A Step towards Adaptive Multimodal Virtual Social Interaction Platform for Children with Autism

Esubalew Bekele[1], Mary Young[1], Zhi Zheng[1], Lian Zhang[1], Amy Swanson[3], Rebecca Johnston[3], Julie Davidson[2,3], Zachary Warren[2,3], and Nilanjan Sarkar[4,1]

[1] Electrical Engineering and Computer Science Department
[2] Pediatrics and Psychiatry Department
[3] Treatment and Research in Autism Spectrum Disorder (TRIAD)
[4] Mechanical Engineering Department, Vanderbilt University, Nashville, TN, USA
{esubalew.bekele,nilanjan.sarkar}@vanderbilt.edu

Abstract. Recent advances in computer and robotic technology are enabling the application of such technology in assisting traditional intervention in developmental disorders such as autism spectrum disorders (ASD). A number of research studies indicate that many children with ASD prefer technology and this preference can be explored to develop systems that may alleviate several challenges of traditional treatment and intervention. The current work proposes to develop an adaptive virtual reality-based social interaction platform for children with ASD. It is hypothesized that endowing a technological system that can detect the feeling and state of the child and adapt its interaction accordingly is of great importance in assisting and individualization of traditional intervention approaches. The proposed system employs sensors such as eye trackers and physiological signal monitors and models the context relevant psychological state of the user from combination of these sensors together with the performance of the participant.

Keywords: Social interaction, virtual reality, multimodal system, adaptive interaction, eye tracking, physiological processing, autism intervention.

1 Introduction

Autism spectrum disorders (ASD) defines a spectrum of developmental disorders that are associated with social, communicative and language deficits [1]. Children with ASD exhibit difficulties in communicative interactions and generally poor social skills [2]. Researchers have shown that children with ASD have fewer social communication skills and that this deficit is related to executive function skills [3]. It was shown that children with ASD exhibit severe deficits in facial and vocal affect recognition, social judgment, problem solving and social functioning skills [4] and hence a deficit in social interaction is the major defining feature of ASD. Generally, these common social and communicational deficits are observed in most children with autism, however, the manifestation of these deficits is quite different from one individual to another [5]. These individual differences call for approaches to individualize the

C. Stephanidis and M. Antona (Eds.): UAHCI/HCII 2013, Part II, LNCS 8010, pp. 464–473, 2013.

therapy as opposed to one-therapy-fits-all strategies. Among the most common social interaction skills that lacks in children with ASD is the ability to use appropriate language in a social context [6]. Traditional intervention requiring intensive behavioral sessions results in excessive life time costs and inaccessibility of the therapy for the larger population [7]. Recent assistive technologies have the potential to at least lessen the burden of human therapists and increase effectiveness of traditional human-centric autism therapy. Additionally, technology enabled systems are equipped with objective measures that may be better suited than subjective assessments by human therapists in certain situations. Literature suggests that children with ASD are highly motivated by computer-based intervention tasks [8]. Predictability, objectivity, lack of judgmental behavior, consistency of clearly defined task and the ability to direct focus of attention due to reduced distractions from unnecessary sensor stimuli are among the benefits of technology-enabled therapy [6].

Virtual reality (VR) [9,10] have been proposed for ASD intervention. VR platforms are promising for improving social skills, cognition and overall social functioning in autism [11]. Despite the potential to automatically detect and adapt to the social interaction in VR systems, most existing VR systems as applied to autism therapy focus on performance and explicit user feedback as primary means of interaction with the participant [12]. Therefore, adaptive interaction is limited in these systems. Adaptive social interaction using implicit cues from sensors such as peripheral physiological signals [13] and eye tracking [14] was shown to be possible in VR-based autism therapy. In order to undertake naturalistic social interaction several components have to be developed including conversational dialog, body language (gesture), facial emotional expressions and eye contact. Conversational dialog is an important part of social interaction. Recently spoken conversational modules have been incorporated into VR systems to achieve more natural interaction instead of menu driven dialog management. Instead of large vocabulary, domain independent natural language understanding, limited vocabulary question-response dialog management, which is focused on the specific domain, has been shown to be effective [15,16]. Such multimodal interaction helps in individualization and in cases of inaccessibility of trained therapists, it may serve as a self-contained therapeutic system.

This study is aimed at designing and developing an innovative adaptive VR-based multimodal social interaction platform. The platform integrates peripheral psychophysiological signal monitoring for affect detection, eye tracking and gaze metrics for engagement recognition and spoken question-answer-based dialog management for a more naturalistic interaction. The remainder of the paper is organized as follows. Section 2 describes the overall process of developing the system. Section 3 discusses the current status of the system. Finally, Section 4 concludes the discussion by highlighting the future direction of the system.

2 System Design

This section describes an ongoing work on a larger VR-based adaptive intervention system for adaptive multimodal social interaction. The effort to build the overall system involves development at various stages and usability studies are performed at various milestones. This section discusses the overall system in detail and the next sections present the current status of the system. Day-to-day social behaviors are expressions of one's attitude towards social situation and interaction and are manifested through verbal conversations and various non-verbal behavioral cues such as facial expressions, body postures and gestures, and vocal outbursts like laughter (Fig. 1 Left)[17]. We are endowing the system to capture most of these verbal and non-verbal cues to facilitate a more natural, individualistic and adaptive virtual social interaction.

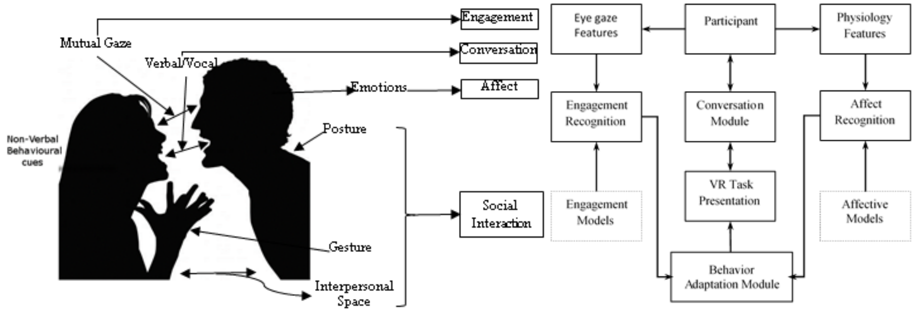

Fig. 1. Left: Verbal and non-verbal components of social interaction. Right: Schematics of the VR-based adaptive multimodal social interaction platform.

The overall system is composed of four major components: (1) an adaptive social task presentation VR module, (2) a spoken conversation management module (Q/A-based natural language processing, NLP module), (3) a synchronous physiological signal monitoring and physiological affect recognition module, and (4) a synchronous eye tracking and engagement detection module (Fig. 1 Right). All separate components of the system run independently in parallel, while sharing data via light-weight network sockets message passing. The VR task presentation engine is built on top of the popular game engine Unity (www.unity3d.com) by Unity Technologies. The peripheral psychophysiological monitoring application was built using the software development kit (SDK) of the wireless BioNomadix physiological signals acquisition device by Biopac Inc. (www.biopac.com). The eye tracker application employed the Tobii X120 remote desktop eye tracker SDK by Tobii Technologies (www.tobii.com).

2.1 The VR Task Presentation Engine

The VR environment is mainly built on and rendered in Unity game engine. However, various 3D software such as online animation and rigging service, Mixamo (www.mixamo.com), and Autodesk Maya were employed for character customization, rigging and animation. The venue for the social interaction task is a virtual school cafeteria (Fig. 2). This environment was chosen for the targeted age group (i.e., 13-17 year old), because it fosters various conversation and interactions for teenagers. The cafeteria was built using a combination of Google Sketchup and Autodesk Maya and was then imported into the Unity. A pack of 12 fully rigged virtual characters (10 teenagers and 2 adults) with 20 facial bones for emotional expressions and several body bones for various gestural animations were developed.

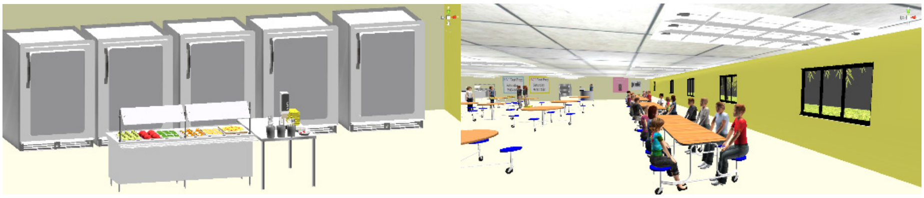

Fig. 2. The cafeteria environment. Food dispensary (left) and dining area (right).

Each avatar's face was animated in Maya to generate Ekman's 7 universal emotional facial expressions [15]. Each of these emotions was generated with 4 degrees of arousal (e.g., low, medium, high and extreme). Fig. 3 (Left) shows some of the characters while displaying some of the gestural animations they are capable of.

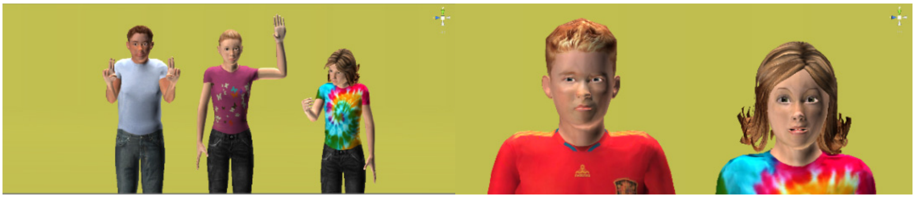

Fig. 3. Left: Representative characters displaying example gestural animations. Right: Anger (left) and surprise (right) facial expression animations.

As facial emotional expressions are major parts of the non-verbal communication cues in social interactions, in a separate study we have conducted a usability study of a VR-based facial emotional expression recognition using 10 typically developing children and 10 children with ASD. The results indicated that there exists inherent pattern difference in the way children with ASD processed the emotional faces and recognize them. Fig. 3 (Right) shows two examples of emotional expressions.

In addition to the facial expressions and gestural and some utility animations, seven phonetic viseme poses were also created for lip-syncing spoken audio by the avatars in the verbal conversation part of the social interaction.

2.2 Spoken Dialog Manager

The verbal conversation is managed by a spoken dialog management module which was developed using the Microsoft speech recognizer from the speech API (SAPI) with domain specific grammar and semantics. The conversation module is based on question-answer dialog and it contained conversational threads for easy (level 1), medium (level 2) and hard (level 3) social tasks with each level having 4 conversational task blocks. Each block in a level has a mission that the participant is expected to accomplish. Each conversation block was represented by a tree of dialog with nodes representing each option and a particular branch in the tree representing the dialog alternative paths from the initial question to the final correct answer. Failure and success is measured in each conversational block and there is a hierarchical scoring mechanism that keeps track of performance in conversation block level as well as mission level. Options in each block of conversation are presented to the participant using a list of items and the participant speaks out their choice through a microphone. Kinect is employed for this purpose as its microphones have superior sound localization and background noise cancellation features.

Overall performance, i.e., success/failure (S/F) is used to switch across missions (levels) as shown in Fig .7 in the "performance only" version of the system. In the adaptive system, physiological affect recognition as well as eye tracking-based engagement detection are combined to adapt the level of difficulty of the interaction in addition to the overall performance of the participant.

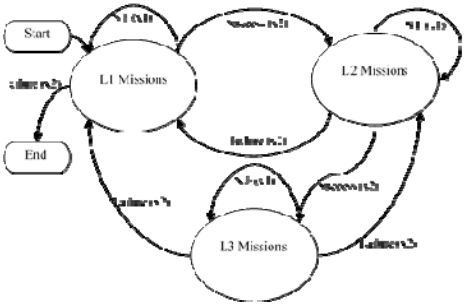

Fig. 4. Finite state diagram showing a level switching logic

2.3 Physiological Monitor

The physiological monitoring application collects 8 channels of physiological data and was developed using the Biopac software development kit (SDK) and BioNomadix wireless physiological acquisition modules with a sampling rate of 1000 Hz. The physiological signals that were monitored were: electrocardiogram (ECG), pulse plethesymogram (PPG), skin temperature (SKT), galvanic skin response (GSR), 3 electromyogram (EMG), and respiration (RSP). Due to social communication impairments in adolescents with autism, they are not usually expressive of their internal affective states and these states often are not visible externally [18,19]. Physiological signals are, however, not affected by these impairments and can be useful in understanding the internal psychological states [20,19]. Among the signals we monitored, GSR, PPG, and ECG are directly related to the sympathetic response of the autonomic nervous system (ANS) [21]. When there is increased sympathetic activity due to external factors and pressures, the heart rate, the blood pressure, and sweating are all elevated [19]. Various features extracted out of these signals are used for supervised training of a machine learning algorithm for later affective state classification in the actual interaction.

2.4 Eye Gaze Tracker

The eye gaze tracker application was developed using Tobii SDK. The remote desktop eye tracker, Tobii X120, is used at 120 Hz frame rate that allows a free head movement of 30 x 22 x 30 cm (width x height x depth) at 70 cm distance. We run two applications: one to monitor the data visually as the experiment progresses and one to record, pre-process and pass the eye tracking data to engagement detection module. The main eye tracker application computed eye physiological indices (PI) such as pupil diameter (PD) and blink rate (BR) and behavioral indices (BI) [14] such as fixation duration (FD) from raw gaze data. The FD is correlated with attention on a specific region of visual stimuli whereas the eye physiological indices PD and BR are indicative of sensitivity to emotion recognition and engagement [14,22-24]. For each data point, gaze coordinates (X, Y), PD, BR, and FD were computed and logged together with the whole raw data, trial markers and timestamps in addition to being used as features for the rule-based engagement detection mechanism. The fixation duration computation was based on the velocity threshold identification (I-VT) algorithm [25]. We chose the I-VT algorithm for its robustness and simplicity. The algorithm sets a velocity threshold to classify gaze points into saccade and fixation points. Generally, fixation points are characterized by low velocities (e.g.: < 100 deg/sec) [25]. We used 35 pixels per sample (~ 60 deg/sec) as our velocity threshold.

2.5 Adaptive Multimodal Interaction

In the actual adaptive interaction two models for physiological affective recognition and rule based engagement detection are combined to give decision of the next level of interaction.

Physiology-Based Affect Modeling

First, the signals were filtered to reject outliers and artifacts and smooth the data. Then, individual baseline mean was subtracted from the data to remove effects of individual variations. For ECG and PPG, peaks were detected after baseline wander removal following the artifact removal. Finally, several features were extracted from the signals.

Feature Extraction

From the 8 channels of physiological signals collected, a total of 51 features were extracted. These features were chosen because of their correlation with engagement and emotion recognition process as noted in psychophysiology literature [18,21,10]. For example, cardiovascular activities such as inter-bit interval (IBI) represents the rate at which the cardiovascular activity changes and can be used to distinguish arousal levels of an emotion. Electrodermal activity as measured via GSR is indicative of response to external stimuli that might make the subject tense or anxious. The pulse transit time (PTT) is a measure of the time the blood takes to travel from the heart to the finger tips. This specific feature was computed using the peaks of both ECG and PPG signals. The features are normalized before being fed to the dimensionality reduction and supervised classification algorithms.

Supervised Training

The extracted features are mapped to a lower dimensional space using principal component analysis (PCA). In a separate study, training data for supervised classification were recorded from 10 subjects with ASD as well as 10 typically developing subjects. These data were recorded while the subjects were playing computer games (Pong and Anagram) with various levels of difficulty that were designed to induce a range of affective states including liking, engagement, frustration, and boredom. They were made to play the games for sufficient amount of time to collect enough training data. After the data collection, these data were trained using support vector machine (SVM) and artificial neural network (ANN) classification learners in a comparative study. Both classification methods resulted in close classification accuracies from upwards of 80 percent to nearly 90 percent. This is considered a high classification accuracy for physiological signals based affective modelling due to the challenging nature of affect detection. Affect detection is challenging due to various factors including daily and individual variations of the signals themselves. Combined effect of various emotions expressed in a single channel also makes it difficult to attribute and learn the effect of a single emotion or affective state separately.

Engagement Modeling

A rule-based system for engagement detection is developed to infer engagement using the behavioral as well as physiological indices from the tracking data as features. The rules use adaptive thresholds and these thresholds are personalized using average baseline values recorded before interaction.

3 Current Status

As described in Section 2, this system is an ongoing development effort which is tested for usability incrementally. The first phase of the system was developing the virtual environment, the characters and endowing facial emotional expressions to the characters. This stage was evaluated with 10 children with ASD and 10 typically developing children in a separate study. After that, this current study, develops more capabilities such as various animations, the cafeteria environment, the speech-based dialog management, affective modeling with supervised training methods and eye tracking based engagement modeling. This section briefly presents the current status of these components of the system and the preparation to the usability study of this multimodal adaptive interactive virtual reality environment.

3.1 System Design and Development Status

The VR social task environment development is completed. All characters now have all the lip-synced speech capabilities, several animations and the environment is populated with enough avatars to undertake the conversational missions (levels). Eventually the system will have both conversational and non-conversational (non-verbal) missions that would capture the social task components depicted in Fig. 1 Right. As described in Section 2.2, the current spoken dialog management system uses already pre-defined conversational threads for each mission. We chose pre-defined conversational threads, at least for this development cycle, in order to keep this mechanism tractable while we could assess the reliability of the avatars and feedback mechanism. The threads were designed by ASD therapists at Vanderbilt University Kennedy Center. We developed an application that automatically created XML grammar files for the speech recognition engine and serializes the conversation thread trees once the therapist created the thread intuitively using the application. We are currently experimenting with probabilistic question and answer dialog management with real natural language processing capabilities. However, this task needs a lot of vocabulary and speech corpus for training to achieve the level of accuracy that is sought for this task.

We have conducted a separate study to collect training physiological data for affect modeling as described in Section 2.5. We are currently training and comparatively evaluating SVM and NN performances for this specific modeling as part of that separate study. Currently, we are also exploring and comparatively studying a rule-based engagement detection and Bayesian belief networks (BBN) based engagement modeling. Nasoz et al. [26] used BBN for modeling emotions as part of driver safety measures.

3.2 Usability Study

We have obtained IRB approval for a usability study with the system. The usability study will evaluate how adolescents with ASD interact with the system. We are in the process of participant recruitment for the study.

4 Conclusion and Future Direction

The main contribution of this work is to present the development of a realistic multimodal VR-based social interaction platform that can be used for ASD intervention. The uniqueness of this platform relies on its ability to gather objective eye gaze and physiology data while a participant is engaged in a closed-loop VR-based adaptive social interaction. The system design and development is complete at this time for this iteration. The system is tested for correct interaction and the components such as the conversational dialog manager are being tested extensively. The results of the usability study for the conversational mission iteration of the system will be available in the near future.

Acknowledgement. This work was supported in part by National Science Foundation Grant [award number 0967170] and National Institute of Health Grant [award number 1R01MH091102-01A1].

References

1. Lord, C., Volkmar, F., Lombroso, P.J.: Genetics of childhood disorders: XLII. Autism, part 1: Diagnosis and assessment in autistic spectrum disorders. Journal of the American Academy of Child and Adolescent Psychiatry 41(9), 1134 (2002)
2. Diagnostic and Statistical Manual of Mental Disorders: Quick reference to the diagnostic criteria from DSM-IV-TR. American Psychiatric Association, Amer. Psychiatric Pub. Incorporated, Washington, DC (2000)
3. McEvoy, R.E., Rogers, S.J., Pennington, B.F.: Executive function and social communication deficits in young autistic children. Journal of Child Psychology and Psychiatry 34(4), 563–578 (2006)
4. Demopoulos, C., Hopkins, J., Davis, A.: A Comparison of Social Cognitive Profiles in children with Autism Spectrum Disorders and Attention-Deficit/Hyperactivity Disorder: A Matter of Quantitative but not Qualitative Difference? Journal of Autism and Developmental Disorders, 1–14 (2012)
5. Ploog, B.O., Scharf, A., Nelson, D., Brooks, P.J.: Use of Computer-Assisted Technologies (CAT) to Enhance Social, Communicative, and Language Development in Children with Autism Spectrum Disorders. Journal of Autism and Developmental Disorders, 1–22 (2012)
6. Gal, E., Bauminger, N., Goren-Bar, D., Pianesi, F., Stock, O., Zancanaro, M., Weiss, P.L.: Enhancing social communication of children with high-functioning autism through a co-located interface. AI & Society 24(1), 75–84 (2009)
7. Ganz, M.L.: The lifetime distribution of the incremental societal costs of autism. Archives of Pediatrics and Adolescent Medicine, Am. Med. Assoc. 161(4), 343–349 (2007)
8. Bernard-Opitz, V., Sriram, N., Nakhoda-Sapuan, S.: Enhancing social problem solving in children with autism and normal children through computer-assisted instruction. Journal of Autism and Developmental Disorders 31(4), 377–384 (2001)
9. Parsons, S., Mitchell, P., Leonard, A.: The use and understanding of virtual environments by adolescents with autistic spectrum disorders. Journal of Autism and Developmental Disorders 34(4), 449–466 (2004)
10. Welch, K.C., Lahiri, U., Liu, C., Weller, R., Sarkar, N., Warren, Z.: An affect-sensitive social interaction paradigm utilizing virtual reality environments for autism intervention. In: Jacko, J.A. (ed.) Human-Computer Interaction, Part III, HCII 2009. LNCS, vol. 5612, pp. 703–712. Springer, Heidelberg (2009)
11. Parsons, S., Mitchell, P.: The potential of virtual reality in social skills training for people with autistic spectrum disorders. Journal of Intellectual Disability Research 46(5), 430–443 (2002)
12. Parsons, T.D., Rizzo, A.A., Rogers, S., York, P.: Virtual reality in paediatric rehabilitation: A review. Developmental Neurorehabilitation 12(4), 224–238 (2009)
13. Kandalaft, M.R., Didehbani, N., Krawczyk, D.C., Allen, T.T., Chapman, S.B.: Virtual Reality Social Cognition Training for Young Adults with High-Functioning Autism. Journal of Autism and Developmental Disorders, 1–11 (2012)
14. Lahiri, U., Warren, Z., Sarkar, N.: Design of a Gaze-Sensitive Virtual Social Interactive System for Children With Autism. IEEE Transactions on Neural Systems and Rehabilitation Engineering (99), 1 (2012)

15. Kenny, P., Parsons, T.D., Gratch, J., Leuski, A., Rizzo, A.A.: Virtual patients for clinical therapist skills training. In: Pelachaud, C., Martin, J.-C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) IVA 2007. LNCS (LNAI), vol. 4722, pp. 197–210. Springer, Heidelberg (2007)
16. Leuski, A., Patel, R., Traum, D., Kennedy, B.: Building effective question answering characters. In: Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue, pp. 18–27. Association for Computational Linguistics (2009)
17. Vinciarelli, A., Pantic, M., Bourlard, H.: Social signal processing: Survey of an emerging domain. Image and Vision Computing 27(12), 1743–1759 (2009)
18. Liu, C., Conn, K., Sarkar, N., Stone, W.: Online affect detection and robot behavior adaptation for intervention of children with autism. IEEE Transactions on Robotics 24(4), 883–896 (2008)
19. Picard, R.W.: Future affective technology for autism and emotion communication. Philosophical Transactions of the Royal Society B: Biological Sciences 364(1535), 3575–3584 (2009)
20. Groden, J., Goodwin, M.S., Baron, M.G., Groden, G., Velicer, W.F., Lipsitt, L.P., Hofmann, S.G., Plummer, B.: Assessing cardiovascular responses to stressors in individuals with autism spectrum disorders. Focus on Autism and Other Developmental Disabilities 20(4), 244–252 (2005)
21. Cacioppo, J.T., Tassinary, L.G., Berntson, G.G.: Handbook of psychophysiology. Cambridge Univ. Pr. (2007)
22. Anderson, C.J., Colombo, J., Shaddy, D.J.: Visual scanning and pupillary responses in young children with autism spectrum disorder. Journal of Clinical and Experimental Neuropsychology 28(7), 1238–1256 (2006)
23. Hsiao, J.H., Cottrell, G.: Two fixations suffice in face recognition. Psychological Science 19(10), 998–1006 (2008)
24. Libby Jr, W.L., Lacey, B.C., Lacey, J.I.: Pupillary and cardiac activity during visual attention. Psychophysiology 10(3), 270–294 (1973)
25. Salvucci, D.D., Goldberg, J.H.: Identifying fixations and saccades in eye-tracking protocols. Paper Presented at the Proceedings of the 2000 Symposium on Eye Tracking Research & Applications (2000)
26. Nasoz, F., Lisetti, C.L.: Affective user modeling for adaptive intelligent user interfaces. In: Jacko, J.A. (ed.) Human-Computer Interaction, Part III, HCII 2007. LNCS, vol. 4552, pp. 421–430. Springer, Heidelberg (2007)