Investigating Social Cue-Based Interaction in Digital Learning Games

Ionut Damian
Human Centered Multimedia
Augsburg University, Germany
damian@hcm-lab.de

Tobias Baur
Human Centered Multimedia
Augsburg University, Germany
baur@hcm-lab.de

Elisabeth André
Human Centered Multimedia
Augsburg University, Germany
andre@hcm-lab.de

ABSTRACT

This paper studies the potential of signal processing techniques to generate social cue-based interaction in the context of a job interview simulation game. To this end, we investigate how social cues can be automatically recognized using state-of-the-art in sensor technology and we provide a perspective on how social cue recognition can be used to generate believable interaction between the user and the system, and thus enhance game experience.

1. INTRODUCTION

One large issue Europe faces is the rising number of young people who are out of employment, education or training (NEETs). NEETs often have underdeveloped socio-emotional and interaction skills [9, 12], such as a lack of self-confidence, lack of sense of their own strengths or social anxiety [15]. This can cause 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 specialised inclusion centres meant to aid young people secure employment through coaching by professional practitioners. One problem of this approach is that it is very 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.

Job interviews are used by the potential future employer as means to determine whether the interviewee is suited for the company's needs. One way the interviewer asses this is using social cues, i.e. actions, conscious or unconscious, of the interviewee that have a specific meaning in a social context such as a job interview.

In this paper we present an approach to using signal processing techniques to generate credible interaction in a digital game meant to help youngsters improve social skills which are pertinent to job interviews. The game is being developed as part of the TARDIS project and it will use virtual agents acting as recruiters during scenario-based job interview simulations. Using a game-like approach for this purpose is especially appealing as it offers the users a motivating and rewarding experience which is known to enhance the learning process [19].

In particular, this paper will be investigating whether it is possible to automatically recognise job interview relevant social cues manifested by the interviewees. Once recognised, such social cues can be used to generate seamless interaction between the user and the system as well as to perform a detailed analysis of the user's behaviour in real-time. This in turn would allow the system to react to the user's behaviour in an intelligent way, e.g. altering the scenario in real-time based on the user's performance.

2. RELATED WORK

A growing amount of literature demonstrates the power of social cues that are consciously or unconsciously shown by people in various situations. Varni et al. [24] studied social cues in small groups to understand the synchronization of affective behaviour and the emergence of functional roles, such as leadership. McGovern and Tinsley [14] as well as Arvey [1] found that nonverbal behaviours, such as eye gaze contact, body movement and voice tone, significantly bias the assessment of the job interviewers. Hence, the use of non-verbal behaviours and their impact on the success of a job interview has become a major focus of research. Curhan and Pentland [4] observed that speech activity, conversational engagement, prosodic emphasis, and vocal mirroring were highly predictive of the outcome of simulated job interviews.

In order to help people train social skills, a variety of techniques have been developed, such as role playing, group discussions or specific exercises [8]. The need for effective social training has also inspired a number of several proposals of computer-based simulation environments as additional platforms for delivering such training for a variety of applications including job interviews [18], inter-cultural communication [6]. Similarly, Pan et al.[15] and Pertaub et al.[17] investigated whether interaction with virtual characters can help people suffering from social anxiety.

The objective of this paper is to investigate the potential of social signal processing in the context of a job interview game. Earlier research in the area of social signal processing typically focused on a limited set of prototypical behaviours[5, 10]. Nevertheless, only few researchers explored techniques from the area of social signal processing in interactive scenarios. An example is Scherer's health care agent Ellie [20] which was designed to help detect behaviours related to depression and post-traumatic stress disorder and to offer related information if needed.

To apply social signal processing in the context of job interviews, a comprehensive set of prototypical and less prototypical behaviour needs to be recognised that provides a sufficient basis for an assessment of the impression an interviewee conveys.

3. THEORETICAL FOUNDATIONS

The strength of social cues lies in the communication of implicit information. In contrast to natural language, the non-verbal channel offers more indirect information which can, in many cases, be even more meaningful than spoken language. We propose the following four categories of information that are implicitly conveyed in social interactions by body language: *Social Attraction*, *Engagement*, *Self Efficacy* and *Attitude*. Each of these categories can be used as a factor that influences a game scenario such as changing the environments settings or an agent's attitude towards the player.

Social Attraction refers to the amount of appreciation a person evokes in others [23]. The relation of social attraction and body language has been investigated in various studies in social sciences. McGinley et al. [13] conclude that open body positions are usually received more positive than positions with arms or legs crossed. According to Schouwstra and Hoogstraten [21] upright postures with the head up receive more positive judgements than the opposite.

According to Sidner and colleagues [22], *Engagement* "is the process by which two (or more) participants establish, maintain and end their perceived connection during interactions they jointly undertake". Pease [16] demonstrates how engagement is portrayed by an orientation of the body and face towards the interlocutor. Another aspect of engagement is the overall activation of the movements. Here it is necessary to distinguish whether the communicator is speaking or listening. While speakers tend to show their engagement by a high amount of overall activity, in the role of a listener, interlocutors should show less overall activity because such a behaviour is usually interpreted as a sign of distraction.

People with a high amount of *Self Efficacy* are confident that they will be able to master difficult situations [2]. Self efficacy is usually conveyed by calm, fluid and high energy movements while quick and jerky movements tend to make a person appear nervous. In addition, a high amount of self manipulations, such as scratching one's head, reveals the anxiety of people in a social situation. Pease [16] provides various examples of body postures that signal a high amount of dominance, such as placing both feet apart or both hands behind the head with the elbows facing outward.

In psychology, the term *Attitude* refers to the expression of favour or disfavour towards a particular person or theme [7]. Usually, open body postures, such as opened arms, are interpreted as a sign of willingness to cooperate while closed body postures, such as crossing one's arms, rather communicate the opposite [16].

Recognising such high level body language interpretations can be, however, very difficult to do automatically as people show great individualism in displaying these. To this end, our work focuses on recognising distinct social cues which can then be mapped to high level states such as *Social Attraction*, *Engagement*, *Self Efficacy* and *Attitude*.

4. THE SYSTEM

We implemented a system meant to record and analyse social and psychological signals from users and recognise predefined social cues in real time in the context of a digital learning game for acquiring job interview related social skills.

The system uses a combination of sensors and software algorithms which offer good results in terms of accuracy, low intrusion, reliability, set-up time and cost. High accuracy ensures that a youngster's social cues are correctly recognised and allows the game itself to correctly react to them. It is equally important that the approach has a low intrusion factor. For example, biological signal sensors are not feasible in this scenario because attaching various sensors to the skin of the users will most likely result in an increase in stress which might have a negative effect on the user's job interview performance, but may not be actually indicative of the user's actual abilities. Therefore, in the context studied, remote sensors are preferred.

The reliability of an approach is also important as the game should be able to be operated in different environments, from schools and specialised inclusion centres to the private homes of the users. The system should also have a short set-up time and a simple set-up procedure to ensure that non-experts can operate the system. The final characteristic we identified as important is the cost of the system. This has a direct influence on who can use the system. An expensive system which uses complex sensors might be more accurate but it will be of no use if the training institutions cannot afford it.

For recording and pre-processing human behaviour data, our system relies on the SSI framework which was developed as part of our previous work [25]. It provides a variety of tools for the real-time analysis and recognition of such data. We chose the Microsoft Kinect as the main sensor as it enables us to recognize a broad range of social cues using only a single sensor. The main advantages of this sensor are: It is low-cost, it does not require any time-consuming configuration, it is relatively robust against lighting conditions and it incorporates a microphone and an RGB camera in addition to the depth camera. Furthermore, because it is a remote sensor it has a minimal intrusion level. There are also software development kits for skeleton and face tracking available which provide a good starting point for human behaviour analysis. In our system, the Kinect is connected to the system using the Microsoft Kinect SDK 1.7 .

As a first step, we implemented the recognition of the following six social cues:

- *Hands to face.* This is a self manipulation type social cues which has a negative correlation with the *self efficacy* of the user [2].
- *Looking away* is an important cue for determining the level of *engagement.*
- *Posture: Arms crossed, Arms open, Hands behind head. Arms crossed* and *arms open* have been found to correlate with both *social attraction* and *attitude* [16, 13] whereas in Section 3 we argue that *hands behind head* is associated with dominance, and thus has an impact of the *self efficacy* dimension.
- *Leaning back, Leaning forward.* These behaviours are often coupled with *engagement.*

- *Voice activity.* It can be used to determine whether the user is currently the speaker or the listener. This is important, for example, when determining the *engagement* level. *Voice activity* can also be used to compute other social cues such as interrupting the interviewer, short answers or long silence.

In addition to these social cues, our system is also able to compute the expressivity of the user's movements. This gives further information on the *engagement and* self efficacy level of the user as discussed in 3.

Looking Away. To detect head gaze, we use the face tracking library provided with the Microsoft Kinect SDK Toolkit . This library uses both the RGB information and the depth information to track the face of the user and compute several characteristics . Out of these characteristics, the most important one to us is the head pose data. This allows our system to determine the orientation of the user's head. After this is determined, we use a threshold-based approach to detect when the user is looking straight ahead, to the left or to the right.

Voice Activity. In order to detect when the user is talking our system looks at the audio signal provided by a microphone. To ensure accurate results, we decided to use a close-talk microphone instead of the one incorporated in the Microsoft Kinect. The main advantage of the close-talk microphone is that it filters out most of the environmental noise. For the voice activity detection itself we use a binary Signal-To-Noise filter, which uses a threshold-based approach to categorize an audio sample into noise and non-noise, in our case voice activity. The filter also enforces a minimal duration of 0.1s and minimal silence duration of 1.0s. This makes the system more robust towards environmental noise, interjections or short pauses in speech. Preliminary tests showed that the system is able to accurately detect when the user is talking and is largely unaffected by environmental noise.

Gestures, Postures and Leaning. To recognise gestures, postures and leaning occurrences we use the FUBI component [11] of our framework. It is able to recognise predefined postures and gestures using a state machine like approach. The gestures and postures to be recognised can be defined using an XML-based specification language. For the initial version of our system we focus on the following postures: Arms open (a), hands behind head (b), left/right hand close to face (c), leaning forward/back (d). These are exemplified in Fig. 1.

Expressivity Features. The system is able to determine the energy, overall activation, spatial extent and fluidity of the user's movement. This is done in accordance to the work by [3].

5. EVALUATION

With the help of a user study, we evaluated the robustness of each recogniser. For this, we recruited 11 persons (10 male and 1 female) with an average age of 30.4. The participants had a medium amount of experience with the Microsoft Kinect (mean value 3.09 on a scale from 1, no experience at all, to 5, practically daily usage).

Design and Procedure. Each participant was shown a series of social cue descriptions and was asked to perform the specified social cue as soon as it appears, hold it for a couple of seconds and then return to a normal body position. The social cues we used were: *look right, arms open, hands behind head, arms crossed, hand to face, lean back, lean forward* and *voice activity.* The order in which the social cue descriptions were displayed was counterbalanced between the participants to compensate for learning effects or any possible stress users might experience at the start of the study. A Microsoft Kinect was positioned in front of the participants at a distance of approximately 1.1m from the participants' head and 1.3m of the ground. The participants also wore an AKG C 444 close-talk microphone.

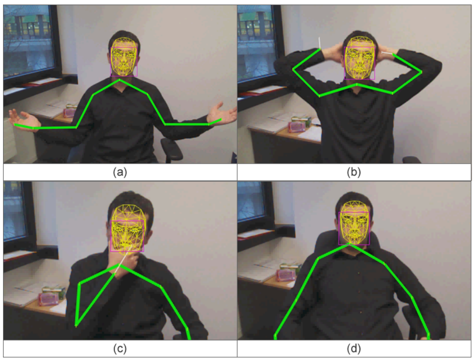

Figure 1: Examples of the gestures our system can recognise.

Table 1: Results of the evaluation showing the mean recognition rates of each social cue.

Social Cue	Recall
Arms open	100.0%
Hands behind head	100.0%
Arms crossed	81.8%
Hand to face	90.9%
Lean backward	72.7%
Lean forward	81.8%
Voice activity	100.0%

Results and Discussion. The evaluation of the data yielded an overall mean recognition rate of 88%, with 3 social cues (*arms open*, *hands behind head* and *voice activity*) achieving 100%. The worst recognition rate was seen for the *lean back* social cue with 72.7%. The results are shown in Table 1. The main reason for the non-perfect recognition rate was the rather unstable tracking provided by the Microsoft Kinect SDK with the users sitting down. However, if we consider the benefits of the Microsoft Kinect (low cost, minimal intrusion), it becomes immediately clear that it is the best solution currently available in this context. Other skeleton tracking sensors, such as motion capturing systems, have a much higher intrusion level and an increased set-up time and complexity.

6. CONCLUSION

This paper shows an approach to using social signal processing techniques in the context of a digital game with the ultimate goal of helping people not in employment, education or training (NEETs) improve job interview pertinent social skills. To this end, we investigated how social cues can be automatically recognised in real-time in order to generate seamless interaction between the user and the system as well as enabling a detailed analysis of the user's behaviour during a simulated interview.

We implemented a system based on the SSI framework [25] which can recognise various social cues using consumer depth cameras such as the Microsoft Kinect. The accuracy of the system was evaluated using a user study. The results of the study (recognition rate > 88%) suggest that bodily behaviours pertinent to job interviews can be reliably recognised by low cost depth sensors. Problems encountered for the recognition components were mainly due to the sensitivity of the depth sensor.

The system presented in this paper can recognise a small set of social cues. While this is a clear limitation which we plan to rectify as part of our future work, it did not affect the goal of the paper which was to investigate the potential of such a system in a job interview scenario. Our evaluation study showed that social cue recognition is viable in such an environment. Our ultimate goal is to develop a digital game that allows the user to participate in a job interview simulation where the job recruiter is a virtual character. The social cue recognition will provide a basis for credible interaction between the recruiter and the user. It will allow the virtual recruiter to recognize and react to the user's social cues in real-time similar to how a real interviewer might do. The goal of this is to generate immersion, and thus improve the learning effect.

As part of our future work we intend to implement recognisers for the additional social cues (i.e. smiling, laughter, voice features) and to investigate the use of other channels (i.e. eye gaze). Additional user studies using actual NEETs are also planned. These will allow us to more accurately test the capabilities of our system in the desired context.

Acknowledgement. This work was partially funded by the European Comission within FP7-ICT-2011-7 (Project TARDIS, grant agreement no. 288578).

7. REFERENCES

[1] R. D. Arvey and J. E. Campion. The employment interview: A summary and review of recent research. *Personnel Psychology*, 35(2):281–322, 1982.

[2] A. Bandura. *Self Efficacy: The Exercise of Control*. Palgrave Macmillan, New York, N.Y., 1997.

[3] G. Caridakis, A. Raouzaiou, K. Karapouzis, and S. Kollias. Synthesizing gesture expressivity based on real sequences. *Workshop on multimodal corpora: from multimodal behaviour theories to usable models, LREC Conf. Genoa, Italy*, Mai 2006.

[4] J. Curhan and A. Pentland. Thin slices of negotiation: predicting outcomes from conversational dynamics within the first 5 minutes, 2007.

[5] W. Dong, B. Lepri, A. Cappelletti, A. S. Pentland, F. Pianesi, and M. Zancanaro. Using the influence model to recognize functional roles in meetings. In *Proc. 9th Intl. Conf. on Multimodal interfaces*, ICMI '07, pages 271–278, New York, NY, USA, 2007. ACM.

[6] B. Endrass, E. André, M. Rehm, and Y. Nakano. Investigating culture-related aspects of behavior for virtual characters. *Autonomous Agents and Multi-Agent Systems*, 2013.

[7] J. P. Forgas, J. Cooper, and W. D. Crano. *The Psychology of Attitudes and Attitude Change*. Taylor & Francis Group, New York, NY, 2010.

[8] J. Greene and B. Burleson. *Handbook of Communication and Social Interaction Skills*. LEA's Communication Series. L. Erlbaum Associates, 2003.

[9] T. Hammer. Mental health and social exclusion among unemployed youth in scandinavia. a comparative study. *Intl. Journal of Social Welfare*, 9(1):53–63, 2000.

[10] H. Hung and D. Gatica-Perez. Estimating cohesion in small groups using audio-visual nonverbal behavior. *Trans. Multi.*, 12(6):563–575, Oct. 2010.

[11] F. Kistler, B. Endrass, I. Damian, C. T. Dang, and E. André. Natural interaction with culturally adaptive virtual characters. *Journal on Multimodal User Interfaces*, 6:39–47, 2012.

[12] R. MacDonald. Disconnected youth? social exclusion, the underclass and economic marginality, 2008.

[13] H. McGinley, R. LeFevre, and P. McGinley. The influence of a communicator's body position on opinion. *Journal of Personality and Social Psychology*, 31(4):686–690, 1975.

[14] T. V. McGovern, B. W. Jones, and S. E. Morris. Comparison of professional versus student ratings of job interviewee behavior, 1979.

[15] X. Pan, M. Gillies, Barker, D. M. C. M. Clark, and M. Slater. Socially anxious and confident men interact with a forward virtual woman: An experiment study. *PLoS ONE*, 7(4):e32931, 2012.

[16] A. Pease. *Body Language: How to read other's thoughts by their gestures*. Sheldon Press, London, tenth impression 1988 edition, 1988.

[17] D.-P. Pertaub, M. Slater, and C. Barker. An experiment on public speaking anxiety in response to three different types of virtual audience. *Presence: Teleoper. Virtual Environ.*, 11(1):68–78, Feb. 2002.

[18] H. Prendinger and M. Ishizuka. The empathic companion: A character-based interface that addresses users' affective states. *Applied Artificial Intelligence*, 19(3-4):267–285, 2005.

[19] M. Prensky. *Digital Game-Based Learning*. Paragon House, 2007.

[20] S. Scherer, S. Marsella, G. Stratou, Y. Xu, F. Morbini, A. Egan, A. Rizzo, and L.-P. Morency. Perception markup language: Towards a standardized representation of perceived nonverbal behaviors. In *Intelligent Virtual Agents*, volume 7502 of *LNCS*, pages 455–463. Springer Berlin Heidelberg, 2012.

[21] S. Schouwstra and J. Hoogstraten. Head position and spinal position as determinants of perceived emotional state. *Perceptual and Motor Skills*, 81:673–674, 1995.

[22] C. L. Sidner, C. D. Kidd, C. Lee, and N. Lesh. Where to look: a study of human-robot engagement. In *IUI '04: Proc. 9th Intl. Conf. on Intelligent user interfaces*, pages 78–84, NY, USA, 2004. ACM.

[23] J. A. Simpson and B. A. Harris. Interpersonal attractione. In A. L. Weber and J. H. Harvey, editors, *Perspectives on close relationships*, pages 45–66. Prentice Hall, 1994.

[24] G. Varni, G. Volpe, and A. Camurri. A system for real-time multimodal analysis of nonverbal affective social interaction in user-centric media. *Multimedia, IEEE Transactions on*, 12(6):576–590, 2010.

[25] J. Wagner, F. Lingenfelser, and E. André. The social signal interpretation framework (SSI) for real time signal processing and recognition. In *Proc. Interspeech 2011*, 2011.