Automatic Low-Level Overlays on Presentations to Support Regaining an Audience's Attention

Walter Ritter[1](✉), Guido Kempter[1], Isabella Hämmerle[1], and Andreas Wohlgenannt[2]

[1] Vorarlberg University of Applied Sciences, Hochschulstr. 1, 6850 Dornbirn, Austria
walter.ritter@fhv.at
[2] WolfVision GmbH, Oberes Ried 14, 6833 Klaus, Austria

Abstract. In a world full of distractions, keeping an audience focused on a presentation is getting increasingly difficult. In this paper, we propose a system that supports presenters in a nearly subliminal way to regain attention of the overall audience. The system uses a measure of motion complexity inside the audience area as an estimate for overall attention. It then applies low-level visual overlays over presentations if the estimated level of attention is getting too low. Ideally, these dynamically adapted visual overlays can be detected in the peripheral field of view but not in the foveal field of view. In a pilot study with 14 participants, we tested the feasibility of this approach with a simplified version of the system, limiting stimuli to red colored overlays up to an opacity of 20%. First results show that motion complexity can indeed be a good indicator of distractions and low-level visual overlays can lead to a higher perceived level of agitation. However, the visual effects used in this pilot study have been partly perceived by the audience. Further work is needed to identify visual stimuli that are best fitted for recapturing attention without irritating those already focused on the presentation.

Keywords: Presentation support · Subliminal · Adaptive systems
Visual overlays

1 Introduction

Remaining focused on presentations in today's world is getting increasingly difficult, given the countless distractions offered by smartphones and other personal electronic devices, in addition to the various external and ambient distractions. This imposes a big challenge on presenters to perfectly prepare their presentations for their audience.

In a study, Ward et al. (2017) report that even having phones in silent mode makes it harder to stay focused and reduces cognitive performance compared to having no phone at all. But also without smartphones the available attention span is limited. Reports on the typical attention span duration vary broadly. McKeachie and Svinicki (2013) mention a 10-min attention span, which they base on a study about note taking during lectures by Hartley and Davies (1978). Bunce et al. (2010) mentioned that students self-reported an attention decline already after 4–5 min into a presentation after the initial settling in period, and that they would also alternate between attention and nonattention

M. Kurosu (Ed.): HCI 2018, LNCS 10901, pp. 429–440, 2018.
https://doi.org/10.1007/978-3-319-91238-7_35

states in briefer and briefer cycles during a lecture. In general Wilson and Korn (2007) criticize the lack of consideration of individual differences in various attention span studies. A widely-reported study conducted by Microsoft, which described an attention span of just 8 s in 2013, has been disregarded in this context, as it basically referred to the time people typically spent on a web page before moving on, and not to actual attention span durations (Bradbury 2016).

Independent of the actual length of the attention span, it's not only a matter about attention itself, it's also about where members of the audience divert their attention to, or expressed differently, what might distract them. Pratto and Oliver (1991) describe negative social information as a powerful source to grab attention. Smith et al (2003) reported about a negativity bias in attention allocation, meaning negative stimuli can divert the attention away from the current focus more easily than positive ones.

As soon as some individuals of an audience get distracted, this state might be passed on to others subconsciously by mood transfers, which could possibly lead to an agitated audience (Tudor et al 2013). This process is caused by the same mechanisms as the ones underlying for example contagious yawning, which is described as nonverbal communication to synchronize behavior within a group (Guggisberg 2011). The mechanism of mood transfers has originally been discussed in birds' research, where one bird fleeing causes all others of the group to flee too at virtually the same time (Lorenz 1935).

There are, of course many ways presenters can prepare for these challenges in first place. Having the information perfectly selected and prepared according to the interests of the expected audience is an important first step. Obeying best practices for preparing presentations is another important step towards captivating the attention of an audience.

However, an interesting question arises. Could a presentation system possibly support the presenter in a subliminal way to regain the attention of an audience? What would be needed to distract the audience from the countless distractions back to the presentation? In this paper, we present a concept for such a system and describe a brief pilot study to test the feasibility of this approach.

Such a system should be able to judge the overall level of attention of an audience and to present stimuli in a way, that recaptures the interest of distracted people without distracting those people already focused on the presentation. In the next section, we look at some of the related work before we present our approach.

2 Related Work

Measuring the level of attention in classroom settings has been of interest to researchers since a long time. One approach is described by Helmke & Renkl in their Münchner Aufmerksamkeitsinventar (MAI). The basic idea is to observe students in a classroom during a lecture and sequentially code the attention state of individual students in fixed time intervals of 5 s. For this, attention states are classified in two layers. The first one discriminates between no-task, on-task, and off-task states. In the second layer the on-task state is further divided into compliance with task-requirements, self-initiated, or externally initiated activities. The off-task state is further divided into active (e.g. distracting others) or passive (e.g. dreaming) state (Helmke and Renkl 1992). Hommel presented an adjusted variant that's modified to better fit higher education and makes use of videography to analyze all students in 30 s intervals over the whole time (Hommel 2012). Both approaches also consider the lecture context in addition to attention state.

More advanced approaches make use of emotion recognition systems that analyze facial expressions (Magdin et al. 2016; Magdin and Prikler 2018; Heyjin 2017; Ayvaz and Gürüler 2017; Daouas and Lejmi 2018), eye movements (Krithika and Lakshmi 2016), voice characteristics, written text or a combination of different modalities (Nedji et al. 2008; Bahreini et al. 2016) to deduce the emotional state of a person. These systems are mostly used and tested in an e-learning context, where the knowledge of the emotional state should allow for better feedback by the remote lecturer. However, most of these systems are optimized for observations of individuals, but not suited as an overall measure for a whole audience. Furthermore, members of an audience (e.g. students in a lecture or in general listeners to presentations) would not feel comfortable if they would be tracked individually by cameras.

The human eye features different capabilities in the foveal and the peripheral field of view. In its periphery, the eye has a higher temporal resolution and sensitivity for light changes than in the foveal area. This higher sensitivity to light changes is caused by a higher rod density in the periphery (Curcio et al. 1990; Jonas et al. 1992; Wells-Grey et al. 2016). Rod mediated vision also seems to cause more activation in the brain area that is associated with motion, compared to cone mediated vision (Hadjikhani and Tootell 2000). The reason for the difference between cone- and rod-density could be originated in evolution: humans with this feature were able to detect dangerous situations outside of their current field of view faster and could react to the situation in time and thus survived.

These characteristics might also be exploitable by a system that targets people not directly looking at it.

Adaptive systems that react to some form of user input in an indirect/non-obtrusive way to better support users have been used in various fields. Ward et al. (2000) demonstrated a novel text input interface, which used previous input and learned language statistics to make it easier to select probable succeeding characters. This mechanism enables efficient text input even by only using an input modality with 1- or 2-dimensional input capabilities. The same principle is used for modern touch keyboards that automatically increase the sensitive area around certain keys that are more probable to be typed next than others, and therefore make them easier to reach. Kempter et al. (2003) discussed the benefits of analogue communication in the human computer interaction context, where a system's behavior can be influenced by analogue channels with no explicitly/pre-defined meanings. Ritter (2011) described different approaches where subliminal feedback loops in human computer interaction, based on these analogue communication patterns, could be used to support users in a transparent and unobtrusive way. In a widely-criticized study for its ethical aspects, Kramer et al. (2014) have shown that emotional states can be transferred over social networks. Users tend to take on the same emotions as conveyed in their news stream, even without the users being aware of this process. These studies demonstrate that systems can be used to influence people in subliminal ways.

While people might feel uncomfortable, when they hear they can be influenced in a subliminal way, the process of influencing people is one of the corner stones of advertising and is already happening all around the clock.

3 Concept and Implementation

Based on the insights of the work described in the previous section, we aimed to create a presentation system that observes the audience, deduces an estimate for overall attention and applies evolving visual stimuli to recapture the attention of people looking away from the presentation.

The system shown in Fig. 1 depicts the basic system stages. The input processing layer consists of an image analysis module that takes input from a webcam and calculates a motion complexity index. This index is then forwarded to the adaptation loop stage. The input processing layer also contains an environment module that monitors air conditions inside the room, like CO_2 level, temperature, and humidity. These measures are currently logged but not yet used as input for the adaptation loop. The adaptation loop module implements a simple genetic algorithm that uses the motion complexity input as fitness values for its candidate adaptations. Once fitness values are available for each candidate in the population, a new generation of adaptation candidates is generated using a roulette wheel algorithm for parent selection. Each candidate is then applied for a situation with exceeding motion complexity and again assigned a corresponding fitness value based on the motion complexity. In our current system, candidate solutions are encoded using a 10-bit string shown in Table 1.

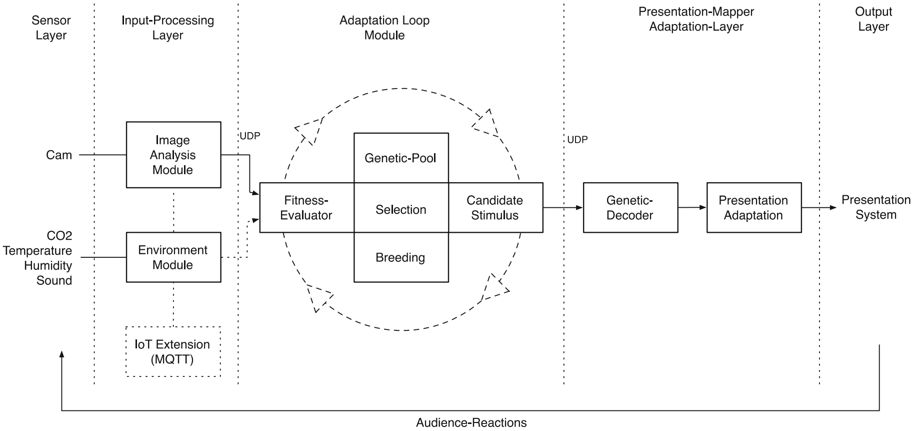

Fig. 1. System overview of an adaptive presentation system, consisting of an input processing stage, an adaptation loop module, a presentation mapping stage, and the output layer.

Table 1. Example genetic encoding of visual overlay parameters, each consisting of 2 bits

Parameter	Values
Opacity	0.0, 0.1, 0.15, 0.2
Transition time	0.025, 0.050, 0.100, 10
Fade-In curve	EaseInOut, Linear, EaseIn, EaseOut
Fade-Out curve	EaseInOut, Linear, EaseIn, EaseOut
Color	black, red, blue, white

New evaluation cycles are triggered every 60 s during situations with too much motion. The following 30 s are used as timeframe for measuring the reactions of the audience following a stimulus (see Fig. 2).

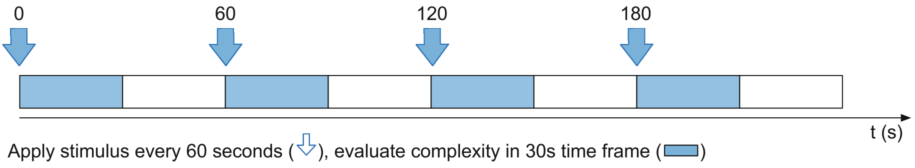
0
60
120
180
t (s)
Apply stimulus every 60 seconds (⇩), evaluate complexity in 30s time frame (▬)

Fig. 2. Scheduled stimulus application and evaluation timeframes

The presentation mapper takes a candidate bit string as input and maps this to specific adaptation animations that are then overlaid over the presentation. The main idea of these adaptations is to be very low-level and fast, effectively producing a lighting change, and therefore being more noticeable in the peripheral field of view while hardly being visible for people looking directly at the presentation.

Motion complexity (MC) inside the audience area is derived via an image processing algorithm, taking a video stream from a webcam as input and performing grayscale conversion, background subtraction, thresholding, and morphing operators to the individual images (see Fig. 3) using OpenCV. This results in a black and white image with connected and independent motion areas.

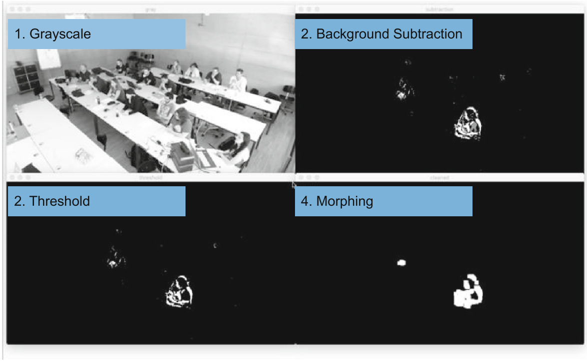

Fig. 3. Image processing steps for retrieving motion complexity

From this, two different indices are calculated (see Eqs. 1 and 2).

$$MC_{rel} = \text{image area with changes} \,/\, \text{total image area} \tag{1}$$

$$MC_{cnt} = \text{count of independent movement areas} \tag{2}$$

MC_{rel} should give an overview of the overall action taking place in the camera's field of view and can also be used to detect lighting changes inside the room (e.g. a percentage higher than 50%). MC_{cnt} reflects the number of independent disturbance areas and helps to counter the effect that moving objects closer to the camera have a bigger influence on $MC_{rel.}$

We set up a small pilot study to explore the usefulness of the two different motion complexity indices for estimating an audience's attention level, and to test the feasibility of this approach in a real-life setting.

4 Pilot Study

In a small pilot study, we used a simplified version of the system described in the previous section, to mainly investigate the plausibility of the motion complexity indices and to see if a background color overlay would cause a noticeable effect. Since the genetic algorithm adaptation loop would have added too much variability, we chose to apply a red background overlay of varying opacity with a maximum value of 20%, depending on the current motion complexity. Red was chosen because of the common signal-meaning of it. This overlay will also result in a lighting change depending on the opacity, which should mainly be recognizable from the visual periphery.

4.1 Method

The system was installed in a lecture room with the camera placed so that it was observing the audience but not covering the lecturer area. During the presentation, text-based and image-based slides were shown, both with and without dynamic overlays. Figure 4 shows the sequence of the lecture slides. The audience was not told in advance that there would be overlays. They were informed however, that they would have to fill out a brief questionnaire after some slides. Each slide was shown for exactly 60 s.

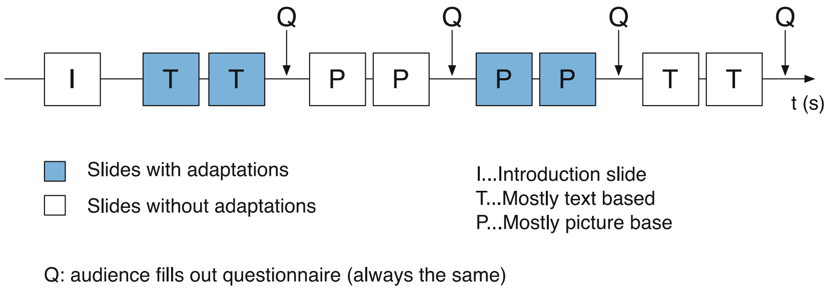

Fig. 4. Timeline of the test sequence. Duration of each slide is 60 s

In the pilot study 14 persons (6 female, 8 male) attended the test presentation. After each segment the participants had to rate the following properties on a 7-step rating scale:

Q1: “I followed the presentation…” from “exactly (1)” to “not at all (7)”
Q2: “At the moment I feel…” from “very calm (1)” to “very agitated (7)”
Q3: “The presentation content was…” from “very exciting (1)” to “very boring (7)”
Q4: “The slides were…” from “very intrusive (1)” to “very conservative (7)”

4.2 Results

Table 2 lists the descriptive statistics of the questionnaires. The text-based slides with ongoing adaptations were rated the most boring (M = 4.36, SD = 1.50). Text based slides without overlays were rated the most conservatives (M = 4.0, SD = 1.47). The reported

Table 2. Descriptive statistics of ratings

	Q1		Q2		Q3		Q4	
Stimulus	M	SD	M	SD	M	SD	M	SD
Adaptive text	3.00	2.00	3.71	1.82	4.36	1.50	3.64	1.15
Static pictures	1.93	0.92	2.71	1.54	3.29	1.49	3.71	1.14
Adaptive pictures	2.62	1.50	3.15	1.82	3.77	1.36	3.31	1.38
Static text	2.71	1.20	2.86	1.56	3.57	1.34	4.00	1.49

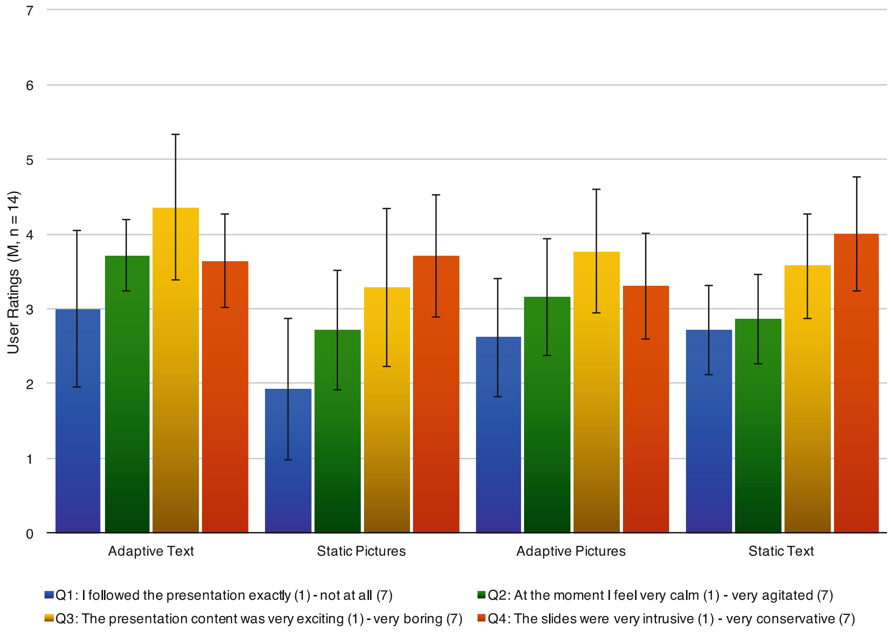

Fig. 5. Mean rating values on a scale from 1 to 7 of the different slide types (n = 14). Error bars show the 95% confidence interval.

attention (Q1) was rated highest for the static pictures (M = 1.93, SD = 0.92) and adaptive pictures (M = 2.62, SD = 1.50). Agitation was rated lowest for the two static slide types ($M_{pictures}$ = 2.71, $SD_{pictures}$ = 1.54, M_{text} = 2.86, SD_{text} = 1.56). See Fig. 5 for a graphical illustration.

T-tests were performed between the static and dynamic phases. The results show that there were significant differences between the reported agitation-level ($M_{adaptive}$ = 3.57, M_{static} = 2.79, p = .019) and the feeling-bored-level of the slides ($M_{adaptive}$ = 4.18, M_{static} = 3.43, p = .007). See Table 3 for more results and Fig. 6 for a graphical overview.

Table 3. Rating differences between adaptive and static slides

Question	$M_{adaptive}$	$SD_{adaptive}$	M_{static}	SD_{static}	$M_{paired-diff}$	p
Q1	2.96	1.90	2.32	1.12	0.64	.098
Q2*	**3.57**	**1.89**	**2.79**	**1.52**	**0.79**	**.019**
Q3*	**4.18**	**1.52**	**3.43**	**1.40**	**0.75**	**.007**
Q4	3.50	1.23	3.86	1.29	−0.36	.321

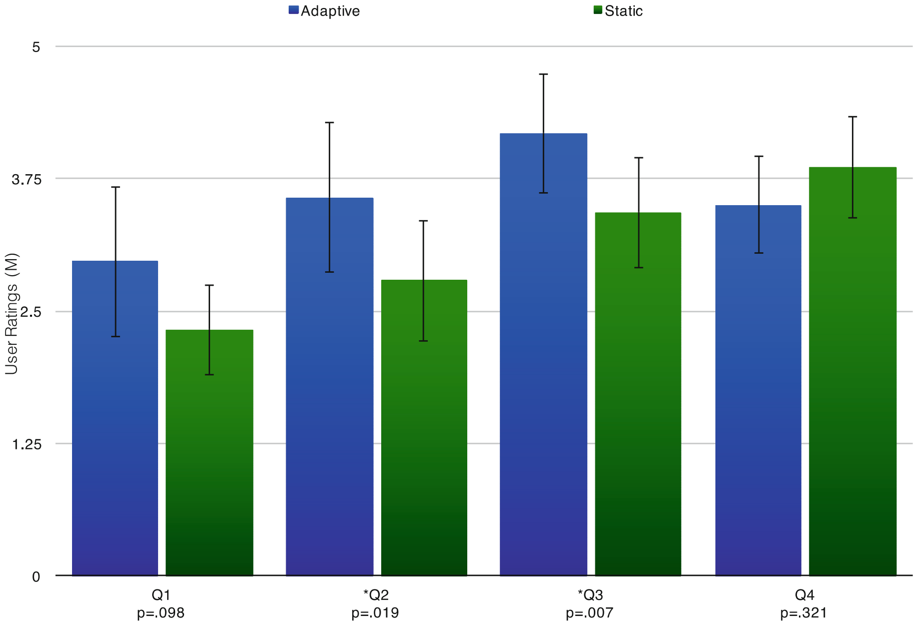

Fig. 6. Differences between adaptive and static slides. Significant differences could be found for agitation (Q2) and excitation (Q3). Error bars show the 95% confidence interval.

To also check the influence of the slide type (picture or text), a t-test has been performed between these groups. However, no significant differences could be identified.

Both motion complexity indices show a peak at the beginning of the lecture and then at the questionnaire slots 3, 6, and 9 (see Fig. 7), indicating that both measures reflect the settling in period of a presentation and react to distractions during the presentation (filling out the questionnaires). It's also interesting to note that the motion complexity levels for the picture slides (4,5 & 7,8) are above the ones for text slides (1,2 & 10,11). This could relate to the higher self-reported attention level of picture slides.

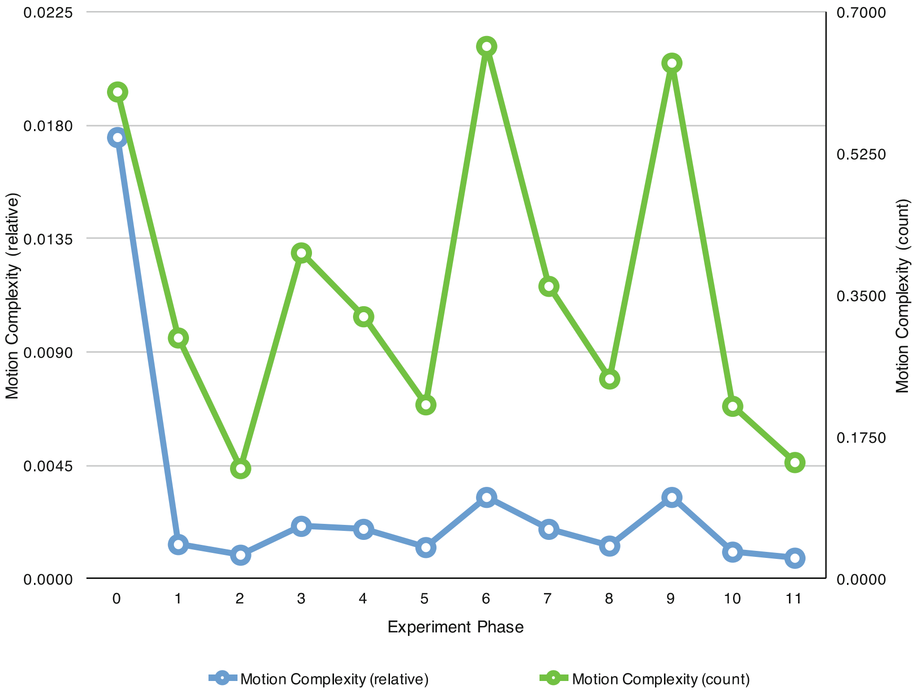

Fig. 7. Motion complexity indices through the course of the presentation. Peaks can be observed at the start, and at phases 3, 6, and 9 (questionnaire action).

5 Discussion and Conclusion

The results show that there was a significant difference between slides with and without adaptions taking place, hinting that we might indeed be able to influence the perception of slides using low-level visual overlays. However, the difficult part is to find the right balance between the noticeability from the peripheral field of view and the center field of view. In the mapping of the pilot study, the strongest visual overlays were too noticeable when directly looking at it – this was probably the reason for the relatively high self-reported agitation level of the slides with adaptions. Overall, the derived motion complexity for picture slides was higher than for text slides. This contrasts with the reported agitation level, where the slides with adaptations were rated higher independent of slide content type, but falls in line with the higher self-reported attention level for picture based slides. A decline from the first to the second slide within a condition could be observed for each variation, which was to be expected due to the settling in phase from the preceding questionnaire section.

One downside of our motion complexity indicators is the susceptibility to overreaction when massive motion occurs (e.g. a person leaving the room or is coming in late). While this is a little improved with MC_{cnt} compared to MC_{rel}, it is still present there too. One workaround could be to set a narrow threshold for valid motion complexity levels, and if this threshold is exceeded, pause the system until things get back to normal. This would also be advisable for filtering out abrupt light changes.

A possible concern is the influence of such a system on people with Epilepsy. Even if our overlays are within the same specs as e.g. movies or games that could also play in the peripheral field of view, further clarifications with experts in this field must be done, before using such a system in real-life settings.

In a next step, different stimuli will be tested in a lab setting regarding their potential to grab peoples' attention while they are focused on a different task, where the stimulus screen is placed outside the foveal field of view. The most promising properties will then be encoded for use with the genetic algorithm. Also, the motion complexity parameters could be compared to attention level ratings using e.g. the ModAI instrument. Once attention levels are rated for specific lectures, the motion complexity parameter algorithms can be fine-tuned using video recordings of the rated lectures.

After this, further tests of the overall system with more participants will have to be done.

Acknowledgements. Financial support for this project was provided by the Austrian research funding association (FFG) under the scope of the COMET program within the research project "Easy to use professional business and system control applications (LiTech)" (contract # 843535). This program is promoted by BMVIT, BMWFJ and the federal state of Vorarlberg.

References

Ayvaz, U., Gürüler, H.: Real-time detection of students' emotional states in the classroom. In: 25th Signal Processing and Communications Applications Conference (SIU), pp. 1–4. IEEE, Antalya (2017)

Bahreini, K., Nadolski, R., Westera, W.: Towards multimodal emotion recognition in e-learning environments. Interact. Learn. Environ. **24**(3) (2016)

Bradbury, N.A.: Attention span during lectures: 8 seconds, 10 minutes, or more? Adv. Physiol. Educ. **40**, 509–513 (2016)

Bunce, D.M., Flens, E.A., Neiles, K.Y.: How long can students pay attention in class? a study of student attention decline using clickers. J. Chem. Educ. **87**(12), 1438–1443 (2010)

Curcio, C.A., Sloan, K.R., Kalina, R.E., Hendrickson, A.E.: Human photoreceptor topography. J. Comp. Neurol. **292**, 497–523 (1990)

Daouas, T., Lejmi, H.: Emotions recognition in an intelligent elearning environment. Interact. Learn. Environ. (2018). https://doi.org/10.1080/10494820.2018.1427114

Guggisberg, A.G., Mathis, J., Schnider, A., Hess, C.W.: Why do we yawn? the importance of evidence for specific yawn-induced effects. Neurosci. Biobehav. Rev. **35**(5), 1302–1304 (2011)

Hadjikhani, N., Tootell, R.B.: Projection of rods and cones within human visual cortex. Hum. Brain Mapp. **9**(1), 55–63 (2000)

Hartley, J., Davies, I.K.: Note taking: a critical review. Program. Learn. Educ. Technol. **15**, 207–224 (1978)

Helmke, A., Renkl, A.: Das Muenchener Aufmerksamkeitsinventar (MAI): Ein Instrument zur systematischen Verhaltensbeobachtung der Schueleraufmerksamkeit im Unterricht. Diagnostica **38**(2), 130–141 (1992)

Hommel, M.: Kodierhandbuch des Beobachtungsinventars zur systematischen und videobasierten Erfassung der Aufmerksamkeit von Lernenden (m/w): modifiziertes Aufmerksamkeitsinventar (ModAI). In: Dresdner Beiträge zur Wirtschaftspädagogik. - Dresden : Technische Universität, vol. 2012, p. 1 (2012). ISSN 0945-4845, ZDB-ID 21890250

Hyejin, K.: Learner's intelligent emotion detection system in U-learning environment. Int. J. u- and e-Serv. Sci. Technol. **10**(8), 91–98 (2017)

Jonas, J.B., Schneider, U., Naumann, G.O.: Count and density of human retinal photoreceptors. Graefes Arch. Clin. Exp. Ophthalmol. **230**(6), 505–510 (1992)

Kempter, G., Weidmann, K.H., Roux, P.: What are the benefits of analogous communication in human computer interaction? In: Stephanidis, C. (ed.) Universal Access in HCI: Inclusive Design in the Information Society, pp. 1427–1431. Lawrence Erlbaum Associates, Mahwah (2003)

Kramer, A.D.I., Guillory, J.E., Hancock, J.T.: Emotional contagion through social networks. Proc. Nat. Acad. Sci. **111** (24), 8788–8790 (2014)

Krithika, L.B., Lakshmi, P.G.G.: Student emotion recognition system (SERS) for e-learning improvement based on learner concentration metric. Proc. Comput. Sci. **85**, 767–776 (2016)

Lorenz, K.: Der Kumpan in der Umwelt des Vogels – Der Artgenosse als auslösendes Moment sozialer Verhaltensweisen. J. für Ornithologie **83**, 137–213 (1935)

Magdin, M., Turčáni, M., Hudec, L.: Evaluating the Emotional State of a User Using a Webcam. Int. J. Interact. Multimed. Artif. Intell. **4**(1), 61–68 (2016)

Magdin, M., Prikler, F.: Real time facial expression recognition using Webcam and SDK affectiva. Int. J. Multimed. Artif. Intell. (2018, in Press)

McKeachie, W.J., Svinicki, M.: McKeachie's Teaching Tips: Strategies, Research, and Theory for College and University Teachers, 14th edn. Wadsworth Publishing, Belmont (2013)

Nedji Milat, I., Seridi, H., Sellami, M.: Towards an intelligent emotional detection in an e-learning environment. In: Woolf, Beverley P., Aïmeur, E., Nkambou, R., Lajoie, S. (eds.) ITS 2008. LNCS, vol. 5091, pp. 712–714. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-69132-7_86

Pratto, F., Oliver, J.: Automatic vigilance: the attention-grabbing power of negative social information. J. Pers. Soc. Psychol. **61**(3), 380–391 (1991)

Ritter, W.: Benefits of subliminal feedback loops in human-computer interaction. Adv. Hum.-Comput. Interact. **2011**, Article ID 346492 (2011)

Smith, N.K., Cacioppo, J.T., Larsen, J.T., Chartrand, T.L.: May I have your attention, please: electrocortical responses to positive and negative stimuli. Neuropsychologia **41**(2), 171–183 (2003)

Tudor, A.D., Poeschl, S., Doering, N.: What do audiences do when they sit and listen? Stud. Health Technol. Inform. **191**, 120–124 (2013)

Ward, A.F., Duke, K., Gneezy, A., Bos, M.W.: Brain drain: the mere presence of one's own smartphone reduces available cognitive capacity. J. Assoc. Consum. Res. **2**(4), 140–154 (2017)

Ward, D.J., Blackwell, A.F., MacKay, D.J.C.: Dasher—a data entry interface using continuous gestures and language models. In: Proceedings of the 13th Annual ACM Symposium on User Interface Software and Technology (UIST 2000), pp. 129–137. ACM, New York (2000)

Wells-Gray, E.M., Choi, S.S., Bries, A., Doble, N.: Variation in rod and cone density from the fovea to the mid-periphery in healthy human retinas using adaptive optics scanning laser ophthalmoscopy. Eye **30**, 1135–1143 (2016)

Wilson, K., Korn, J.K.: Attention during lectures: beyond ten minutes. Teach. Psychol. **34**(2), 85–89 (2007)