a spoken utterance, or the displayed animation (Fig. 17 right). In the case of recognition errors or the necessity of adding additional annotations that are not covered by the automated coding procedure, NovA offers the possibility to manually add or delete event-based annotations or to edit their temporal position and duration with intuitive mouse

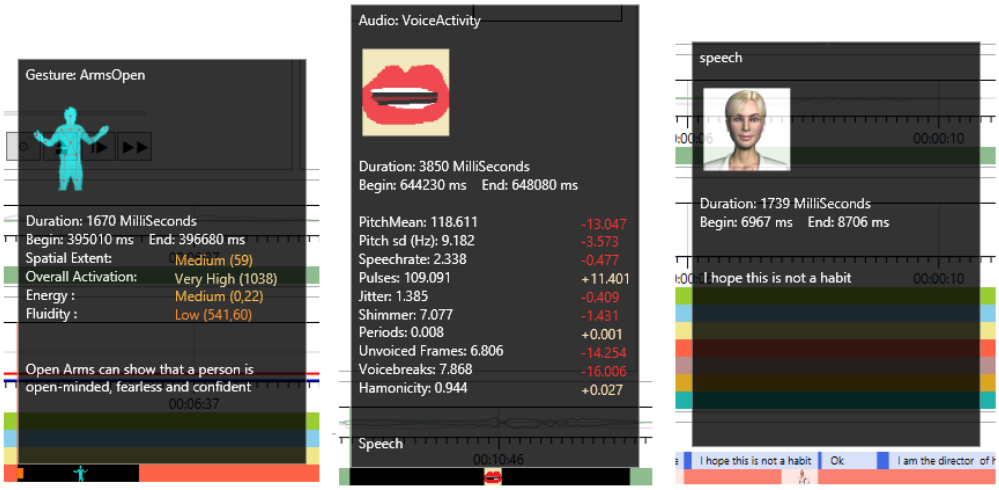

Fig. 17. Detailed information for each annotation depends on the annotation being a gesture/posture (left), a spoken utterance (middle) or agent behavior (right).

gestures. Further, we created export functions for event-based annotations to other tools like ELAN, or to SSI machine learning annotations for training models based on labeled data.

7.4. Illustrating Examples

In the following, we present two examples to illustrate behavior analysis in NovA. Figure 18 illustrates how the system determines the level of engagement of the user, showing the interface with detected annotation labels and continuous graphs.

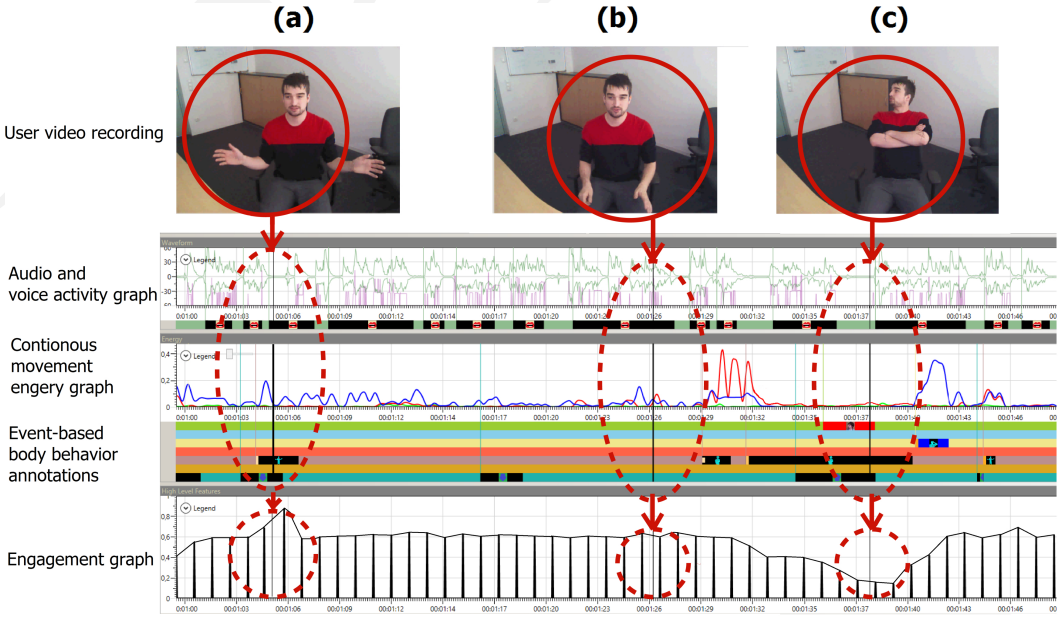

Fig. 18. Comparison of detected cues for high (a), medium (b) and low engagement (c).

On the left (a), the participant has an open body posture, while looking towards the agent and orientating his body in the same direction. (b) shows the user in a neutral position. The right part of the figure (c) demonstrates the outcome when the participant uses body language specific to low engagement (see Section 6), such as leaning back, looking away and crossing the arms. In this instance, bar charts are representing the outcome of the engagement recognition for each calculation, which is performed every second.

The second example shows a typical interaction with a virtual agent. Referring to our description in Section 5 the agent behavior time-line (Figure 19 a) shows annotations of the agent's speech and animation events, while the annotations of the user's behavior are presented alongside on the lower time-line (Figure 19 b). The segments below the engagement graph (Figure 19 c) represent the different conditions/phases of the interaction, received from the interaction modeling component. As seen in Figure 19 the interaction starts with a *Welcome* phase. While the participant begins with closed body language, she partly opens up during the *Welcome* phase as she continuous talking with the agent which raises the level of engagement. When it comes to the phases *Elaboration* and *Questions & Answers*, she turns back to a half closed body posture, by crossing one arm on a side position. Finally the agent tells her she did well and she turns to talk with the experimenters in the room, smiling and showing an open posture which increases the detected engagement.

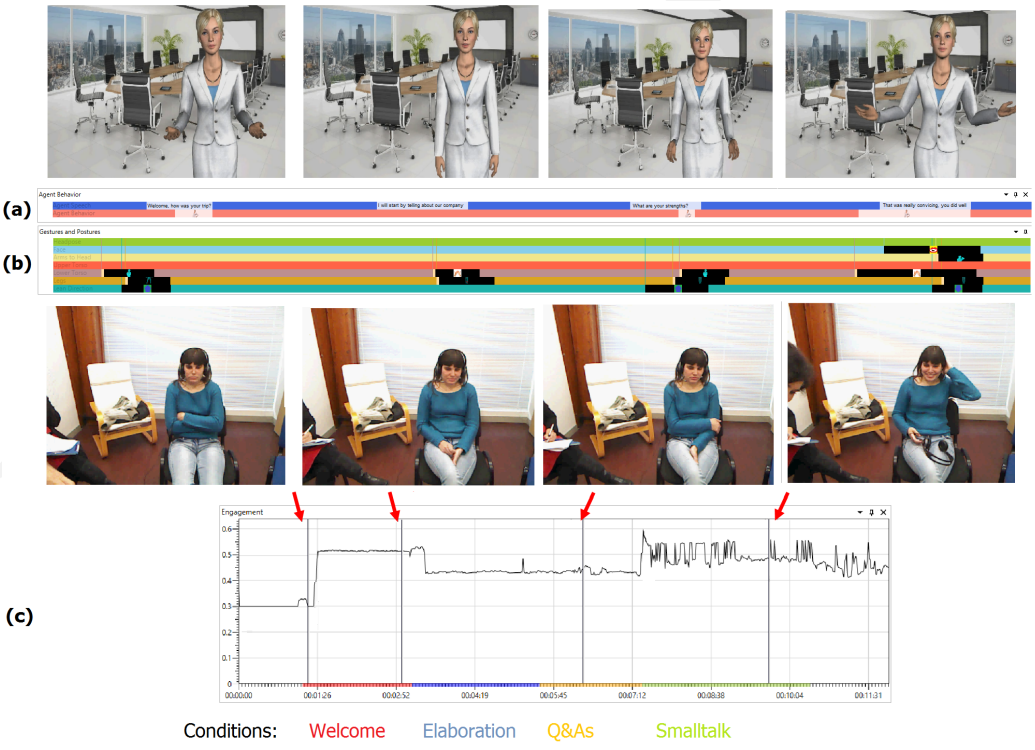

Fig. 19. Analysis of a bi-directional interaction with a Virtual Agent. The Figure illustrates the Tracks that show the logged behavior from the agent (a), the user (b) and the engagement graph (c) with automatically created annotations of phases (Welcome, Elaboration, Questions & Answers and Smalltalk).

The data from this example was picked from real-world recordings that were taken within the TARDIS project. In the next section we will describe our experiences with the system in field study.

8. USE CASE STUDY AND DISCUSSION

We evaluated the performance of the system's single components as part of our previous work in various studies and conditions [Gebhard et al. 2014] [Baur et al. 2013a] [Damian et al. 2013] [Porayska-Pomsta et al. 2014]. To show the usefulness of the system for coaching, we decided to evaluate the approach as a whole and most importantly in the field with pupils and practitioners, instead of a controlled lab study. The highlight of our approach is the integration of social signal processing techniques with interaction modeling in combination with an easy to use visualization interface. Consequently, we found it important to evaluate the visualization and analysis tools not in isolation, but as parts of an integrated interactive system.

To evaluate the impact of the training and coaching system, we conducted a study over the course of three days at a secondary modern school (Mittelschule Stadtbergen, Augsburg, Germany) using the TARDIS game application presented in Section 3. Within the TARDIS game, NovA serves to analyze the learner's social cues when interacting with a virtual recruiter during a virtual job interview training with respect to dialogue dynamics and context information.

The participants of the study were 20 pupils (10 male and 10 female) in their final or pre-final graduation year, aged between 13 and 16 (mean = 14.37; SD = 0.94) that have been categorized by their teachers as being at risk of exclusion. Most of them already started looking for employment. Two professional career councilors from the Career Service of Augsburg University volunteered to support us in the study. The main objective of the system was to evaluate the impact of the TARDIS game and the NovA coaching interface on the pupils. Furthermore, we were also interested in seeing the reception such a system receives from both students and teachers to get a first impression of the feasibility of a long-term deployment of technology-enhanced training systems in schools.

8.1. Procedure

The user study was conducted over the course of three days. An overview of the procedure can be seen in Table III. For each session, pupils were picked up from their classroom.

Table III. Procedure of user study over three days.

	experimental group (EG)	control group (CG)
day 1	mock job interview	mock job interview
day 2	interaction with training system	training with printed job interview guide
day 3	mock job interview	mock job interview

On the first day, each student participated in a mock job interview led by one of the professional career trainers. Career trainers were instructed to be as objective as possible and to focus on the nonverbal behavior of the participants. Two interviews were carried out in parallel in separate rooms whilst each lasted for approximately 7 minutes. After each mock interview, both career trainers and pupils filled in questionnaires *A* and *B* respectively (presented below). The purpose of these first interviews was to establish a baseline regarding the job interview performance of the pupils before their interaction with the system.

- *Questionnaire A:* On a seven-point Likert scale career trainers rated the pupil's 1) overall performance, 2) recommendation for the job based on their behavior, 3) appropriate usage of smiles, 4) appropriate usage of eye contact, 5) appropriate usage of gestures, as well as the pupil's 6) nervousness 7) interest and 8) focus.
- *Questionnaire B:* Pupils rated on a seven-point Likert scale whether they thought they 1) performed well in the interview, 2) were nervous, 3) used a lot of filler words such as "er" or "uhm", 4) were not focused, 5) were aware of their non-verbal behavior and 6) performed appropriate non-verbal behavior.

On the second day, pupils were randomly assigned to either the control group (CG) or the experimental group (EG) ($N(CG) = 9$, 5 female, 4 male; $N(EG) = 10$, 4 female, 6 male). The data of

Fig. 20. Mock interviews have been performed on day 1 and 3 with professional career councilors.

one participant in the CG had to be removed due to extraordinary circumstances on the first day of the study resulting in nervous and unfocused behavior (giving a friend company to the hospital after a minor accident before her first session).

The EG interacted with the TARDIS Game and the NovA coaching interface. The system was installed in two typical class rooms which allowed parallel sessions. The participants were seated at a school desk on which a 22" monitor was positioned. A Microsoft Kinect was placed behind the monitor at a height of approximately 1.5 m of the ground and a distance of 1.5 m from the participant. During the interaction with the training game (see Fig. 21, left), the user was wearing a SHURE WH20 microphone which was paired with a TASCAM US322 audio interface. Recordings were performed on desktop computers with Intel Core i7-3930k processors. Each training lasted for about 15 minutes, split between game interaction and debriefing. During the session, their nonverbal behavior was recorded and analyzed by the system. A debriefing phase followed each interaction with the game. In this phase, a researcher assisted the pupils in reviewing the interaction using the NovA coaching interface (see Fig. 21, right). However, the researcher was only allowed to repeat the information already provided by the system and provide no further interpretation. This was done to avoid having the expertise of the researchers impact the study.

Fig. 21. Setup of the system with a participant acting with the TARDIS Game, and during the debriefing session with the NovA user interface

Pupils of the CG were reading a printed job interview guide[6] for the same amount of time the EG interacted with the system. The written guide was published by a renowned German youth advisory institution with which the school regularly cooperates for their employment preparation classes.

[6] **https://www.aok-on.de/bayern/berufseinsteiger/beruf-zukunft/koerpersprache-im-vorstellungsgespraech/**

On the third day, a second set of mock job interviews with the professional career trainers was conducted with each participant. Pupils of both groups (EG and CG) were brought to the career trainers in random order, who were unaware to which condition the pupils were assigned during the second day. After each mock interview, career trainers and pupils filled in the same questionnaire they filled in during day 1 (*Questionnaires A* and *B* respectively). This allowed us to make a direct comparison of the participants' performance between day 1 and 3.

8.2. Results

Overall, we recorded 40 mock-interviews and 10 interactions with the TARDIS system. Analyzing the first day of our experimental setup, no significant differences were found in questionnaires A and B using the independent two-tailed t-test to compare pupils that were later assigned to either join the EG or CG. These results suggest there were no prior differences between the groups in their rating by the career counselors as well as in their self-assessment. Table IV shows the mean values of the rating on the first and on the third day. Questionnaire A was filled in by the counselors, questionnaire B by the participants. Significant differences between groups on a particular day are written in bold and marked with *. Significant differences within groups between days are written in italic and marked with ‡.

Table IV. Mean values of control group (CG) and experimental group (EG) on first and third day.

	day 1		day 3	
	CG	EG	CG	EG
Questionnaire A				
overall performance	4.44	*4.90*‡	**5.33***	***6.20***‡*
recommendation	4.55	*4.70*‡	5.33	*6.20*‡
smiles	4.44	*4.20*‡	5.33	*5.70*‡
eye contact	4.44	*4.60*‡	5.66	*5.70*‡
gestures	3.60	2.80	4.00	4.10
nervousness	4.33	*4.00*‡	3.55	*2.70*‡
interest	5.00	5.00	5.55	5.80
focus	5.00	5.10	5.55	5.90
Questionnaire B				
overall performance	4.66	4.60	5.33	5.20
nervousness	4.77	*4.20*‡	4.33	*2.20*‡
use of filler words	4.88	3.40	4.22	3.00
not focused	2.77	3.10	2.44	2.40
aware of n.v. behavior	5.00	5.00	5.77	5.40
performed n.v. behavior	5.22	4.70	5.77	5.10

Comparing the two groups again after the third day revealed interesting insights. Pupils that interacted with the training system were rated better by the career councilors on all dimensions compared to the CG (see Table IV). Please note, a lower score for nervousness is considered being better. An independent two-tailed t-test with Bonferroni-Holm error adjusted significance levels yielded statistically significant differences for the career counselor's ratings on overall performance ($p = 0.004$, $\alpha = 0.006$). A strong trend was also found for the recommendation dimension ($p = 0.012$, $\alpha = 0.007$). Figure 20 summarizes the findings for Questionnaire A. For the pupils' self-assessment, results show that members of CG rated themselves slightly better than the EG although not significant.

To evaluate the improvement of performance for each group individually, we compared the results within groups between day one and three.

Pupils of the EG were rated better on all dimensions by the career counselor on the third day compared to the first day (see Table IV). Performing paired two-tailed t-tests (again with Bonferroni-Holm error adjusted significance levels) revealed significant differences for the dimensions recommendation ($p = 0.005$, $\alpha = 0.006$), overall performance ($p = 0.006$, $\alpha = 0.007$), nervousness

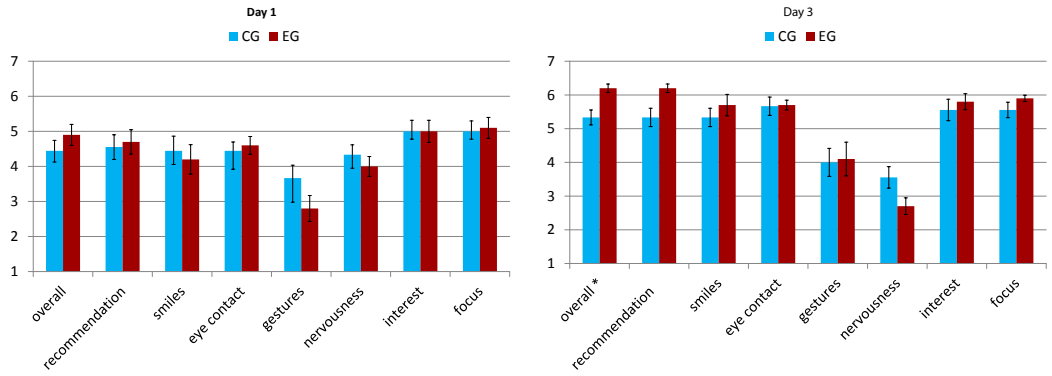

Fig. 22. Practitioners' ratings of day one (left) and day three (right) comparing CG and EG. Dimensions marked with * present significant differences between the two groups.

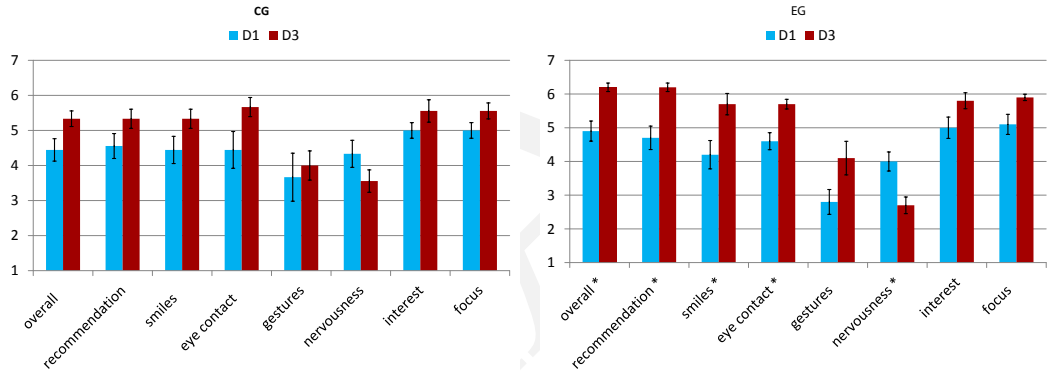

Fig. 23. Practitioners' ratings of CG (left) and EG (right) across day one and three. Dimensions marked with * present significant differences between the two days.

($p = 0.006$, $\alpha = 0.007$), eye contact ($p = 0.007$, $\alpha = 0.010$) and smiles ($p = 0.012$, $\alpha = 0.013$). Pupils' self reports revealed significant differences on the nervousness dimension only ($p = 0.001$ $\alpha = 0.008$), with participants rating themselves being less nervous on the third day compared to the first day.

Pupils of the CG were not rated significantly better in any category when comparing day three to day one (see Table IV). Also, no significant improvements were found in the self reported questionnaires.

For the experimental group, we further asked about the participant's impressions right after interacting with the system on day 2:

- *User Experience Questionnaire:* Pupils rated on a seven-point Likert scale whether they 1) found the video self reflection with the NovA Coaching interface useful, 2) had the impression they learned from the self-reflection 3) would use the training system for job preparation, 4) found the gaming cards helpful 5) had fun playing the game

The participants rated they had fun playing the game with a mean result of 5.6 which we consider a good result for a training and learning environment. The pupils further rated the helpfulness of the game cards high (mean = 6.4) which suggests that pupils are thankful for direct guidelines for specific conversational topics which we assumed by designing the application. Pupils rated they would

use the system at home to prepare for a real job interview with mean = 6.1. The post interview with the NovA coaching interface also received high ratings for helpfulness (mean = 5.7) and learning effect (mean = 5.5).

8.3. Discussion

The analysis of the questionnaire data on the first day of our experiment revealed no significant differences, which suggests that the pupils were comparable in their job interview performance. We can thus consider differences observed on the third day between the groups to be caused by the training completed on the second day. In general, both groups improved from the first day to the third. This is not surprising, considering the fact that all participants received some sort of training in job interviews over the course of three days. However only for the EG were the differences significant. Furthermore, comparing the two groups on the third day, the EG was rated significantly better in terms of overall performance than the CG. This suggests the technology-enhanced training had a greater effect on the pupils' job interview performance than the traditional method. We consider this encouraging, especially since the reading material the CG was using on the second day is issued by a respectable local youth organization and is regularly used by our cooperating school.

These findings are also reflected by the within group analysis of both groups comparing the first and the third day of our experiment. While both groups were rated better by the career counselors on the third day, only the EG showed significant differences on the dimensions in terms of overall performance, recommendation for the job, smiles, eye contact and nervousness. As the goal of any job interview training technique is to increase the user's chances for employment, we consider these results as motivating. The only statistical difference found in the pupils ratings was the self-reported nervousness of the EG. This is also interesting as it indicates the virtual job training environment might help users feel more comfortable during job interviews. The system also left a good impression on the school teachers who stated *"using the system, pupils seem to be highly motivated and able to learn how to improve their behavior"*. As a possible reason for this they mentioned the technical nature of the system, which *"transports the experience into the pupil's own world"* and the technology-enhanced debriefing phase *"makes the feedback be much more believable"*.

Pupils seemed to like interacting with the system as well. One participant even asked for permission to photograph the game cards so she would be able to study them at home. Furthermore, that particular pupil had a job interview after her session on the second day which led to a successful employment at a local fashion store chain. While we certainly cannot draw any conclusion regarding the impact of the training on this fortunate outcome, it only goes to show how pertinent such training exercises are for this particular user group. Especially during the debriefing phase with the NovA coaching interface, pupils seemed eager to explore the effects of their behavior. A participant stated: *"It was weird to see myself in the software in the beginning. At first I didn't want to watch it, but in the end it was really helpful. I tried to be less nervous and I concentrated on my posture and facial expressions in the second role-play interview"*. We believe if pupils were exposed to the system for a longer period of time, their job interview performance could be improved even more than shown by our experiment.

Even though we showed the suitability and effectiveness of our system with the use-case job interview application, we would like to point out the techniques and concepts could be generalized to other scenarios, involving not only virtual characters but also physical robots or other intelligent conversational agents. Examples are smart homes, automotive systems, elderly care, education or similar areas. While such adaptations require adjustments in the interaction scripts and Sceneplayer communication protocols to the respective system and, depending on the scenario, adaptations in the social cue recognition module, the overall concept remains the same.

We also would like to address some of the weaknesses of the system that still need to be considered. A large problem when analyzing user behavior is recognition performance. Compared to offline classification, our social cue recognizers are required to work in real-time for as many users as possible, independent from sex, age, body size or other factors. We designed all recognizers to

work user independently, but nevertheless, they are still limited by hardware restrictions to some extent. That said, subtle movements are hard to automatically detect with the current hardware available, if we do not want to put extra sensors on the users' bodies. Body-worn sensors possibly lead to high intrusion and also cause additional stress and therefore we tried to avoid them as much as possible. For the TARDIS game especially, we used game cards to give advice on how to behave in different phases of an interview. This is, on the one hand helpful for users of the system, as they learn what behavior they should show, but on the other hand it limits the amount of social cues from which we can infer social attitudes. For the selection and the automated high-level interpretation of social cues we rely on various models extracted from social theories. These models may vary in integrity and controversiality, but deliver a practical starting point for the automated analysis of human nonverbal behavior. Context Information described in this article is limited to the actual conversation. It is also imaginable to consider classical context information from the area of mobile computing, like location and time based information for the analysis of interaction in future work.

9. CONCLUSION

In this article we presented a system for the automated analysis and annotation of social interactions between humans and conversational agents. To consider the interaction context during the interpretation of social cues, the system combines a social cue recognition module with an interaction modeling component. The social cue recognition module extracts human behavioral cues in real time while the agent's interaction cues are logged by an interaction modeling component. In addition, the interaction modeling component allows us to track meta information on the dialogue, such as the topic of the conversation.

By combining this information our system infers higher-level social attitudes to point out the most critical parts of the interaction, but also to allow the agent to adapt to the user. This is used to automatically label corpora and to create automated statistical analyses.

As an use case, we presented the TARDIS project where the nonverbal behavior of young job seekers is analyzed while interacting with a virtual recruiter. In a field study we showed that a combination of real-time behavior recognition and visualization are useful in social coaching. Participants that interacted with the system improved significantly compared to the control group that used traditional preparation methods. In this project, NovA's primary use is the recognition of behaviors, as well as debriefing and post-hoc analysis of social interactions. It allows users to reflect on their behavior, and thus learn to perform better in social situations. Further, it helps to point out significant differences in behavior over different phases of an interaction.

Our system supports researchers from multiple disciplines to design interaction studies with conversational agents to get deeper insight into human behavior in modern HCI applications. Also, automated behavioral analysis and coding can help researchers, related to annotation of corpora by reducing their workload.

Even though automated annotation from social signal processing techniques provides many advantages for objective labeling of corpora, it is still a young discipline that relies on the availability of accurate sensory devices. For specific sensors, such as eye tracking glasses the necessity of wearing them on the body can lead to a high intrusion factor which may have negative influences on the naturalness of the interaction. Further, the sensor setup requires a certain expertise to assure correct tracking, which is necessary for automated recognition. Given the fact that modern sensors have difficulties detecting subtle movements, it is not possible at this stage to recognize all nuances of human behavior. To address this, our system can be extended with new devices and algorithms using plug-ins. Further, our annotation interface supports manual annotations for any cues that are not covered by automated recognition. Nevertheless, the current recognizers deliver a huge amount of information on human behavior which is an important step towards entire automated multimodal analysis and annotation of human behavior in interactions.

To maximize our contribution to the research community we made the NovA user interface available for download at *http://openssi.net/nova*. Furthermore all recognition modules have been integrated in the OpenSSI Framework (*http://openssi.net/*). Visual Scenemaker is also open source

and available for download at *https://github.com/SceneMaker/VisualSceneMaker*.

REFERENCES

Keith Anderson, Elisabeth André, Tobias Baur, Sarah Bernardini, Mathieu Chollet, Evy Chryssafidou, Ionut Damian, Cathy Ennis, Arjan Egges, Patrick Gebhard, Hazael Jones, Magalie Ochs, and others. 2013. The TARDIS framework: intelligent virtual agents for social coaching in job interviews. In *Proceedings of the Tenth International Conference on Advances in Computer Entertainment Technology (ACE-13). Enschede, the Netherlands, November 2013. Lecture Notes in Computer Science 8253*.

Ligia Maria Batrinca, Giota Stratou, Ari Shapiro, Louis-Philippe Morency, and Stefan Scherer. 2013. Cicero - Towards a Multimodal Virtual Audience Platform for Public Speaking Training. In *Intelligent Virtual Agents - 13th International Conference, IVA 2013, Edinburgh, UK, August 29-31, 2013. Proceedings (Lecture Notes in Computer Science)*, Ruth Aylett, Brigitte Krenn, Catherine Pelachaud, and Hiroshi Shimodaira (Eds.), Vol. 8108. Springer, 116–128.

Tobias Baur, Ionut Damian, Patrick Gebhard, Kaska Porayska-Pomsta, and Elisabeth André. 2013a. A Job Interview Simulation: Social Cue-Based Interaction with a Virtual Character. In *2013 IEEE/ASE International Conference on Social Computing (SocialCom)*. Washington D.C., USA, 220–227.

Tobias Baur, Ionut Damian, Florian Lingenfelser, Johannes Wagner, and Elisabeth André. 2013b. NovA: Automated Analysis of Nonverbal Signals in Social Interactions. In *Human Behavior Understanding*, AlbertAli Salah, Hayley Hung, Oya Aran, and Hatice Gunes (Eds.). Lecture Notes in Computer Science, Vol. 8212. Springer International Publishing, 160–171.

Paul Boersma and David Weenink. 2005. Praat: doing phonetics by computer (version 4.3.14) [computer program]. (2005).

Antonio Camurri, Gualtiero Volpe, Giovanni De Poli, and Marc Leman. 2005. Communicating expressiveness and affect in multimodal interactive systems. *Ieee Multimedia* 12, 1 (2005), 43–53.

George Caridakis, Amaryllis Raouzaiou, Kostas Karpouzis, and Stefanos Kollias. 2006. Synthesizing Gesture Expressivity Based on Real Sequences. *Workshop on multimodal corpora: from multimodal behaviour theories to usable models, LREC Genoa, Italy* (2006).

Cristina Conati and Heather Maclaren. 2009. Empirically building and evaluating a probabilistic model of user affect. *User Modeling and User-Adapted Interaction* 19, 3 (2009), 267–303.

Roddy Cowie, Ellen Douglas-Cowie, Susie Savvidou*, Edelle McMahon, Martin Sawey, and Marc Schröder. 2000. 'FEELTRACE': An instrument for recording perceived emotion in real time. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*.

Roddy Cowie, Gary McKeown, and Ellen Douglas-Cowie. 2012. Tracing Emotion: An Overview. *International Journal of Synthetic Emotions (IJSE)* 3, 1 (Jan. 2012), 1–17.

Jared R. Curhan and Alex Pentland. 2007. Thin Slices of negotiation: predicting outcomes from conversational dynamics withing the first 5 minutes. *Journal of Applied Psychology* 92, 3 (2007), 802–811.

Ionut Damian, Tobias Baur, and Elisabeth André. 2013. Investigating Social Cue-Based Interaction in Digital Learning Games. In *Proceedings of the 8th International Conference on the Foundations of Digital Games*. SASDG.

Nivja H. de Jong and Ton Wempe. 2009. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods* 41, 2 (2009), 385–390.

Wen Dong, Bruno Lepri, Alessandro Cappelletti, Alex Sandy Pentland, Fabio Pianesi, and Massimo Zancanaro. 2007. Using the influence model to recognize functional roles in meetings. In *Proceedings of the 9th international conference on Multimodal interfaces (ICMI '07)*. ACM, New York, NY, USA, 271–278.

Birgit Endraß, Elisabeth André, Matthias Rehm, and Yukiko I. Nakano. 2013. Investigating culture-related aspects of behavior for virtual characters. *Autonomous Agents and Multi-Agent Systems* 27, 2 (2013), 277–304.

Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent Developments in openSMILE, the Munich Open-source Multimedia Feature Extractor. In *Proceedings of the 21st ACM International Conference on Multimedia (MM '13)*. ACM, New York, NY, USA, 835–838.

Patrick Gebhard, Tobias Baur, Ionut Damian, Gregor Mehlmann, Johannes Wagner, and Elisabeth André. 2014. Exploring interaction strategies for virtual characters to induce stress in simulated job interviews. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 661–668.

Patrick Gebhard, Gregor Mehlmann, and Michael Kipp. 2012. Visual SceneMaker - A Tool for Authoring Interactive Virtual Characters. *Special Issue of the Journal on Multimodal User Interfaces: Interacting with Embodied Conversational Agents, Springer-Verlag* 6, 1-2 (2012), 3–11.

Jonathan Gratch, Arno Hartholt, Morteza Dehghani, and Stacy C. Marsella. 2013. Virtual Humans: A New Toolkit for Cognitive Science Research. In *Cognitive Science*. Berlin, Germany.

Hatice Gunes, Massimo Piccardi, and Maja Pantic. 2008. From the lab to the real world: Affect recognition using multiple cues and modalities. (2008).

Torild Hammer. 2000. Mental Health and Social Exclusion among Unemployed Youth in Scandinavia. A Comparative Study. *International. Journal. of Social Welfare* 9, 1 (2000), 53–63.

Jinni A. Harrigan and Kristy T. Taing. 1997. Fooled by a Smile: Detecting Anxiety in Others. *Journal of Nonverbal Behavior* 21, 3 (1997), 203–221.

Mohammed (Ehsan) Hoque, Matthieu Courgeon, Jean-Claude Martin, Bilge Mutlu, and Rosalind W. Picard. 2013. MACH: my automated conversation coach. In *The 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing, UbiComp '13, Zurich, Switzerland, September 8-12, 2013*, Friedemann Mattern, Silvia Santini, John F. Canny, Marc Langheinrich, and Jun Rekimoto (Eds.). ACM, 697–706.

Hayley Hung and Daniel Gatica-Perez. 2010. Estimating Cohesion in Small Groups Using Audio-Visual Nonverbal Behavior. *Multimedia, IEEE Transactions on* 12, 6 (Oct. 2010), 563–575.

Sin-Hwa Kang, Jonathan Gratch, Candy L. Sidner, Ron Artstein, Lixing Huang, and Louis-Philippe Morency. 2012. Towards building a virtual counselor: modeling nonverbal behavior during intimate self-disclosure. In *Proceedings of the 2012 international Conference on Autonomous Agents and Multiagent Systems, Valencia, Spain, June 4-8, 2012 (3 Volumes)*, Wiebe van der Hoek, Lin Padgham, Vincent Conitzer, and Michael Winikoff (Eds.). IFAAMAS, 63–70.

Jonghwa Kim and Elisabeth André. 2008. Emotion recognition based on physiological changes in music listening. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 30, 12 (2008), 2067–2083.

Michael Kipp. 2013. ANVIL: The Video Annotation Research Tool. In *Handbook of Corpus Phonology*. Oxford University Press, Oxford, UK.

Felix Kistler, Birgit Endrass, Ionut Damian, Chi-Tai Dang, and Elisabeth André. 2012. Natural interaction with culturally adaptive virtual characters. *Germany Journal on Multimodal User Interfaces Heidelberg/Berlin* (2012).

Andrea Kleinsmith and Nadia Bianchi-Berthouze. 2011. Form as a Cue in the Automatic Recognition of Non-acted Affective Body Expressions. In *Affective Computing and Intelligent Interaction*, Sidney D'Mello, Arthur Graesser, Björn Schuller, and Jean-Claude Martin (Eds.). Lecture Notes in Computer Science, Vol. 6974. Springer Berlin / Heidelberg, 155–164.

Robert E. Kraut and Robert E. Johnston. 1979. Social and Emotional Messages of Smiling: An Ethological Approach. *Journal of Personality and Social Psychology* 37, 9 (1979), 1539–1553.

Marwa Mahmoud, Louis-Philippe Morency, and Peter Robinson. 2013. Automatic multimodal descriptors of rhythmic body movement. In *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 429–436.

Maurizio Mancini, Laurent Ach, Emeline Bantegnie, Tobias Baur, Nadia Berthouze, Debajyoti Datta, Yu Ding, Stéphane Dupont, Harry J Griffin, Florian Lingenfelser, and others. 2014. Laugh When You're Winning. In *Innovative and Creative Developments in Multimodal Interaction Systems*. Springer, 50–79.

Gary McKeown and Ian Sneddon. 2014. Modeling continuous self-report measures of perceived emotion using generalized additive mixed models. *Psychological methods* 19, 1 (2014), 155.

Gary McKeown, Ian Sneddon, and William Curran. 2015. Gender differences in the perceptions of genuine and simulated laughter and amused facial expressions. *Emotion Review* 7, 1 (2015), 30–38.

Gregor Mehlmann and Elisabeth André. 2012. Modeling Multimodal Integration with Event Logic Charts. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction, ICMI 2012, Santa Monica, California, USA, October 22-26, 2012 (ACM International Conference Proceedings)*. ACM, New York, NY, USA, 125–132.

Gregor Mehlmann, Birgit Endraß, and Elisabeth André. 2011. Modeling Parallel State Charts for Multithreaded Multimodal Dialogues. In *Proceedings of the 13th International Conference on Multimodal Interaction, ICMI 2011, Alicante, Spain, November 14-18, 2011 (ACM International Conference Proceedings)*. ACM, New York, NY, USA, 385–392.

Gregor Mehlmann, Kathrin Janowski, Tobias Baur, Markus Häring, Elisabeth André, and Patrick Gebhard. 2014. Modeling Gaze Mechanisms for Grounding in HRI. In *Proceedings of the 21th European Conference on Artificial Intelligence, ECAI 2014, Prague, Czech Republic, August 18-22, 2014 (Frontiers in Artificial Intelligence and Applications)*. IOS Press Ebooks, Amsterdam, The Netherlands, 1069–1070.

Stephane Michelet, Koby Karp, Emilie Delaherche, Catherine Achard, and Mohamed Chetouani. 2012. Automatic Imitation Assessment in Interaction. In *Human Behavior Understanding*, AlbertAli Salah, Javier Ruiz-del Solar, etin Merili, and Pierre-Yves Oudeyer (Eds.). Lecture Notes in Computer Science, Vol. 7559. Springer Berlin Heidelberg, 161–173.

Louis-Philippe Morency, Candace Sidner, Christopher Lee, and Trevor Darrell. 2007. Head gestures for perceptual interfaces: The role of context in improving recognition. *Artificial Intelligence* 171, 8 (2007), 568–585.

Kevin P. Murphy. 2002. Dynamic bayesian networks. *Probabilistic Graphical Models, M. Jordan* (2002).

Yukiko I. Nakano and Ryo Ishii. 2010. Estimating user's engagement from eye-gaze behaviors in human-agent conversations. In *Proeedings. of the 15th inernational conference on Intelligent user interfaces*. ACM, New York, NY, USA, 139–148.

Radoslaw Niewiadomski, Jennifer Hofmann, Jérôme Urbain, Tracey Platt, Johannes Wagner, Bilal Piot, Huseyin Cakmak, Sathish Pammi, Tobias Baur, Stephane Dupont, Matthieu Geist, Florian Lingenfelser, Gary McKeown, Olivier Pietquin, and Willibald Ruch. 2013. Laugh-aware virtual agent and its impact on user amusement. In *Proceedings of the 2013 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 619–626.

Xueni Pan, Marco Gillies, Chris Barker, David M. Clark, and Mel Slater. 2012. Socially anxious and confident men interact with a forward virtual woman: an experimental study. *PloS one* 7, 4 (2012), e32931.

Maja Pantic, Nicu Sebe, Jeffrey F. Cohn, and Thomas Huang. 2005. Affective Multimodal Human-computer Interaction. In *Proceedings of the 13th Annual ACM International Conference on Multimedia (MULTIMEDIA '05)*. ACM, New York, NY, USA, 669–676.

Allan Pease. 1988. *Body Language*. Sheldon Press, London.

Alex Pentland. 2007. Automatic mapping and modelling of human networks. *Physica* A, 378 (2007), 59–67.

Kaka Porayska-Pomsta, Paola Rizzo, Ionut Damian, Tobias Baur, Elisabeth Andr, Nicolas Sabouret, Hazal Jones, Keith Anderson, and Evi Chryssafidou. 2014. Whos Afraid of Job Interviews? Definitely a Question for User Modelling. In *User Modeling, Adaptation, and Personalization*, Vania Dimitrova, Tsvi Kuflik, David Chin, Francesco Ricci, Peter Dolog, and Geert-Jan Houben (Eds.). Lecture Notes in Computer Science, Vol. 8538. Springer International Publishing, 411–422.

Charles Rich, Brett Ponsleur, Aaron Holroyd, and Candace L. Sidner. 2010. Recognizing engagement in human-robot interaction. In *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction (HRI '10)*. IEEE Press, Piscataway, NJ, USA, 375–382.

Tobias Ruf, Andreas Ernst, and Christian Küblbeck. 2011. Face detection with the sophisticated high-speed object recognition engine (SHORE). In *Microelectronic Systems*. Springer, 243–252.

Georgia Sandbach, Stefanos Zafeiriou, Maja Pantic, and Lijun Yin. 2012. Static and dynamic 3D facial expression recognition: A comprehensive survey. *Image and Vision Computing* 30, 10 (Oct. 2012), 683–697.

Stefan Scherer, Stacy Marsella, Giota Stratou, Yuyu Xu, Fabrizio Morbini, Alesia Egan, Albert(Skip) Rizzo, and Louis-Philippe Morency. 2012. Perception Markup Language: Towards a Standardized Representation of Perceived Nonverbal Behaviors. In *Intelligent Virtual Agents*, Yukiko Nakano, Michael Neff, Ana Paiva, and Marilyn Walker (Eds.). Lecture Notes in Computer Science, Vol. 7502. Springer Berlin Heidelberg, 455–463.

Thomas Schmidt. 2004. Transcribing and annotating spoken language with EXMARaLDA. In *Proceedings of the International Conference on Language Resources and Evaluation: Workshop on XML based richly annotated corpora, Lisbon 2004*. ELRA, Paris, 879–896. EN.

Marc Schröder, Elisabetta Bevacqua, Roddy Cowie, Florian Eyben, Hatice Gunes, Dirk Heylen, Mark Ter Maat, Gary McKeown, Sathish Pammi, Maja Pantic, Catherine. Pelachaud, Björn. Schuller, Etienne de Sevin, Michel Valstar, and Martin Wöllmer. 2012. Building Autonomous Sensitive Artificial Listeners. *Affective Computing, IEEE Transactions on* 3, 2 (April 2012), 165–183.

Nicu Sebe, Ira Cohen, Theo Gevers, and Thomas S. Huang. 2006. Emotion Recognition Based on Joint Visual and Audio Cues. In *Proceedings of the 18th International Conference on Pattern Recognition - Volume 01 (ICPR '06)*. IEEE Computer Society, Washington, DC, USA, 1136–1139.

Candace L. Sidner, Cory D. Kidd, Christopher Lee, and Neal Lesh. 2004. Where to look: a study of human-robot engagement. In *Proceedings of the 9th international conference on Intelligent user interfaces*. ACM Press, New York, NY, USA, 78–84.

Anne Loomis Thompson and Dan Bohus. 2013. A Framework for Multimodal Data Collection, Visualization, Annotation and Learning. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction (ICMI '13)*. ACM, New York, NY, USA, 67–68.

David R. Traum, David DeVault, Jina Lee, Zhiyang Wang, and Stacy Marsella. 2012. Incremental Dialogue Understanding and Feedback for Multiparty, Multimodal Conversation. In *Intelligent Virtual Agents - 12th International Conference, IVA 2012, Santa Cruz, CA, USA, September, 12-14, 2012. Proceedings (Lecture Notes in Computer Science)*, Yukiko Nakano, Michael Neff, Ana Paiva, and Marilyn A. Walker (Eds.), Vol. 7502. Springer, 275–288.

Thurid Vogt, Elisabeth André, and Nikolaus Bee. 2008. EmoVoice - A Framework for Online Recognition of Emotions from Voice. In *Perception in Multimodal Dialogue Systems, 4th IEEE Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems, Kloster Irsee, Germany (Lecture Notes in Computer Science)*. Springer, 188–199.

Johannes Wagner, Florian Lingenfelser, Tobias Baur, Ionut Damian, Felix Kistler, and Elisabeth André. 2013. The Social Signal Interpretation (SSI) Framework - Multimodal Signal Processing and Recognition in Real-Time. In *Proceedings of ACM MULTIMEDIA 2013*. Barcelona.

Harald G. Wallbott. 1998. Bodily expression of emotion. *European Journal of Social Psychology* 28 (1998), 879–896.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: A professional framework for multimodality research. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)* (2006), 879–896.

Xuehan Xiong and Fernando De la Torre. 2013. Supervised descent method and its applications to face alignment. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 532–539.

Online Appendix to: Context-Aware Automated Analysis and Annotation of Social Human-Agent Interactions

TOBIAS BAUR, Augsburg University
GREGOR MEHLMANN, Augsburg University
IONUT DAMIAN, Augsburg University
PATRICK GEBHARD, DFKI GmbH
FLORIAN LINGENFELSER, Augsburg University
JOHANNES WAGNER, Augsburg University
BIRGIT LUGRIN, Augsburg University
ELISABETH ANDRÉ, Augsburg University

Expressivity Features

Energy/Power (EN) represents the dynamic properties of a movement (e.g. weak versus strong). It is calculated from the first derivative of the motion vectors in all three dimensions where $\vec{m}()$ is the motion of the specified joint relative to the torso joint and n is the number of frames considered for the calculation.

$$EN = \sqrt{\sum_{i=0}^{n}((\vec{m}(i).x^2 + \vec{m}(i).y^2 + \vec{m}(i).z^2)/3)/n}$$

Fluidity (FL) differentiates smooth movements from jerky ones. This feature aims to capture the continuity between movements. It is calculated as the sum of the variance (Var) of both hands' motion vectors' norms ($\vec{l}, \vec{r}$) (respectively feet for leg postures).

$$FL = Var(\sum_{i=0}^{n}\vec{l}(i)/n) + Var(\sum_{i=0}^{n}\vec{r}(i)/n)$$

Spatial extent (SE) is modeled as the space used for gesturing in front of the recorded person. It is calculated as the maximum Euclidean distance of the position of the two hands (l,r) (respectively feet for leg postures).

$$SE = max(d(\mid r(i) - l(i) \mid))$$

Overall activation (OA) represents the quantity of the movement (passive versus active). It is calculated as the sum of the motion vectors' norm of both hands (respectively feet for leg postures):

$$OA = \sum_{i=0}^{n} \mid \vec{r}(i) \mid + \mid \vec{l}(i) \mid$$

Temporal extent (TP) represents the duration of a gesture (short vs sustained). The duration of each gesture is computed from the starting and end points synchronized with the recording time in the SSI framework.

DOI : http://dx.doi.org/10.1145/0000000.0000000

Questionnaire for Interviewers

Questionnaire A (Day 1 and Day 3)

The overall performance of the person during the interview was:

Very bad ○ ○ ○ ○ ○ ○ ○ Very good

Related to behaviour, would you recommend to hire the participant?

I don't agree ○ ○ ○ ○ ○ ○ ○ I agree

The participant smiled appropriately during the Interview

I don't agree ○ ○ ○ ○ ○ ○ ○ I agree

The participant kept appropriate eye contact during the interview

I don't agree ○ ○ ○ ○ ○ ○ ○ I agree

The participant gesticulated appropriately during the interview

I don't agree ○ ○ ○ ○ ○ ○ ○ I agree

The participant appeared nervous during the interview

I don't agree ○ ○ ○ ○ ○ ○ ○ I agree

The participant appeared interested during the interview

I don't agree ○ ○ ○ ○ ○ ○ ○ I agree

The participant appeared focused during the interview

I don't agree ○ ○ ○ ○ ○ ○ ○ I agree

Questionnaire for Participants

Questionnaire B (Day 1 and Day 3)

I think I performed well in the job interview

I don't agree ○ ○ ○ ○ ○ ○ ○ I agree

I was nervous during the interview

I don't agree ○ ○ ○ ○ ○ ○ ○ I agree

I used a lot of filler words like "uhm" or "er"

I don't agree ○ ○ ○ ○ ○ ○ ○ I agree

I was not focused during the interview

I don't agree ○ ○ ○ ○ ○ ○ ○ I agree

I paid particular attention to my nonverbal behaviour

I don't agree ○ ○ ○ ○ ○ ○ ○ I agree

I think my nonverbal behaviour was appropriate

I don't agree ○ ○ ○ ○ ○ ○ ○ I agree

User Experience Questionnaire

System interaction: experiment group (Day 2)

I found the video self-reflection useful

I don't agree ○ ○ ○ ○ ○ ○ ○ I agree

I learned something by watching my video session

I don't agree ○ ○ ○ ○ ○ ○ ○ I agree

I would use the training system for preparing a real job interview

I don't agree ○ ○ ○ ○ ○ ○ ○ I agree

I found the gaming cards helpful

I don't agree ○ ○ ○ ○ ○ ○ ○ I agree

I had fun playing the game

I don't agree ○ ○ ○ ○ ○ ○ ○ I agree