Using Virtual Agents to Cue Observer Attention

Assessment of the impact of agent animation

Santiago Martinez, Robin J.S. Sloan, Andrea Szymkowiak and Ken Scott-Brown
White Space Research
University of Abertay Dundee
Dundee, DD1 1HG. UK
s.martinez@abertay.ac.uk, r.sloan@abertay.ac.uk, a.szymkowiak@abertay.ac.uk, k.scott-brown@abertay.ac.uk

***Abstract*— This paper describes an experiment developed to study the performance of virtual agent motion cues within digital interfaces. Increasingly, agents are used in virtual environments as part of the branding process and to guide user interaction. However, the level of agent detail required to establish and enhance efficient allocation of attention remains unclear. Although complex agent motion is now possible, it is costly to implement and so should only be routinely implemented if a clear benefit can be shown. Previous methods of assessing the effect of gaze-cueing as a solution to scene complexity have relied principally on manual responses. The current study used an eye-movement recorder to directly assess the immediate overt allocation of attention by capturing the participant's eye-fixations following presentation of a cueing stimulus. We found that fully animated agents speed up user interaction with the interface. When user attention was directed using a fully animated agent cue, users responded 35% faster when compared with stepped 2-image agent cues, and 42% faster when compared with a static 1-image cue. These results inform techniques aimed at engaging users' attention in complex scenes such as computer games or digital transactions in social contexts by demonstrating the benefits of gaze cueing directly on the users eye movements, not just their manual responses.**

Keywords: agents, interfaces, computer animation, reaction time, eyetracking

I. Introduction

The allocation of attention by a human observer is a critical yet ubiquitous aspect of human behaviour. For the designer of human-computer interfaces, the efficient allocation of operator attention is critical to the uptake and continued use of their interface designs. Historically, many human-computer interfaces (HCI) have relied on static textual or pictorial cues, or a very limited sequence of frames loosely interconnected over time (for example, on automated teller device menus, or on websites). More recently, the increased power of computer graphics at more cost effective prices has allowed for the introduction of high resolution motion graphics in human computer interfaces. Until now, psychological insights on attention and the associated cognitive processes have mirrored the HCI reliance on either static or stepped pictorial stimuli, where stepped pictorial stimuli consist of a few static frames displayed over time to imply basic motion. Again, this legacy can be attributed to limitations in affordable and deployable computer graphics.

The reported study is centered on the evaluation of fully animated (25 frames per second) virtual agents, where both the head and eye-movements of the agent are animated to allocate user attention. In contrast to most previous studies that have relied on manual responses to agent gaze, the current study uses the captured eye-gaze of the participant as a response mechanism, following on from the work of Ware and Mikaelian [15].

Where observers look in any given scene is determined primarily by where information critical to the observer's next action is likely to be found. The visual system can easily be directed to guide and inform the motor system during the execution of information searching. Consequently, a record of the path observer gaze takes during a task provides researchers with what amounts to a running commentary on the changing information requirements of the motor system as the task unfolds [4]. This is the underlying principle of the reported experiment, which is an expansion of the cognitive ethology concept expressed by Smilek et al. [3] to virtual agents. The experiment is based on the deictic gaze cue – the concept that the gaze of others acts like a signal that is subconsciously interpreted by an observer's brain, and that it can transmit "information on the world" [10]. The gaze of another human agent is inherently difficulty to avoid, and it can be used as a specific pointer to direct an observer's attention [8]. The incorporation of this concept can be easily implemented into an agent-based interface.

The efficiency of such an interface can be assessed based on the speed of observer response to cues. In the case of the current study, the cues are presented as fully animated (dynamic) agents, stepped agents (two images), or static agent images. Coupled with appropriate software, a virtual agent can anticipate user's goals, and point (using gaze) to the area where the next action has to be performed. An agent with animated gaze may therefore be useful to adopt in digital interfaces to guide user attention and potentially increase the speed of attention allocation, or where the work space of human physical action may have many possible choices and the possibility of not selecting the right one is high.

In the following sections we will explain in detail the application of the virtual agent to cue observer attention. In

Section 2 we will describe the existing literature reviews from two different research fields. In Section 3 we will explain the method used to develop the experiment. In Section 4 we will present the results of that experiment. Finally, in Section 5, we will discuss the overall results, the effects of 3D compared with 2D agents and the impact on user engagement and agent animation.

II. LITERATURE REVIEW

Previous studies belong to two different but related research fields: namely cognitive psychology and computer interface design. Psychological studies have reviewed attention and its relationship with the cues. Posner [11] describes the process of orienting attention. Relative to neutral cue trials, participants were faster and/or more accurate at detecting a target given a valid cue, and they were slower and/or less accurate given an invalid cue. Friesen and Kingstone [5] worked with faces and lines drawn following the gaze direction towards the target area. They found that subjects were faster to respond when gaze was directed towards the intended target. This effect was reliable for three different types of target response: detection, localization and identification. Langton and Bruce [3], and more recently Langton et al. [9], investigated the case of attention in natural scene viewing. They concluded that facial stimuli which indicate direction by virtue of their head and eye position produce a reflexive orienting response in the observer. Eastwood et al. [3] produced experimental findings which led to the conclusion that facial stimuli are perceived even when observers are unaware of the stimuli. In 2006, Smilek et al. [3] focused on isolating specific processes underlying everyday cognitive failures. They developed a measure for attention-related cognitive failures with some success, and introduced the term of cognitive ethology.

Studies in HCI and computing are focused on proving the validity of eye-gaze as an input channel for machine control. Ware and Mikaelian [15] used an eye-tracker to compare the efficacy of gaze an as an input channel with other more usual inputs, such as manual input using physical devices. They found that the gaze input was faster with a sufficient size of target. Sibert and Jacob [12] studied the effectiveness of eye gaze in object selection using their own algorithm and compared gaze selection with a traditional input – a hand operated mouse. They found that gaze selection was 60% faster than mouse selection. They concluded that the eye-gaze interaction is convenient in workspaces where the hands are busy and another input channel is required.

The above research shows how eye-gaze can be used to assess the response of a user when accurate tracking is possible. In addition, it has been demonstrated that the eye-gaze of an agent can effectively allocate attention. However, the interplay between pictorial cues to gaze allocated attention (and subsequent assessment of allocated attention) is still to be fully explored. In particular, for the reported experiment, two goals were set by the authors; to assess the extent to which the gaze of the observer can be used to record their selection of targets and response time to agent cues, and to determine whether fully animated agents would offer an advantage over standard static (1-image) or stepped (2-image basic motion) agents when directing attention using gaze. By focusing on gaze as a means of target selection, this removes as much motor response as possible from the observer. Manual responses inevitably introduce uncertainty in establishing the true response time since they are an indirect response to the gaze cue (requiring over allocation of attention and eye-gaze, followed by translation of the response signal to the sensory modality of touch). Therefore, when it comes to assessing the effectiveness of animated versus static and stepped agent cues, directly recording the eye-movements of observers and using this data to determine the speed of their response and their selection of objects offers a significant advantage.

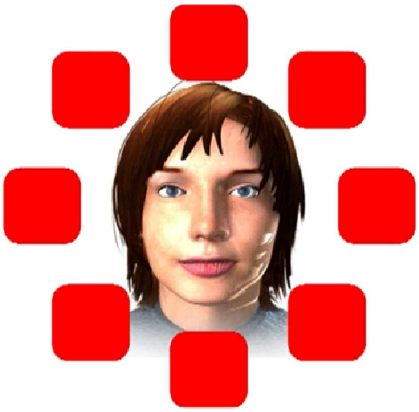

Figure 1: The appearance of the virtual agent, surrounded by eight target squares, arranged on both the cardinal and oblique axes.

III. METHOD

A. Task description

In this experiment, participants were asked to perform an object selection task (using their gaze alone) on a series of twenty-four different agent animations, presented on a monitor at a resolution of 1024 x 768. Each of the videos showed a virtual agent's head in the centre of the screen surrounded by eight different possible target areas (see Fig. 1). Each agent was displayed on screen for 3000 ms. Over the course of the video, the agent would orient its head and eyes aim at a particular target square. The point at which the agent oriented its head and gaze (and the nature of the agent's movement) was determined by the type of agent cue (see below). Of the eight target areas in each video, only one was the right choice in each trial – the one that was specifically indicated by the agent. If the participants selected that specific area with their eye-gaze, it was counted as a success. If the participant selected any of the other seven areas, it was counted as incorrect. Fixations to areas outside the 8 target areas were coded as no target selected. The target areas were red squares approximately 150 x 150 pixels in size, and were all equidistant from the center of the screen.

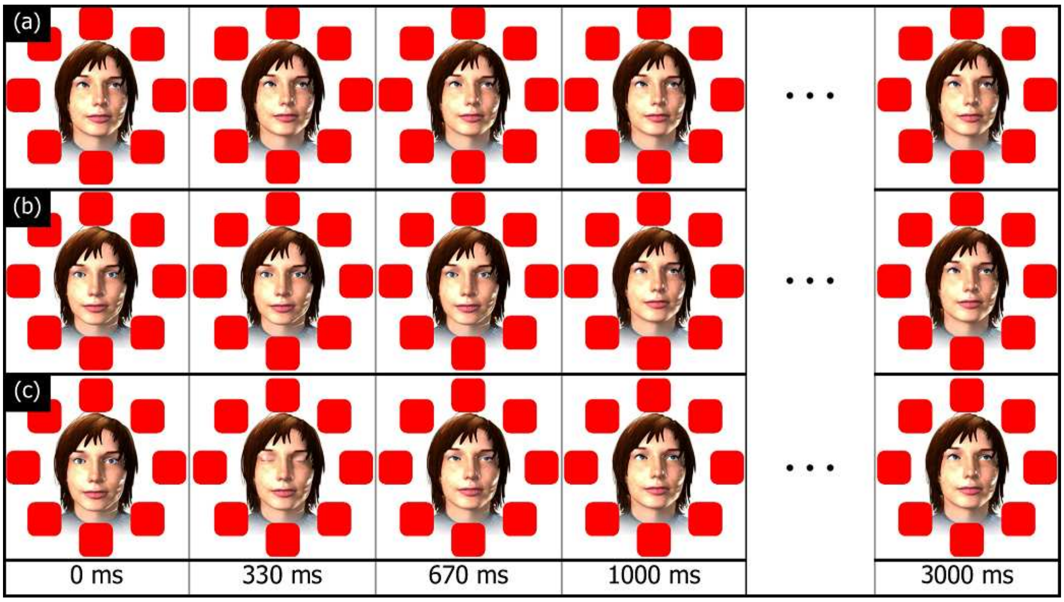

Figure 2. The appearance of the three types of helper agents over 1000 ms. Helper agents used head orientation and gaze to highlight one of eight targets. In the above example, three types of helper agent are shown highlighting the NE target. (a) shows a static (1-image) helper agent, which highlights the NE target from 0 ms onwards. (b) shows the stepped (2-image) helper agent, which looks towards the observer in frame 1 (from 0 ms) before changing to highlight the NE target in frame 2 (from 960 ms). (c) shows the dynamic (25-image, 25 fps) agent, which begins at 0 ms by looking at the observer, and is animated with natural movement so that the head and gaze shift towards the NE target at 960 ms. All helper agents are shown. to participants for a total of 3000 ms, so that the appearance of the agent at 1000 ms is held for two seconds.

B. Agent Cues

There were three different types of agent cues (see Fig. 2):

a) Static cue: A single image of an agent. The agent's head and eyes were aimed at the target area for the duration that the stimulus is displayed. The orientation cue was therefore presented from 0 ms till 3000 ms.

b) Stepped cue: Two images of an agent, sequenced to imply movement. The agent's head and eyes were looking straight forward from 0 ms, before the second image was displayed from 960 ms. In the second image, the agent's head and eyes were aimed at the target from 960 ms till 3000 ms.

c) Dynamic cue: A fully animated agent, showing naturalistic movement from 0 ms to 960 ms. The agent's head and eyes were pointing straight forward at 0 ms, before the agent moved (at 25 fps) to aim its head and eyes at the target area. The agent's gaze and head were aimed at the target at 960 ms. The full orientation cue was therefore presented from 960 ms till 3000 ms.

C. Participants

A total of sixteen participants were recruited from students and staff at the University of Abertay-Dundee. There was no compensation and all had normal or corrected-to-normal vision. During the experiment, two of them used contact lenses.

D. Apparatus

To capture participant gaze data, a modified (fixed position) SMI IView HED eye-movement recorder with two cameras was used. One camera recorded the environment (the target monitor) and the other tracked the participant's eye by an infrared light recording at a frequency of 50 Hz and accuracy of 0.5° of visual angle. Stimuli were presented on a TFT 19'' monitor with a 1024 x 768 resolution and 60Hz of frequency controlled by a separate PC. The monitor brightness and contrast were set up to 60% and 65% respectively to ease the cameras' recordings and avoid unnecessary reflections. Also, both devices were individually connected to two different computers. Viewing was conducted at a distance of 0.8 meters in a quiet experimental chamber.

Each participant underwent gaze calibration controlled by the experimenter prior to the start of data collection. The participant was sat down in a height adjustable chair with their chin on the chin rest and in front of the monitor at 0.9 meters distance. Firstly, the calibration of the eyetracker was completed by presenting a sequence of five separate screens with dots in the center and in the corners. The calibration covered the same surface occupied by the target areas.

A final image with the set of five points was shown to double check the calibration by the operator. The calibration

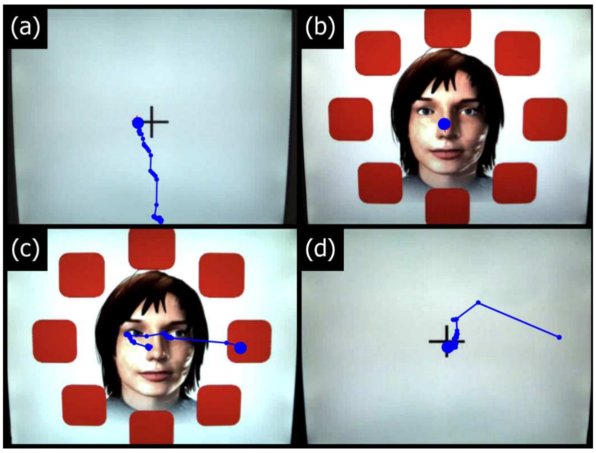

Figure 3: The eye tracking data of one participant, where the blue circles represent fixations. In image (a), the participant looks towards the cross before the agent appears in image (b). In image (c), the agent highlights the East target, at which point the participant looks towards this target, before fixating on the cross again in image (d).

was repeated if necessary following adjustments to the camera positions to ensure good calibration. The experiment started with a ten seconds countdown sequence. After that, the series of twenty-four videos (3 agent cue types x 8 target areas) were presented to participants in a randomized order. The duration of each task video was three seconds, and each video was shown one by one full screen. Before each task video, a central black cross over a white background was shown for two seconds to center the gaze of the participant. This ensured that the participant was looking at the centre of the screen at the start of each video. Fig. 3 shows sample screen captures from the eye-tracker.

E. Data analysis

The participant gaze data was analyzed using the software BeGaze 2.3. The data stored in BeGaze contained all the fixations' timestamps. Only trials where the participant's gaze started on the cross in the center of the screen were considered valid. Target selection was defined by the first full-gaze fixation occurring in the eight predefined areas of interest overlying the 8 target destinations. The fixation duration criterion for an observer response is defined in the light of previous literature. Ware and Mikaelian (1987) used 400 ms; Sibert and Jacob (2000) considered 150 ms. Because extended forced fixation (400 ms) can become laborious, we established a criterion for a successful cognitive response to fixation as equal or greater than to 250 ms, i.e., a fixation that locates on the target area at least for 250 ms.

Based on this concept, of the total number of possible cognitive responses, 92.18% were successfully tracked. Of the successfully tracked data, correct responses accounted for 95.2% of the total and mismatches accounted for 4.9%. The definition of a mismatch was when there was a fixation of 250 ms or more inside an incorrect target area. In 8.47% of the total mismatches, no clear target was selected – i.e., there was no fixation of 250 ms or more in any of the target areas.

TABLE I. PARTICIPANT SELECTION OF TARGETS

Type	Correct	Incorrect	No Target selected	Corrupt (Exclusions)
Static	95 %	5.7 %	0.8 %	7 / 128
Stepped	92.5 %	5.8 %	1.7 %	8 / 128
Dynamic	94.2 %	5 %	0.8 %	7 / 128

IV. RESULTS

Only one participant presented problems during the tracking because of the unexpected movement of her contact lens in the tracked eye. This resulted in four non-tracked responses in the same participant.

For each agent type a total of 128 eye tracking recordings were made. Recordings were then evaluated and allocated to one of four categories: Correct (where the observer clearly selected the intended target), Incorrect (where the observer clearly selected an unintended target), No Target (where it was not clear which target the observer had selected), and corrupted (where the eye tracking data had been disrupted resulting in lost data, for instance interference from reflections or other light sources). After excluding the corrupted recordings, it was clear that observers were able to accurately select the intended target regardless of whether the virtual agent was static (95%), stepped (92.5%), or dynamic (94.2%) (see Table I). This would suggest that, in general, the type of virtual agent (in terms whether it was static, stepped, or fully animated) did not substantially impact upon how effective it was at communicating what the intended target was.

A repeated measures analysis of variance (ANOVA) was used to determine whether agent type had an effect on how long it took participants to look at and select the intended target square. The response times for static agent cues - which contained agents which were oriented towards the target 960 ms earlier than both stepped and dynamic cues – were corrected to account for this difference. The analysis showed that the type of agent did have a significant effect on participant response time, F(2, 30) = 52.73, $p < .001$. Participants responded most quickly to the dynamic (fully animated) agent type (M = 1220, SE 95) than they did to either the stepped (2 frame) agent type (M = 1874, SE 61) or the static (1 frame) agent type (M = 2091, SE 59) (see Fig. 4).

Comparisons between agent types were assessed using the Bonferroni post-hoc test. The results showed that participants responded to the dynamic agent type significantly more quickly than both the static (Mean Deviation (MD) = 870, $p < .001$) and the stepped (MD = 654, $p < .005$) agent types. Furthermore, participants also responded to the stepped agent type significantly more quickly than the static agent type (MD = 217, $p < .005$) (see Table II). These results not only underline that static agent

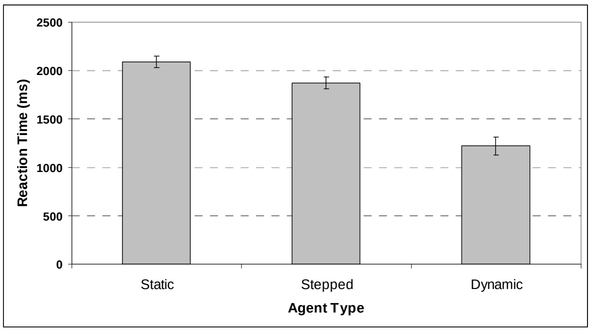

Figure 4: The mean response times for static, stepped, and dynamic agents indicate that participants reacted most quickly to the fully animated, dynamic agents

types are significantly less effective at cueing observer attention than either stepped or dynamic agents, but also that stepped agent types are significantly less effective than fully animated, dynamic agents.

V. DISCUSSION AND FUTURE WORK

Using a paradigm where the criterion for correct response to pictorial or animated agent gaze is the eye-gaze of the participant we found that the presence of full-motion in the gaze inducing agent drives the observer's attention the fastest. Gaze recorded responses for 25 frame stimuli were 35% faster than stepped and 42% faster than static stimuli. This result is consistent with previous research on gaze cueing [9]. The current paradigm provides the most direct route to the establishment of the overt allocation of gaze location since it subverts the need for a translation to a manual response. This confirms Ware and Mikaelian's [15] assertion that participants eye-gaze itself can be used to indicate responses.

The presence of movement in gaze cueing stimuli seems to drive the user's attention more quickly. One prediction arising from this is that, when compared with 2D agents, 3D agents make the expectations of more believable behaviour. The combination of additional pictorial cues and natural motion may make the appearance of the agent more akin to that of a human conversation partner. The additional realism possible with modern computer animation techniques may make agents more believable and engaging [14].

The present study indicates how the animation of an agent can be linked to the sequencing of the social 'script' or 'narrative' of a HCI interface experience. Previous investigators such as Kendon [6] observed a hierarchy of body movements in human speakers; while the head and hands tend to move during each sentence, shifts in the trunk and lower limbs occur primarily at topic shifts. They discovered the body and its movements as an additional part of the communication, participating in the timing and meaning of the dialogue. Argyle and Cook [1] discuss the use of deictic gaze in human conversation. They argued that during a conversation the gaze serves for information seeking, to send signals and to control the flow of the conversation. They explained how listeners look at the speaker to supplement the auditory information. Speakers on the other hand spend much less time looking at the listener, partially because they need to attend to planning and do not want to load their senses while doing so. Preliminary work from our laboratory suggests that experience in the gaze task over time may lead to a learning effect whereby extended exposure to these stimuli leads to improved gaze allocation, this analysis will form part of a wider study including a sequence of guided navigation prompts in a naturalistic setting. Only by creating a natural sequence of user choices with a combination of gaze cues and items competing for attention (including distractors) can we fully confirm the efficacy of an agent-based cue in human computer transactions in the natural environment. The research here is consistent with the wider conclusions of other investigators [14], which indicate that vivid, animated emotional cues may be used as a tool to motivate and engage users of computers, when navigating complex interfaces. The results of this experiment provide guidance for agent design in consumer electronics such as computer games or animation. In order to avoid an unpleasant robotic awareness, natural motion and the correct presentation of the cue contribute to increase the deictic believability of the agent. Deictic believability in animated agents requires design that considers the physical properties of the environment where the transaction occurs. The agent design must take account of the positions of elements in and around the interface. The agent's relative location with respect to these objects, as well as social rules known from daily life, are critical to create deictic gestures, motions, and speech that are both effective, efficient and unambiguous. All these aspects have an effect in addition to the core the response time measure. They easily trigger natural and social interaction of human users, reaching the right level of expectations. Furthermore, they make the system errors, human mistakes and interaction barriers more acceptable and navigable to the user [2].

TABLE II. MULTIPLE COMPARISONS BETWEEN AGENT TYPES

Type	Comparison	Mean Deviation	Std. Error	Sig.
Static	Stepped	217 ms	54.3	.004
	Dynamic	870 ms	85.8	.000
Stepped	Static	-217 ms	54.3	.004
	Dynamic	654 ms	114.3	.000
Dynamic	Static	-870 ms	85,8	.000
	Stepped	-654 ms	114.3	.000

ACKNOWLEDGMENT

Martinez is grateful for support from the Alison Armstrong Studentship. Sloan is grateful for support from a University funded studentship. The authors gratefully acknowledge the supportive interdisciplinary research environment provided by Abertay's Whitespace Research Group, Institute for Arts Media and Computer Games and Centre for Psychology.

REFERENCES

[1] M. Argyle and M. Cook, “Gaze and mutual gaze”, New York: Cambridge University Press, 1976, 221 pages, ISBN-13: 978-0521208659.

[2] Diederiks, E. M, “Buddies in a box: animated characters in consumer electronics”,IUI '03, 2003, pp. 34-38, doi: http://doi.acm.org/10.1145/604045.604055

[3] J. D. Eastwood, D. Smilek, and P. M. Merikle, “Differential attentional guidance by unattended faces expressing positive and negative emotion”, Perception & Psychophysics, vol. 63, 2001, pp. 1004-1013.

[4] J. M. Findlay and I. D. Gilchrist, “Active vision: the psychology of looking and seeing”, Oxford University Press, Oxford. 2003.S. R. H.

[5] C. K. Friesen and A. Kingstone, “The eyes have it! reflexive orienting is triggered by nonpredictive gaze”, Psychonomic Bulletin and Review, vol. 5, 1998, pp. 490-495.

[6] A. Kendon, “Some relationships between body motion and speech: ana anlysis of an example”, In: A. Siegman and B. Pope, eds, Studies in Dyadic Communication, pp. 177-210, Elmsfor, NY: Pergamon Press, 1972.

[7] S. R. H. Langton and V. Bruce, “Reflexive visual orienting in response to the social attention of others”, Visual Cognition, vol. 6, 1999, pp. 541-567., doi:10.1080/135062899394939

[8] S. R. Langton, R. J. Watt, and V. Bruce, “Do the eyes have it? Cues to the direction of social attention”, Attention And Performance, vol. 4(2), 2000, pp. 50-59, ISSN 1364-6613, DOI: 10.1016/S1364-6613(99)01436-9

[9] S. R. Langton, C. O’Donnell, D. M. Riby, and C. J. Ballantyne, “Gaze cues influence the allocation of attention in natural scene viewing”, Experimental Psychology, vol. 59(12), 2006, pp. 2056-2064, doi: 10.1080/17470210600917884.

[10] I. Poggi and C. Pelachaud, “Signals and meanings of gaze in animated faces,”. In: S. Nuallain, C. Muhlvihill and P. McKevitt, eds, Language, Vision and Music. Amsterdam: John Benjamins, 2001

[11] M. I. Posner, “Orienting of attention”, Quarterly Journal of Experimental Psychology, vol. 32, 1980, pp. 3–25.

[12] L. E. Sibert and R. J. Jacob, “Evaluation of eye gaze interaction”, In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 00), ACM, 2000, pp. 281-288, doi: 10.1145/332040.332445.

[13] D. Smilek, E. Birmingham, D. Cameron, W. Bischof, and A. Kingstone, “Cognitive ethology and exploring attention in real world scenes,” Brain Research, vol. 1080, Issue 1, Attention, Awareness, and the Brain in Conscious Experience, 2006, pp. 101–119.

[14] T. Vanhala, V. Surakka, H. Siirtola, K. Raiha, B. Morel, and L. Ach, “Virtual proximity and facial expressions of computer agents regulate human emotions and attention”, Computer Animation And Virtual Worlds, vol 21(3-4), 2010, pp. 215-224, doi: 10.1002/cav.336.

[15] C. Ware and H. H. Mikaelian, “An evaluation of an eye tracker as a device for computer input”, SIGCHI Bull., 17, May. 1986, pp. 183-188, doi:10.1145/30851.275627.