Evaluating a minimally invasive laboratory architecture for recording multimodal conversational data

Spyros Kousidis[1], *Thies Pfeiffer*[2], *Zofia Malisz*[3], *Petra Wagner*[3], *David Schlangen*[1]

[1]Dialogue Systems Group, [2]A.I. Group, Faculty of Technology, [3]Phonetics and Phonology Group
[1,2,3] Bielefeld University, Germany
spyros.kousidis@uni-bielfeld.de, tpfeiffe@techfak.uni-bielfeld.de

Abstract

This paper presents ongoing work on the design, deployment and evaluation of a multimodal data acquisition architecture which utilises minimally invasive motion, head, eye and gaze tracking alongside high-quality audiovisual recording of human interactions. The different data streams are centrally collected and visualised at a single point and in real time by means of integration in a virtual reality (VR) environment. The overall aim of this endeavour is the implementation of a multimodal data acquisition facility for the purpose of studying non-verbal phenomena such as feedback gestures, hand and pointing gestures and multi-modal alignment. In the first part of this work that is described here, a series of tests were performed in order to evaluate the feasibility of tracking feedback head gestures using the proposed architecture.

Index Terms: Multimodal interaction, feedback, virtual reality

1. Introduction

The acquisition of annotated multimodal conversational data is nowadays considered essential for the better understanding of human discourse [1], but also in the context of interaction between humans and ECAs [2]. However, scientific interest in multimodal corpora extends beyond computational linguistics into the fields of behavioural and social sciences, while the problems that arise in constructing, maintaining and reusing such databases have become the subject of research in computer science [3].

Two major issues that often arise when designing multimodal corpora are the inhibition of natural discourse by the presence of sensory equipment (a problem that also exists in traditional, audio-only corpora [4]), and the lack of standardisation in storing, annotating and querying the data. In addition, the use of visual data is also non-standard, as the angle and zoom of the camera(s) are often chosen to serve specific purposes thus limiting re-usability of the content. Finally, annotation of the additional signal streams is costly, often limiting the size of the corpus and introducing compromises that may also limit the usefulness of the acquired content [3].

The data collection architecture described here addresses these issues by using minimally invasive motion tracking sensors and a VR environment which is used both as a collection point of all sensory data, as well as an additional annotation tool. The purpose of this work is to collect multimodal conversational data in order to study various interaction phenomena. One type of non-verbal behaviour that is of particular interest are visual feedback gestures which are deemed essential in interaction management, complementary to spoken feedback dialogue acts, such as backchannels [2]. Visual feedback gestures include eye and head movements, facial expressions, hand gestures and body posture [5]. The ability to automatically detect and model such gestures is highly desirable in ECA design [6]. Another planned use is the study of alignment between interlocutors, which has previously been studied in a number of modalities, including posture and gaze [7]. However, few studies have looked at more than one modality at a time (e.g. [8]), perhaps due to the lack of sufficient data aggregation and synchronisation. The proposed architecture also addresses this issue by exploiting the immersive capabilities of VR.

Technology for the real-time assessment of multimodal human actions has long been a corner-stone of VR research. Together with the capabilities of simulating cognitive models of communication in virtual agents, the combination of VR and linguistic research is very promising [9]. In previous work, assessment of human pointing behaviour has been achieved through the implementation of an experimental-simulative loop using VR technology with the tool IADE [10]. A study on human-human interactions, in which both participants' gestures and speech were tracked [11] was re-simulated in VR in order to aggregate and review all collected and annotated data in one place. The use of VR technology allowed experimenters and annotators to immerse into the recorded setting and to be situated right within the original interaction context. Later work also included the tracking of gaze and the real-time identification of the objects of interest [12]. Although tracking technology has often proved inhibiting to natural behaviour on the part of participants [13] in the past, technology has recently become less obtrusive, and remote sensing capabilities for eye gaze and 3D gestures are commercially available. The following section describes our data collection architecture which utilises these technologies in order to capture visual feedback gestures.

2. Data collection architecture

The laboratory setup is shown in Figure 1. The data stream from each sensor is independently transmitted to the VR environment via LAN. This allows immersive viewing of the recorded interaction from any angle, including a real-time updating first perspective view of tracked subjects. Logging is also performed at this central point, ensuring synchronisation of the sensory components described below.

2.1 Motion tracking

Motion tracking is performed by the Microsoft Kinect[1], an interaction device based on a *depth camera* produced by PrimeSense[2]. The Kinect does not require any attachments to the tracked subject, but projects a structured light and then uses its distortion to create a depth image. As a second step, the provided software frameworks, Microsoft Kinect SDK[1] or

[1] MSDN 2010 Microsoft Kinect SDK http://www.microsoft.com/en-us/kinectforwindows/

[2] PrimeSense Ltd, http://www.primesense.com/

OpenNI[3], extract skeletal information. This skeleton model is still rather coarse and does not contain fingers or the orientation of the head. This technology is quite novel, so more precise versions of Kinect-like systems and better software frameworks for skeleton extraction can be expected.

2.2 Head/eye/gaze tracking

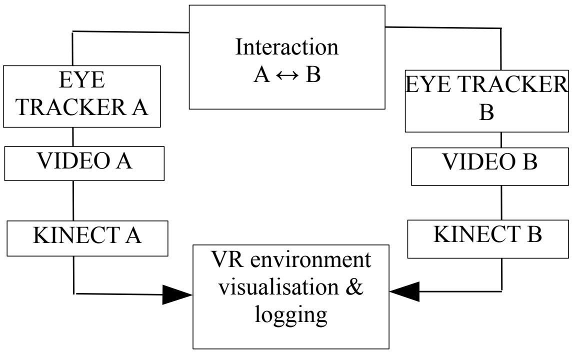

Figure 1: Schematic of data collection architecture

Head, eye and gaze tracking is performed using Facelab 5 by Seeingmachines[4] which is a set of two (or more) eye-tracking cameras and an infrared light that is projected onto the face. Facelab uses the reflection of this light to track the position and orientation of the subject's head, the direction of gaze, the motion of several facial features, and several derived measurements such as the percentage of eye closure, fatigue, blinks, or the vergence point of the two gaze vectors of the subject's eyes. Additional components such as zoom lenses allow for a number of different positioning configurations, moving the cameras away from the tracked subject.

2.3 Audiovisual recording

Traditional audiovisual recording is performed by a set of three synchronized Canon XHG1S HD cameras and a choice of either directional or close-contact Sennheiser microphones. For a dyadic interaction, two cameras provide close-up front views of the interactants, while the third camera provides a panoramic view of the interaction.

2.4 Virtual Reality

The different devices are connected using InstantReality[5], a VR framework and the underlying InstantIO network-transparent technology. Specific implementations of InstantIO modules for the Kinect and FaceLab were developed, along with an XML-based data format to log the events from all connected devices in an integrated fashion. The logging process is managed by a custom-built software tool, which is part of an effort to create a complete, publicly available tool chain for manual and semi-automatic recording and annotation of multimodal experiments.

3. Evaluation test procedures

The evaluation plan of the system consisted of two parts: in the first part – described in this section – a number of procedures were designed in order to acquire gold standard data against which to assess the accuracy of the tracking sensors. The second part – described in the next section – comprised data acquisition in real conditions. All procedures were performed by one female (F) and two male (M1, M2) subjects.

3.1 Head position and orientation accuracy

A laser-pointing device with a precision of ±2 mm was used to measure distances from a person's head to flat panels placed around the person. Rotations of the head both in up-down (similar to nodding) and side-ways (tilting) directions were measured with a pitch-angle measuring device with a precision of 1 degree. The devices were fixed on a lightweight helmet that could be firmly strapped to the person's head. With the assistance of a lab technician, subjects moved or rotated their head in the tracked 3D space and measurements from both the laser-pointing device and the eye-tracking sensor were taken at 36 random points for each subject. Left-to-right rotation (yaw) angles were inferred using the difference in distance from the subject's head to a flat panel in front of them, as the head rotates.

3.2 Head position tracking range

The eye-tracker allows the subject to be seated in a range of distances from the tracking cameras, depending on the configuration of zoom lens provided. This distance, theoretically at least, has an effect on the range of movements tracked persons can perform before they move out of range and tracking is lost. In order to measure this range, subjects were instructed to perform movements around the tracked space, reaching the limits in each direction. Each of the three subjects was placed at three different distances from the tracker: near (~.75m), mid (~.90m), and far (~1.05m). These positions represent the extremes and mid-point of distances allowed by the focus calibration range of the tracking cameras.

3.3 Gazed object detection

The gaze-tracking function of the eye-tracking sensor, combined with a VR model of objects in the subject's field-of-view allows for detection of the object the subject is gazing at. Because this detection is based on whether a 'gaze vector' coming out of the subject's eyes intersects the modelled objects, the distance of the person from the sensors can theoretically affect accuracy. As in the previous procedure, subjects were placed at three different distances, while gazing at a set of five 40x40mm coloured cubes that were fixed on a flat table. A score (0-3) was given for each object, depending on whether the gaze vector pointed to the modelled object itself or one of a set of progressively larger proxy objects at the same location in the VR model. An overall success rate was calculated as a percentage of acquired points over the maximum possible points for all objects combined.

3.4 Body limb and hand position accuracy

The motion tracking sensor precision was measured by placing subjects in three predefined positions marked on the lab floor. A snapshot of the skeleton tracking data was taken at each position. The procedure was repeated three times, moving the motion sensor to a new position each time. The accuracy was assessed by comparing the calculated distances between the three positions. A similar procedure was followed for detecting whether subjects were holding a specific object. Three 80mm-diameter spheres were positioned in the tracked area and subjects were asked to hold them in their hands. Again, a comparison between the actual and the calculated distance between the spheres (derived from the tracked hand positions) yields a measure of the tracking accuracy.

[3] OpenNI, http://www.openni.org/

[4] seeingmachines, http://www.seeingmachines.com/product/facelab/

[5] IGD Fraunhofer, 2010, Instant reality, http://www.instantreality.org

4. Tracking feedback gestures

A pilot experiment was performed in order to evaluate the ability of the architecture to track non-verbal feedback gestures. Three subjects, one female and two male (different from the three in the previous section), participated in this test. The setting that was used for feedback gesture elicitation is a simplified version of the one used in [14]. Briefly, one of the subjects (speaker) is asked to narrate a story from their own experience, such as a holiday story while the other subject (listener) is encouraged to actively listen, paying attention to detail, with a hint that they might be asked questions at the end of the narration.

In each interactive session, both participants were video-taped and recorded with contact microphones, while the listener was also tracked with the eye-tracking sensor (the motion-capture sensor was not used in this test). The data from the sensor was read real-time into the VR model and logged using the logging software described in section 2.4. Each of the three sessions lasted for about 10 mins. The annotation of the head gestures was performed by two expert annotators following the schema in [14] using the video only (audio was muted). The annotation labels consist of a *gesture category* (nod, tilt, jerk, shake etc) and the *number of cycles*, i.e. the number of repetitions of the gesture. For example, a 'nod-2' denotes a nod gesture with two cycles (a “double nod”). In total, 210 gestures were found and visually compared to the tracking data, as a first assessment of the detail captured with the described method.

5. Results and discussion

Table 1 shows the results from the accuracy evaluation procedures. For head tracking, the highest accuracy is acquired for the Z axis (towards or away from the tracking cameras) and the lowest for the Y axis (raising-lowering head). The difference in error is quite large, however this maybe attributed to the fact that moving the subject's head without simultaneously rotating it is progressively more difficult in the reverse order of that of the error magnitudes (Z, X, Y).

Subject		Male 1	Female	Male 2	All
Head tracking Position Accuracy (cm)	X	2.15	2.28	1.16	1.86
	Y	3.65	2.75	2.22	2.87
	Z	0.15	2.4	0.15	0.9
Head tracking Orientation Accuracy (DEG)	Pitch	1.87	1.96	1.71	1.85
	Yaw	2.75	3.62	3.41	3.26
	Roll	0.78	1.76	1.34	1.29
Motion tracking position accuracy (cm)		1.37	2.83	3.12	2.44
Motion tracking object holding (cm)		3.16	5.4	3.4	3.99

Table 1: Mean absolute error of tracking sensors

Similarly, an error margin of ~2° is common for 2 of three angles (pitch and roll) which were measured with the pitch-angle measuring device, while the left-right head-direction angle (heading) which was inferred rather than directly measured shows a higher error margin (3.26°). Thus, a value of ~2° is a better estimate of the error margin for the angles. These errors are larger than those specified by the equipment vendor (positional and angular accuracy of ±1 mm and ±1° respectively). The difference is most likely due to the fact that the realtime data stream was read from the tracker instead of the more accurate one that comes with a latency of 2.5 seconds.

Figure 2 shows the effect of distance of the subject from the eye-tracking sensor on the tracking range. As predicted, this yields an increase of the effective tracking space, allowing for more freedom in subject movements: the “far” position yields a 25% larger tracking space compared to the “near” position (the tracking range is computed by adding the three dimensions together). There is a similar 20% increase (“far” vs “near”) in the range of head rotations (not shown).

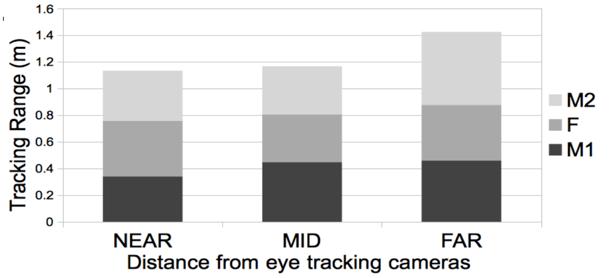

Figure 2: Effect of distance on head-tracking range

On the other hand, the distance between subject and eye-tracker does not have an obvious effect on gazed object detection accuracy (see Figure 3). The effect is balanced by the fact that a sharper angle is required to gaze at the objects at the near position compared to the far position. Objects with an area of at least 6x6cm facing the viewer at a distance of 1m can be consistently detected.

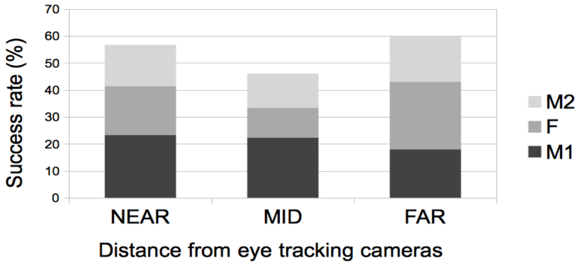

Figure 3: Effect of distance on gazed object detection

The motion-capture sensor yielded a comparable error margin of ±2.44 cm (see table 1) when comparing the positions of ankles or shoulders in a standing posture, while the position of the hands showed a larger error margin (±3.98 cm). This position is not well-defined (a hand can hold an object in various ways) and therefore and may have differed significantly between subjects. Applications of detecting the position of a subject's hand in real time can be either hand gesture detection or monitoring whether a subject's gaze follows a displayed object (by combining motion and eye-tracking data). The results suggest that this is feasible provided that the objects are large enough to ensure a high success rate for gazed object detection and allow for hand position error. The error margins for the Kinect sensor are larger than those reported in [15], but the latter reported using special apparatus to hold subjects in place, while the focus of the work reported here has been more towards “real” conditions with naïve, unconstrained subjects.

Tracking during the pilot experiment proved reliably robust, as the signal from the eye-tracker had no break-ups in the first two sessions, while tracking was lost 3.27% of the time in the third session. This was mainly due to a bad position of the subject relative to the tracker, causing his head to move out of range during a few extreme movements. A similar result has been reported for the Kinect sensor [15], i.e. tracking is never lost unless subjects move out of range.

The results from the pilot experiment also suggest that successful head gesture detection may be expected. Figure 4 shows head tracking data (approximately 0.8 seconds) for which the corresponding video was labeled as a nod with two cycles. These can clearly be discerned in the plot at 500 and 800ms, where the slope of the change in pitch angle is the steepest. Importantly, this nod is visually very subtle according to the annotators.

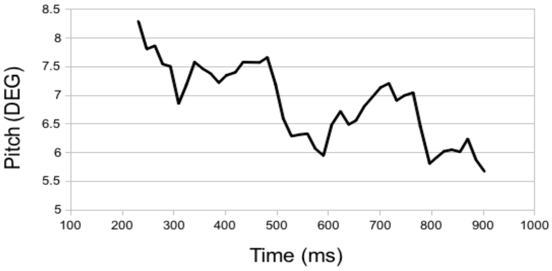

Figure 4: Nodding gesture (2 cycles)

Similarly, the interval (2 sec) shown in Figure 5 was annotated as a complex gesture (nod + 2-cycle shake). Again, the plot shows a 'dip' around 900ms on the top line (pitch) which corresponds to the nod, and two more dips at 1600 and 1800 ms on the bottom line (yaw) which are the left-to-right head shakes. The simultaneous dip at the top line at the time of the second shake shows that the head rotated diagonally both down and to the right.

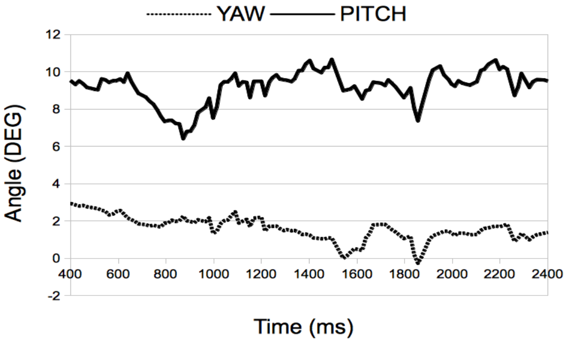

Figure 5: Nod and subsequent shake (2-cycles) head

It would be possible to use the VR environment to playback the tracked head movements on a 3D-head, offering advantages such as zooming/navigating freely around it and thus providing more view angles in comparison to traditional annotation using video image. This was not performed, because the logging tool is currently work in progress and has limited playback/seeking capabilities. However, this type of functionality is an expected outcome of the work described here. Another planned improvement is integration of the VR environment with widely-used annotation tools such as ELAN [16]. Finally, further improvements are expected in the upcoming releases of the MS Kinect SDK[1], with more detailed skeleton and head tracking, making this sensor even more attractive to use due to its low cost and minimum disturbance to the interaction setting.

6. Conclusions & Future work

We have presented a multimodal data collection architecture that uses minimally invasive sensors and virtual reality as a central connecting point of the data streams, opening possibilities to study multimodal phenomena in a well-designed environment. In an evaluation of the setup both under ideal and under realistic conditions we found the accuracy of the collected data to be adequate for capturing multimodal behaviour such as feedback head gestures. Our aim is to further explore the combination of the immersive capabilities of VR with real time motion-tracking data, towards a fully integrated multimodal annotation environment. These tools will eventually be released under an open source license.

7. Acknowledgements

This research is partly supported by the Deutsche Forschungsgemeinschaft (DFG) in the CRC 673 "Alignment in Communication". The authors would like to thank Felix Hülsmann, Florian Hofmann, Joanna Skubisz and Michael Bartholdt.

8. References

[1] P. Paggio, J. Allwood, E. Ahlsen, K. Jokinen, and C. Navarretta, "The NOMCO multimodal Nordic resource - goals and characteristics," , LREC 2010, Valetta, Malta, 2010.

[2] M. Boholm and J. Allwood, "Repeated head movements, their function and relation to speech," LREC 2010, Valletta, Malta, 2010.

[3] D. Knight, "The future of multimodal corpora," Revista Brasileira de Linguistica Aplicada, vol. 11, pp. 391-415, 2011.

[4] N. Campbell, "Databases of emotional speech," in ISCA Workshop on Speech and Emotion, Northern Ireland, 2000, pp. 34–38.

[5] K. Jokinen, "Gaze and Gesture Activity in Communication Universal Access in Human-Computer Interaction. Intelligent and Ubiquitous Interaction Environments.", C. Stephanidis, Ed., Springer Berlin / Heidelberg, 2009, pp. 537-546.

[6] I. Mlakar and M. Rojc, "Towards ECA's Animation of Expressive Complex Behaviour Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues," A. Esposito, A. Vinciarelli, K. Vicsi, C. Pelachaud, and A. Nijholt, Eds., ed: Springer Berlin / Heidelberg, 2011, pp. 185-198.

[7] D. C. Richardson, R. Dale, and K. Shockley, "Synchrony and swing in conversation: Coordination, temporal dynamics, and communication," in Embodied Communication, G. Knoblich, Ed., ed: Oxford University Press, 2008.

[8] N. Campbell, "An Audio-Visual Approach to Measuring Discourse Synchrony in Multimodal Conversation Data,", Interspeech 2009, Brighton ,UK, 2009.

[9] T. Pfeiffer, "Using virtual reality technology in linguistic research," IEEE Virtual Reality Workshops (VR), 2012, pp. 83-84.

[10] T. Pfeiffer, A. Kranstedt, and A. Lücking, "Sprach-Gestik Experimente mit IADE , dem Interactive Augmented Data Explorer," in Dritter Workshop Virtuelle und Erweiterte Realität der GIFachgruppe VRAR, 2006, pp. 61--72.

[11] A. Kranstedt, A. Lücking, T. Pfeiffer, H. Rieser, and I. Wachsmuth, "Deictic object reference in task-oriented dialogue," in Situated Communication, ed Berlin: Mouton de Gruyter, 2006, pp. 155-207.

[12] T. Pfeiffer, "Understanding multimodal deixis with gaze and gesture in conversational interfaces," PhD, Bielefeld, Bielefeld, 2010.

[13] K. Jokinen, M. Nishida, and S. Yamamoto, "Eye-gaze experiments for conversation monitoring,", 3rd International Universal Communication Symposium, Tokyo, Japan, 2009.

[14] M. Włodarczak, H. Buschmeier, Z. Malisz, S. Kopp, and P. Wagner, "Listener head gestures and verbal feedback expressions in a distraction task,", Submitted, 2012.

[15] M. A. Livingston, J. Sebastian, Z. Ai, and J. W. Decker, "Performance Measurements for the Microsoft Kinect Skeleton,", IEEE Virtual Reality, Orange County, CA, USA, 2012.

[16] H. Brugman, A. Russel, and X. Nijmegen, "Annotating multi-media / multimodal resources with ELAN",LREC 2004,Lisbon, Portugal, 2004, pp. 2065—2068.