MODELLING USER ATTENTION FOR HUMAN-AGENT INTERACTION

Christopher Peters — Coventry University, Coventry CV1 5FB, United Kingdom

Stylianos Asteriadis — National Technical University, Athens, Greece

Genaro Rebolledo-Mendez — Serious Games Institute, Coventry CV1 2TL, United Kingdom

ABSTRACT

In this work, we propose a design for a user attention model featuring three core components. Our system components can work in real-time, offering indications of user attention from different sensory inputs (both visual and neurophysiological). Intention of the current work is to keep the equipment as unintrusive as possible, while keeping the confidence of the inputs as high as possible. We discuss potential applications of such a system, particularly with respect to evaluating user attentive behaviour during human-agent interactions and as a more natural interface for interacting with agents.

1. INTRODUCTION

In this paper, we propose a system for modelling user attention during real-time human-agent interactions. The three components of the proposed work consist of a gaze detector based on standard web-camera input, detection of attention using neurophysiological signals, and a user attention module for interpreting and representing attention paid by a user to a scene over a time period, representing where, to what degree and when the user is and has been attending to the scene. The user attention model has the potential be used for enhancing interactions with an Embodied Conversational Agent, or ECA, in two ways: first of all, for evaluating human behaviour during human-agent interactions, and secondly for acting as an interface enabling interactions.

Following a description of background work in Section 2, more details of each of the three core components and an overview of the integrated system are presented in Section 3. We describe a prototype scenario in Section 4 and discuss possible implications of the use of these components for HCI in Section 5.

2. BACKGROUND

Brain Computer Interfaces (BCI) constitute an emerging field of research. This paper considers sensor-based BCI that measure and interpret brain activity as a means to interact with computers. Sophisticated BCI, such as MEG, NIRS, PET and fMRI, have been identified as unsuitable for real–time, human-computer interaction [1] due to their intrusiveness and bulkiness. Portable, real-time, bio-recording BCI (mainly EEG-based technology) have become a topic of research interest both as a means for obtaining user input and studying responses to stimuli. For example, EEG biofeedback is being investigated as a tool to aid individuals with learning difficulties [2]. Another approach consists of identifying brain activation patterns associated to functional tasks, such as walking or moving an arm, to translate them into computer commands. This approach has the potential to provide communication mechanisms to control interactions in virtual environments [3] or wheelchair control for individuals with a degree of paralyses [4].

Numerous approaches have been proposed for the estimation of head pose and eye gaze, requiring different types of equipment. The method reported in [5], for example, uses specialised hardware for estimating the focus of attention. Employing pure image/video processing techniques, based on single-camera systems, without the need for calibration tends to be easier to develop, but such systems require more constraints in terms of lighting and user behaviour for robust use. A typical image analysis based technique is reported in [6].

In work concerning eye-gaze for interaction with conversational agents, some approaches have cast the ECA in the primarily role of a listener, for example, as a SAL, or sensitive artificial listener [7], providing feedback to a user. Attentive presentation agents [8] use eye gaze to alter the ongoing behaviour of embodied agents in real-time. In a similar vein, [9] and [10] have been investigating systems for estimating user engagement based on gaze behaviour during interaction with a conversational agent. Unlike previous work, we are considering the use of neurophysiological measurements of attention in conjunction with gaze behaviour for managing engagement during interaction scenarios.

3. CORE COMPONENTS

The three core components consist of a gaze detector (Section 3.1), neurophysiological detection (Section 3.2)

and a user attention representation module for storing, integrating and interpreting attention paid by a user to a scene over a time period (Section 3.3).

3.1. Gaze Behaviour

This is concerned with the detection and tracking of facial feature points in order to establish the user's gaze direction, encoded as head pose and eye gaze vectors (see Fig. 1).

3.1.1. Facial Points Detection and Tracking

Our method for detecting user gaze relies on video processing techniques, using a variant of the method described in [11]. A web-camera is the only specialised hardware necessary for the system to operate. At start-up, it is necessary to detect the user's face and facial features in order to proceed with tracking and further analysis. Apart from the facial points detected in [11], our system also detects the upper and lower point of the mouth, as well as the point between the nostrils and two points on each eyebrow. Eyebrows and nostrils are detected as the darkest regions in a neighbourhood defined by features already detected: For eyebrow detection, an area is expanded above each eye corner, and the vertical projection of its luminance values is calculated. The position of each eyebrow point is considered as the row that corresponds to the lowest projection on the vertical axis. The vertical position of the nostrils is found in a similar way. A three-pyramid Lukas-Kanade algorithm is employed for tracking the detected facial features, while models of natural human motion allow recovery from erroneous tracking, as will be discussed later in this Section.

3.1.2. Gaze Direction Detection

Head pose is estimated using the translation of the point in the middle of the inter-ocular line. The fraction of the inter-ocular distance with the vertical distance between eyes and mouth is monitored. If maintained within certain limits, the translation of the point between the eyes is considered to be due to translational movements of the head, and not rotational. When the above differs significantly with respect to the frontal pose, the translation of the point between the eyes is calculated in pixels to its position when the user was turned frontally. This translation, normalised with the inter-ocular distance at start-up to cater for scale variations, is the *head pose vector*. In cases where one of the two eye trackers provides incorrect information due to large head rotations, the head pose vector reduces in length when the user reassumes a frontal position, but does not reach zero. When this occurs, the system must reinitialise in order to detect the face once again: If the user is oriented frontally, the system will restart successfully.

Fig. 1. Gaze direction detection. Here, the white line represents the head pose vector and the black line the eye gaze vector. Black dots are tracked feature points.

Assuming an orthographic projection, all features would be considered to shift in a similar way. Finding outliers from this uniformity and placing them at their expected positions aids in the recovery from erroneous tracking, but requires the use of many features in our algorithm. Eye gaze is calculated based on the translation of the eye centres in relation to the four points defining the eye corners and eyelids. Averaging these measurements for both eyes and normalising with the inter-ocular distance produces the *eye gaze vector*.

Operationally, the system requires the user to face the camera frontally at start-up. The user can then move freely while our video processing techniques track the movements in order to calculate the head pose and eye gaze vectors. The method works in real-time: Feature tracking requires 13msec per frame on average for an input video of resolution 288x352 pixels, using a Pentium 4 CPU running at 2.80GHz. A reinitialisation requires 330ms to resume proper operation when it occurs.

3.2. Neurophysiological Data

Another core component in our toolkit for attention modelling is a neurophysiological detection component. We are currently investigating the use of EEG, using the *NeuroSky* sensor, as a complement to the gaze behaviour component described in Section 3.1. NeuroSky is a non-invasive, dry bio-sensor that detects electrical neuron-triggered activity to determine states of attention and meditation based on the interpretation of alpha and beta brain waves through EEG. NeuroSky operates at 1Hz via a single electrode and signal-processing unit in a headband arrangement. It is used as a headset with three sensors touching the skin at three locations: beneath the ears and the forehead. The electrical signals read at these points are used by Neurosky as inputs to an interpretative network to determine levels of attention and meditation, output as discrete readings on a scale from 0 to 100. The system is capable of distinguishing electrical activity caused by the

neural activity of interest from other sources, such as electromagentic interference in the environment, although care must be taken in the procedure, particularly with mobile phone technologies. There are a number of compelling reasons for investigating the use of Neurosky for HCI: it is a low-cost, easy-to-use EEC reader developed for the leisure industry. The principle advantage is its unobtrusive nature, providing the potential to conduct accurate user studies in more practical and naturalistic settings without inducing the stress or distractions of more elaborate scanning processes. However, as highlighted in a recent usability study [12], there are tradeoffs: for example, the device provides a much coarser reading of brain activity than multi-electrode EEG or the other BCI technologies mentioned in Section 2. Given that the NeuroSky provides only partial information about attention, we believe it may be effective as a BCI in conjunction with one or more other modalities of detection.

3.3. Attention Representation

In contrast to the previous components, the attention representation component acts to store, integrate and interpret where, when and to what degree the user is and has been attending to the scene. In terms of gaze behaviour (see Section 3.1), this component uses either the eye direction or the head direction to establish the screen-space coordinates of where the user is looking. Special coordinates are used to signal that the user is looking outside of the screen area, detailing which screen border they are looking outside. For example, *top* signifies that the user is looking above the top of the screen, and thus outside the valid screen area. We envisage Neurosky output (see Section 3.2) to be integrated with gaze behaviour derived from the gaze detector in order to provide additional measurements of the degree to which attention is being paid. It is hoped this data will enhance spatial and temporal information relating to the user's detected gaze behaviour, to provide more information about the possible depth of processing in addition to what is being processed.

3.3.1. Virtual Attention Objects and Level of Attention

Since metrics based on user gaze configuration during a single frame are highly unreliable indicators of attention, we define a *level of attention* metric. It refers to a clustering of a user's focus of interest in a single region over multiple frames. In order to simplify the analysis of what is being looked at, we split the scene into multiple *virtual attention objects*, or VAO's. A single VAO is attached to each object for which we wish to record attention information. If the screen-coordinate of the gaze fixation is located inside a VAO, then its corresponding level of attention is updated to reflect this. Thus, as the user's gaze moves around the screen, each VAO maintains a history of how many times and when the user has fixated it. Over a larger time-frame, and for a specific set of VAO's, the *level of interest* of the user for that set can be computed based on the stored attention levels for each VAO of the specified set.

These are the basic elements with which more complicated attention calculations may be conducted. For example, by defining a set of VAO's containing only those objects currently relevant to the interaction, such as a recently pointed to or discussed object, and comparing the attention paid to these objects with the rest of the scene, we can obtain a measurement of how engaged the user has been with respect to the interaction, rather than to superficial or irrelevant scene details.

4. AGENT INTERACTION SCENARIO

In this Section, we provide a description of a generic scenario template forming the basis of a number of experiments we intend to conduct regarding user attention to an agent during interaction. Each of the aforementioned components (see Section 3) will communicate via a *Psyclone* connection - a blackboard system for use in creating distributed multi-modal A.I. systems. This system is modular and amenable to having components plugged in and out, as necessitated by the scenario requirements or experiment under consideration.

We currently model the agent in the role of speaker and the user as listener. While the agent recites a story, user attention is recorded over the scene and stored in the attention representation described in Section 3.3. There are a number of possible ways in which the representation can then be used: Firstly, it can be used solely for user behaviour analysis. Recordings can be made of user behaviour, agent behaviour and the resultant attention representation in order to study possible correlations between user behaviour and agent behaviour across the scenario. This could also be achieved using gaze-tracking equipment, however our approach may allow the user to behave more freely without the burden of a head-mounted tracking device. Furthermore, in most situations it would not be feasible for the user to wear the gaze-tracker and the Neurosky device at the same time.

Secondly, it can be used by the agent to autonomously decide how to adapt to the state of the user. In the role of speaker, for example, if it is detected that the user is looking outside the screen, thus interpreted as not paying attention to the scenario, then the agent may conduct a behaviour to adapt to this, such as pausing the recital to wait until the user pays attention to it again, or even asking the user to pay attention. In order to achieve this, the agent is provided with access to the information regarding all VAO's in the scene. Since the agent is itself a VAO, it therefore has access to a full assessment of the user's gaze history with respect to the scene and specific objects of attention. In this respect, it could be also be used as an interface, allowing the user to

control and direct the scenario using their attention capabilities. For example, when presented with a number of objects in a scene, if the user may pay attention to a specific object in order for the agent to provide information about it. In this way, interaction becomes more subtle, without the need for explicit requests by the user or agent to be made.

5. DISCUSSION

Real-time analyses of user's behaviour and attention levels during agent-based interactions are an important element of user modelling. The analyses will establish patterns of behaviour and attention allocation useful for endowing autonomous agents with more natural interaction capabilities. The analyses will also serve to study the relationship between visual and cognitive attention. The analyses will help to address important questions: are visually distracted users highly attentive cognitively? Are there categories for attention allocation which are conducive of better results? Do people fall into distinguishable patterns of attention? A key factor towards this direction will be correlating the outputs of the gaze tracker and Neurosky in recorded agent-user interaction scenarios. Findings regarding statistical analysis of the two outputs, especially in highly ambiguous situations (e.g. user's gaze fixated at the screen but with low attention levels) should provide interesting outcomes that will enhance the scenario and help the interaction become more natural. Furthermore, finding relationships between the two components will provide an efficient means for evaluating their individual performance and examine those conditions in which more emphasis should be given to either of the two. This will play a particularly important role in the third component, as fusion will be highly influenced by results from the studies.

6. CONCLUSIONS

We have presented a design for a user attention model featuring three core components. The first two components detect user gaze behaviour based on input from a web-camera and measure attention levels using a neurophysiological recording device, respectively. The third component integrates detected information to provide a model of user attention, representing where, when and to what degree the user is and has been attending to the scene. We will be using this model to define a number of scenarios, based on the generic model described in Section 4, to analyse user behaviour, inform agent behaviour and help manage interaction.

We also intend to consider, at a future stage, the agent in the role of listener. In this role, the agent may provide various types of non-verbal feedback to the user (see for example [7]). In conjunction with speaking role capabilities currently being investigated, this would allow for the construction of more sophisticated scenarios.

7. REFERENCES

[1] S.G. Mason and G.E. Birch, "A general framework for brain computer interface design". *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2003. 11: p. 70-85.

[2] M. Linden, T. Habib and V. Radojevic, "A controlled study of the effects of EEG biofeedback on cognition and behavior of children with attention deficit disorder and learning disabilities". *Applied Psychophysiology and Biofeedback*, 1996. 21 (1).

[3] J.D. Bayliss, "Use of the evoked potential P3 component for control in a virtual apartment". *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2003. 11(2): pp. 113-116.

[4] F. Galán, M. Nuttin, E. Lew, P.W. Ferrez, G. Vanacker, J. Phillips, J. del R. Millán, "A brain-actuated wheelchair: Asynchronous and non-invasive Brain–computer interfaces for continuous control of robots". *Clinical Neurophysiology*, 2008. 129: pp. 2159–2169.

[5] C. Hennessey, B. Noureddin, and P. D. Lawrence, "A single camera eye-gaze tracking system with free head motion," in *Proceedings of the Eye Tracking Research & Application Symposium, (ETRA)*, ACM 2006, pp. 87–94, San Diego, California, USA.

[6] J. Wu and M. M. Trivedi, "A two-stage head pose estimation framework and evaluation," *Pattern Recognition*, vol. 41(3), pp. 1138–1158, 2008.

[7] E. Bevacqua, M. Mancini, and C. Pelachaud. "A listening agent exhibiting variable behaviour", in Proceedings of *Intelligent Virtual Agents*: IVA 2008, volume 5208, Lecture Notes in Computer Science, 2008, pp. 262–269. Springer, Tokyo.

[8] H. Prendinger, T. Eichner, E. André, and M. Ishizuka. "Gaze-based infotainment agents", in Proceedings of the International Conference on Advances *in Computer Entertainment Technology*, ACM International Conference Series, 2007, Vol. 203, pp. 87–90.

[9] R. Ishii and Y. I. Nakano. "Estimating user's conversational engagement based on gaze behaviors", in Proceedings of *Intelligent Virtual Agents*: IVA 2008, volume 5208, Lecture Notes in Computer Science, 2008, pp. 200–207. Springer, Tokyo.

[10] C. Peters, S. Asteriadis, K. Karpouzis, E. de Sevin, "Towards a real-time gaze-based shared attention for a virtual agent", *International Conference on Multimodal Interfaces* (ICMI), Workshop on Affective Interaction in Natural Environments, AFFINE, 2008, Chania, Crete.

[11] S. Asteriadis, P. Tzouveli, K. Karpouzis, and S. Kollias. "Non-verbal feedback on user interest based on gaze direction and head pose", 2nd International Workshop on *Semantic Media Adaptation and Personalization* SMAP. 2007, London, United Kingdom.

[12] G. Rebolledo-Mendez and S. de Freitas, "Assessing NeuroSky's usability to detect attention levels in an assessment exercise", in *HCI International*, 2009. San Diego, CA: Springer.