MASTER THESIS

IN

COGNITIVE SCIENCE

DEPARTMENT OF COMPUTER SCIENCE

FACULTY OF SCIENCE

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Master Thesis:

Towards Understanding Attention in Virtual Reality - Analysing Visual Attention in a VR-Classroom Experiment

Supervisors

Prof. Dr. Enkelejda Kasneci

Prof. Dr. Richard Göllner

Author

Philipp Stark

Master of Science in Cognitive Science

Tübingen, October 28, 2020

Abstract

Attention can be seen as a key aspect of learning. Most of children's everyday learning takes place in a classroom. But investigating children's attention and learning in a real-world classroom can be difficult. Therefore, we used an Immersive Virtual Reality classroom to investigate children's attention in a 14 minute virtual lesson. We collected information about the objects children had looked at. With the gazed object information, we analysed the total time spent on specific objects of interest (peer learners, teacher, screen) or investigated children's visual attention behaviour with scanpath analysis (ScanMatch, SubsMatch). The study was conducted as a between design with three different classroom manipulations regarding participants sitting position, the avatar style of the peer learners and their hand raising behaviour. We found significant differences regarding children's visual attention for the position they are seated in the classroom and regarding the visual appearance of the peer learners. Additionally, we found indications that children also process social information in the virtual classroom due to effects of the hand raising condition on children's visual attention. These findings can be seen as a first step towards understanding children's visual attention in an Immersive Virtual Reality classroom.

Key Words: Immersive Virtual Reality, Visual Attention, Scanpath, Classroom.

Contents

List of Figures

List of Tables

Acronyms

ADHD Attention Deficit Hyperactivity Disorder. 7, 9, 11

ANOVA Analysis of Variance. I, IV, 29, 36, 51, 54, 56, 57

ART Aligned Rank Transform. IV, 29, 36, 56, 57

CPT Continuous Performance Task. 9, 11

HMD Head Mounted Display. III, 1, 6, 14, 16, 20, 21, 23, 25–28, 48

IVR Immersive Virtual Reality. III, 1, 2, 5–11, 14–16, 18–21, 24, 27, 48, 51, 53–55

kNN k-Nearest Neighbour. III, 31, 42, 54

OOI Object of Interest. III, 17, 18, 29, 30, 32, 33, 35, 36, 38, 40–43, 46, 48–50, 52–55

SVM Support Vector Machine. III, 33, 44–46, 51, 53, 60, 61

t-SNE t-distributed Stochastic Neighbour Embedding. I, 29, 40, 50

VR Virtual Reality. 6, 8, 9, 14–16, 20

1 Introduction

Attention has been argued, is a key aspect of learning. It is not only the process that helps us to filter between relevant and irrelevant information in our everyday life. Attention is also necessary for learning and knowledge construction. It can be understood as the allocation of limited cognitive resources that allows us to build mental constructs such that information can be integrated into already existing memory structures (Brünken & Seufert, 2006). Therefore, attention plays an important role in education and is necessary to understand how children should be educated.

Most of children's learning takes place in schools, specifically in classrooms, where children spend many hours a day. In these situation, they are not only exposed to critical developmental experiences but the classroom is also a place for social dynamics and interactions (Hamre & Pianta, 2010). Therefore, it is important to understand how children learn in a classroom, but also to investigate social effects in this environment, not only with regard to children's learning outcome.

In scientific practice it is often difficult to evaluate such classroom effects due to several influencing factors that do not allow research with standardized experimental conditions. A promising approach to overcome this problem is Immersive Virtual Reality (IVR) (Blascovich et al., 2002). With this technology, participants can explore a 360° virtual environment by wearing a head-mounted display (HMD) headset. On the one hand IVR environments ensure the exact same experimental condition for all participants and it allows to introduce specific manipulations without disturbance factors. On the other hand it encourages participants to behave naturally in a real and physical way. This allows the design of a virtual field study in a controlled and manipulable environment.

An IVR classroom provides a habitual familiar learning environment for children. It is possible to introduce a teacher as a learning instructor as well as virtual peer learners, who simulate social dynamics in a classroom. Furthermore, different other manipulations can be made in terms of the visual design, as well as participant's sitting position or the specific behaviour of the virtual peer learners.

Since little is known about how children experience and explore IVR class-

rooms (Bailey & Bailenson, 2017), we propose to investigate children's visual attention in this virtual environment. With an integrated eye-tracking device, it is possible to collect children's spatial gaze data during an experiment and to analyse gaze-based attention with regard to the following questions:

- How do students behave during a virtual lesson?
- Towards what do they turn their attention?
- Do they only pay attention to the lesson content or do they give attention to other incidents?

The answers to these questions can hopefully provide a systematic understanding of how IVR classrooms can be used for educational research and how a IVR classroom should be designed for this purpose in future experiments and in practice. This exploratory work can be seen as a first step towards understanding children's visual attention in a IVR classroom.

2 Theory

2.1 Understanding attention

Attention can be defined as the selective process that guides us through the world (Driver, 2001). Every split second we are confronted with an enormous quantity of sensory information. Selective attention facilitate efficient information encoding by selecting certain sensory information while ignoring others (Awh et al., 2006). It has been argued that this selective process is necessary due to the limited capacity of our brain to process information (Carrasco, 2011). We are only able to attend to a limited number of items simultaneously (e.g. Cavanagh & Alvarez, 2005) and sustaining or maintaining attention on relevant tasks over a longer period of time is exhausting due to mental fatigue (e.g. Guo et al., 2016).

Early works of Broadbent (1958) or Treisman & Gelade (1980) have established the idea that different stimuli compete for limited resources, which is also supported by evidence from neuroscience and behavioural studies (e.g. Beck & Kastner, 2009). The cognitive process of attention is also related to the concept of working memory, where attention can be seen as a gatekeeper for biasing our encoding towards potentially relevant items (Awh et al., 2006). Even though the process of information encoding in our brain is a covert, some aspects of attention can be observed from outside by analysing people's behaviour (Goldberg et al., 2019). Concerning the information encoding process that drives people's attention behaviour, two processes can be distinguished.

Attention can be described as a bottom-up as well as top-down process. The process of shifting our attention towards a potentially relevant object or feature can be described as bottom-up mechanism. By processing incoming sensory information, we shift our attention rapidly and involuntarily towards a salient object or feature. An auditory cue, for example a warning signal, shifts our attention immediately towards the relevant target. But attention is also influences by top-down mechanisms. Attention can be biased by our need for certain information (Connor et al., 2004; Beck & Kastner, 2009). For example, in situations where we are keen to get social information we will draw our attention towards other human beings, where we expect to get that kind of information.

Therefore, analysing people's attention can give us information about their

need for certain information due to their attentional bias towards specific objects and features. Despite individual differences in the distribution of attention, people's attention behaviour is also similar in many situations, because of learned social behaviour and from shared experiences (Tomasello, 1995).

2.1.1 Attention and Learning

Attention and learning are codependent structures. On the one hand attention is the selective mechanism that enables us to learn things and on the other hand we learn how to pay and sustain attention. The role of attention in people's learning reaches from basic mechanisms of filtering information for memory consolidation (McCallum, 2015) up to complex mechanisms of knowledge construction or problem solving (Rouinfar et al., 2014). Therefore, attention can be seen as a key factor for learning (Brünken & Seufert, 2006).

In most of the educational research studies attention is measured via scales in a survey after the experiment, with questions like 'I pay close attention to how I do things compared with my classmates' (e.g. Hasenbein et al., 2020). But it is questionable to what extend a student is able to remember and evaluate his or her attention level after a learning period. These surveys can only give an average impression of one's attention level.

Specifically, if we look at children's leaning and education, attention can vary over time. Sustaining attention over a long period of time is exhausting and requires cognitive resources. Therefore, we have to assume that sustaining attention can only be guaranteed for shorter time periods (Wilson & Korn, 2007). This means, a student could be attentive most of the time, but in a critical moment, which is maybe necessary to understand a specific topic, he or she pays attention to something else. Participation and involvement can monitor children's attention (Sezer et al., 2017) but also can their attention be distracted, in a sense that other occurring stimuli pressure them to focus on something else. For example, Lundqvist & Ohman (2005) have shown that emotions, specifically facial expressions influence our attention behaviour and attract our attention towards negative stimuli. Such mechanisms need further investigation especially to analyse and understand attention in real life scenarios and they need to be taken into consideration, when we are exposing children to a certain learning

environment.

2.1.2 Education in Classrooms

The classroom has been understood as a central learning environment for children. It has been examined as children's predominant learning environment. Children not only spend many hours a day in a classroom for the purpose of education, but it is also a place for social dynamics and social interaction. It has been argued that a classroom, with all its entities and structure, can have a significant influence on student's academic outcome and children's classroom experience contributes to their social, cognitive and academic development (Hamre & Pianta, 2010). For example a prominent classroom effect is known under the name 'Big-Fish-Little-Pond' effect and is describing a negative correlation between the perceived academic performance level of a class and student's individual academic self-concept (Marsh, 2005). This means that children do not only acquire pure knowledge in a classroom but also encode social information, which can have an influence on themselves and on their learning experience.

A major problem of investigating classrooms effect is that it is not possible to conduct a classroom study under standardized experimental conditions. Expensive studies with a large sample size are necessary to average out all potential confounding variables in a real classroom environment (cf. Seidel & Shavelson, 2007). These factors can reach from different teaching instructions (Emmer & Stough, 2001), different behaviour of the peer learners up to different designs of a classroom or student's seating location (MacAulay, 1990). All these factors can, but don't have to influence children's attention and learning. So far little is known about these effects, why it seems inevitable to investigate them, but also focus on providing a framework for standardized classroom testing in a controlled environment.

2.2 Immersive Virtual Reality

Immersive virtual reality (IVR) is a technological development that places a person into a virtual 3-D environment, which is delivered by a display system presenting computer generated images. IVR systems present perspective-projected images, which are shown with correct parallax (Slater & Sanchez-

Vives, 2016) according to a person's field of view in the virtual environment. This leads to an experience of presence, the person's feeling of actually being in the virtual surrounding. A necessary requirement to experience presence is the immersion created by IVR systems. "Immersion is a psychological state characterized by perceiving oneself to be enveloped by, included in, and interacting with an environment that provides a continuous stream of stimuli and experiences", (Witmer & Singer, 1998).

In IVR, the virtual environment is presented to the user by a so-called head mounted display (HMD). It is a headset a person is wearing, with two displays placed in front of a person's eyes (one for each eye). Each display delivers a computer generated image showing a perspective with respect to the position and orientation of a person's head in the 3-D virtual environment. Therefore, the images on screen update when a person is moving or rotating his or her head and is showing the person's field of view in their virtual surrounding according to the position and orientation of the HMD. From a user's point of view, the presentation of the virtual images seem to align naturally with how they experience visual exploration in real life. Additionally, most HMDs come with integrated headphones, which allows the user the experience the virtual environment with two senses: seeing and hearing. This leads to the visceral feeling of not only watching an artificial environmental, but physically being in the simulated world (Slater & Sanchez-Vives, 2016).

Technological developments over the past years led to virtual environments in high resolution and a precise adjustment of the virtual images according to people's head movement and orientation. These developments in software and hardware allow the use of IVR for a variety of tasks and low cost application tools are available which can be used in science and practice (Blascovich et al., 2002).

2.2.1 Virtual Reality in Education and Research

IVR in Research. The continuous progress made in VR technology led to immersive, dynamic three-dimensional simulations, which can be useful in research for a number of reasons. VR gives the opportunity to create virtual environments which can be used to simulate everyday experience. The usage and

functioning of IVR requires a user to display behaviour which is similar to real life behaviour (Nolin et al., 2016) and can be observed in real time. Testing participants in IVR is not only cost and time-efficient but allows to control for confounding variables (Adams et al., 2009). It guarantees that all subject's are exposed to the same experimental setting where researchers are able to manipulate only certain aspects of special interest. In comparison to testing in real life situations, IVR presents a controlled environment. For example, experiments with specific treatment conditions do not suffer from methodological flaws often criticised in real-world experiments (Rizzo et al., 2006; Nolin et al., 2016). As Blascovich et al. (2002) stated, IVR allows the perfect trade-off between realism, which can not be guaranteed in lab experiments and a necessary level of experimental control, which is one major issue in experimental field studies. These aspects show the potential of IVR application in scientific research and are probably the reason why IVR has been used in many different scientific areas including psychology, medicine and education (cf. Slater & Sanchez-Vives, 2016).

IVR in Education. There are many IVR applications that are used in educational practice, for example for experimental or discovery learning (Johnston et al., 2018). This also increased the interest in investigating and understanding IVR in education. The advantages of IVR for science mentioned above, can be used to achieve a systematic understanding when conducting studies in educational science. Several studies have investigated educational related topics using a IVR concerning classroom and behaviour management, emotional awareness or curriculum content (cf. Billingsley et al., 2019). Most of these studies conducted research with adults, for example with focus on teacher training. But little is known about the effect of IVR on children (Bailey & Bailenson, 2017). The existing research that used IVR to test children focused on clinical and medical research questions, for example on pain distraction, attention deficit hyperactivity disorder (ADHD) or other impairments. Little attention has been drawn to social dynamics and social interaction in IVR with regard to education and leaning of children (cf. Kamińska et al., 2019). We suggest, that especially in educational science, where researchers are interested in higher order, more complex behaviour up to social interaction, IVR can be seen as a promising tool. The investigation of such processes and critical aspects of IVR has been left out so

far.

The use of IVR devices also gives to opportunity to observe people's eye-movement and vision, due to integrated eye-tracking devices. The functioning and use of these eye-tracking devices are explained later, but need to be mentioned here for a specific reason. Since children are often the subjects of interest in educational studies, ethical issues occur when observing their behaviour for example by taking video recordings during a lesson. IVR does overcome this problem, since we are able to measure children's visual information directly without relying on pictures or images of the children itself. Therefore, IVR guarantees a privacy preserving method of exploring children's visual behaviour (Bozkir et al., 2019). For example, this gives the opportunity to investigate children's visual attention while exposing them to learning environments similar to real one's they are familiar with.

2.2.2 Immersive Virtual Reality Classrooms

Since children have learned what to look for and what to ignore in a real classroom, they should focus their attention properly and understand significant patterns regarding for example the behaviour of their peer learners or the instructions of their teacher (Piontkowski D., 1979). Such habituated attention behaviour should also be recognisable in an IVR classroom since children are exposed to an authentic and familiar learning environment. But in which way attention distributes in the virtual environment does need further investigation.

As mentioned before, IVR gives the advantage of investigating children's behaviour in a controlled experimental setting. The IVR classroom can be used as such a learning environment to investigate the effect of certain classroom manipulations on children, but also their engagement with social peer learners, their teacher or the content of the lesson. Confounding variables often issue educational research, for example classroom and teacher specific differences can be eliminated or controlled in an IVR classroom.

Only few studies have investigated such effects in a virtual classroom and have been used for research concerning attention and learning, but also did not primarily focus on social dynamics in their learning environment. To get an overview of how IVR classrooms have been utilised for research in the past, we

look at some examples.

One of the first studies concerning a virtual reality classroom was conducted by Rizzo et al. (2001). They built a virtual classroom for assessment and rehabilitation of children with attention deficits. These children had to perform different attention related tasks, while different classroom distractions were introduced (e.g. noise, activities outside the window). The authors proposed this VR classroom as a possible tool for investigating and training children with attention deficit disorders. Five years later Rizzo et al. (2006) did a review on the evolution of their virtual classroom environment and the various application that had been introduced over the past years. They also integrated head movement and eye-tracking as additional measurements to investigate children's distraction with focus on group differences between children with and without ADHD. More recent studies concentrated on the topic of hyperactive disorder and used the VR classroom to test attention distraction as well. For example, Adams et al. (2009) tested children's ability on a continuous performance task (CPT) that was presented on the classroom board in the VR environment. Bioulac et al. (2012) also tested ADHD patient's on their performance in time on task and argued that the VR classroom is a reliable method to test sustain performances over time. Later studies in this field used the VR classroom for example as a tool to compare the effectiveness of different attention tasks (Díaz-Orueta et al., 2014). Nolin et al. (2016) tested the reliability of IVR classrooms and extended the approach of testing selective and sustained attention with the CPT for a variety of clinical studies. Recently, Mangalmurti et al. (2020) expanded the analysis on CPT in VR classrooms by looking at the participant's field of view as a mediator between ADHD and associated cognitive anomalies. So far, all these studies focussed on attention in a VR classroom from a clinical perspective regarding deficits in attentional control. However, they did not focus on participant's attention in VR with regard to everyday experience, like social interaction.

Seo et al. (2019) introduced social interaction as an additional level to their analysis of CPT in a virtual classroom. A virtual teacher (or instructor) was included in one experimental condition, which gave instructions and advices to the subjects. Here the authors found out that participant's head movement jumped back and forth between the presented task on the board and the virtual instructor, indicating a social interaction of subjects with a virtual character. Un-

fortunately, they erased other virtual peer learners in their virtual environment and therefore left out one crucial aspect about real life classroom experiences. Also with regard to the design and arrangement of the virtual classroom, so far these studies did not investigated these aspects at all.

Blume et al. (2019) used a virtual classroom to investigate students learning depending on their sitting position. They found evidence, that students seated in the front, proximally to the teacher are better in learning a taught solution strategy than student's seated in the back of the classroom. They measured participant's reaction time in a performance test but did not analyse any eye-movement or visual attention measurements.

Bailenson et al. (2008) conducted four studies in a virtual classroom where they investigated transformation of social interaction due to different manipulation in the IVR environment. The first study contained attentional cues for social interaction of teachers with virtual peers, showing for example that virtual peers in the center were focussed more than virtual peers in the periphery. In the second study, they used the same IVR environment but now subjects were placed in one of the students seats and a virtual teacher gave a lecture. Here they varied the sitting position of the subject regarding the angle they look at the teacher. They computed the total time the subjects kept the teacher in their field of view and found out that students learn better when seated in front of the teacher than in the classroom periphery. In the third experiment, subjects were placed at different positions in the classroom and the classroom population was manipulated as a second variable. Here the authors investigated a positive effect on the learning outcome when subjects are placed in front of the classroom, but no significant effect regarding the population of the classroom. In the last experiment they wanted to investigate the social influence of peer learners on participant's task performance. Virtual peer learners behaved either positive or negative and they found that the behaviour of virtual peer learners influence the pattern of learning of the participants. Subjects were able to remember more details about the room when they were effected negatively from their virtual peers, and concentrated more on the lecture content when they were influenced positively. In their summary they also argued that social peer learners are necessary in a IVR classroom since student's normally also learn in social conditions and assumed that subjects would respond to virtual peers as they respond to human

peer learners.

In summary, as we can also see in Table 1, only few studies were interested in social dynamics in a virtual classroom. But more importantly, studies that utilized a virtual classroom did only in few cases focus on measuring participant's attention with eye or gaze related information.

Table 1: Studies using an Immersive Virtual Reality classrooms

Study	**Research Interest**	**Measurements**	**Eye or head movement measures**
(Rizzo et al., 2001), (Rizzo et al., 2006)	Distraction in IVR, investigated ADHD syndrome	Reaction time, task performance	Head movement and eye-tracking
(Bailenson et al., 2008)	Attention of speaker towards listeners, learning outcome according to sitting position	Task performance	Eye-tracking objects in subject's field of view
(Adams et al., 2009)	Classification of patients by performance in the IVR classroom	Reaction time on CPT	No visual attention or eye-tracking measured
(Bioulac et al., 2012)	Sustaining attention in noisy environment	Reaction time on CPT task	No visual attention or eye-tracking measured
(Díaz-Orueta et al., 2014)	Effectiveness of different attention tasks	Reaction time, correct responses	Head movement
(Nolin et al., 2016)	Testing attention and inhibition in clinical research	Reaction time, correct responses	Head movement
(Seo et al., 2019)	Attention performance task with teacher instructions	Reaction time, correct responses	Head movement
(Mangalmurti et al., 2020)	CPT with field of View as mediator	Reaction time, correct responses	eye-tracking in the field of view
(Blume et al., 2019)	Influence of sitting position in classroom; ADHS syndrome and learning	Reaction time and correct responses on a bisection task	No visual attention or eye-tracking measured

Notes. Overview over studies using an IVR classroom, stating their research interest, the measurements they used to investigate their research question and if they used eye or head movement measuremnts to analyse participants experience in the IVR.

2.3 Measure Attention

To measure individual attention behaviour as an indicator for learning or encoding of social information, we can focus on people's visual behaviour. Investigating people's visual attention can be promising for various reasons.

2.3.1 Attention and Vision

Overview. Visual attention has become a prominent field of study over the last decades. Searching for articles with 'Visual Attention' in the title on PubMed returns over 1600 results. It can be argued that analysing visual attention is highly relevant since vision is mostly our primary sense and peoples eyes are a rich source of information. But also the convenient way visual experiments can be conducted and the current progress made in analysing people's eye-movement and vision might influence the popularity of this field (Hutmacher, 2019). To roughly characterize some relevant aspects of visual attention we can start to distinguish two main types: Overt and covert attention.

Attention can be overt when it coincides with a person's eye-movement towards that location. In overt attention, we assume that a person's focus of attention is similar to where his eyes are fixating on. This straightforward similarity between attention and vision can also be motivated by the anatomy of the eye. Since the eye is a foveal system, perfect acuity is only guaranteed in a small central part of the retina (Jacobs, 1979; Kübler et al., 2017). The direct line connecting the fovea with the outside world is also referred to as the visual line-of sight. For example, this is also the only part of the visual field where we have a resolution high enough to do tasks like reading or recognizing faces (Lodge & Harrison, 2019). Beside overt attention, we are also able to attend to areas in the periphery, without directing our gaze towards it. This phenomena is called covert attention (Carrasco, 2011).

Additionally, there are three categories which allow us to classify different aspects of visual attention: Feature-based attention, object-based attention and spatial attention. Feature-based attention addresses the encoding of feature information like orientation or motion direction to guide our attention. For example, when we are looking for a cap, it could be more helpful to focus on the color of the cars due the the yellow signalling color of most caps. Scanning the scene

for relevant feature information is mostly done covertly. Object-based attention is guided by the structure of an object and is focussing more on the stimulus of an object that we pay attention to. Spatial attention referees to the visual behaviour of people moving their eyes to relevant location in space. It considers the encoding of spatial information with relation to the ability to guide our attention in the world and selectively focus on relevant objects while ignoring others. Spatial attention can be both overt, when the eye-movement coincides with the focus of attention or covert when attention is drawn to relevant location in the visual field without moving our eyes (cf. Carrasco, 2011). In this work, we will focus on overt spatial attention due to the fact that we use gaze direction as an indicator for attention. However, we need to keep in mind, that these different aspects cannot be distinguished completely. They interact with each other, influence each other or build taxonomies of guiding our attention in the world.

To investigate different types of visual attention, the prominent tools of choice are eye-tracking devices, which observe and analyse people's eyes in real time and are able to generate valuable information.

Eye-Tracking. Eye-tracking has become a prominent tool of investigating peoples eye-movement and vision. It is used for a variety of tasks not only in science but also in practice. It's origin goes far back to the beginnings of psychological science. For example, earlier psychologists already measured basic eye-movements with the help of analogue electronics (Kowler, 2011). With the development of computers, eye-tracking can be used to measure eye-movements and vision with high accuracy in real time. Today the general approach of eye-tracking is, that a camera placed at a fixed location records a person's eyes. The information from the recording are further processed to calculate for example people's pupil diameter or the direction a person is looking at. This method can be applied because of the special anatomy of our eyes. In contrast to other animals, our darker coloured pupil, which shows the direction of a person's view, is surrounded by white matter, which makes it possible to identify the visual direction of a person by comparing the position of the pupil according to the position of the head (Singh & Singh, 2012). With this information eye-tracking devices are able to calculate the gaze direction of a person, which can be collected with high accuracy and in real time to analyse people's visual behaviour.

Eye-tracking devices are used to investigate people's eye-movements, when they watch stimuli on a screen, but can also be used in IVR to investigate peoples visual behaviour. Using eye-tracking in IVR also requires information about other variables, which are relevant to analyse eye-movement and vision.

Head Mounted Display Orientation. In IVR the field of view of a person is limited by the head mounted display. Only by moving their head, people are able to observe the whole 360° virtual environment. Therefore, the orientation of the head must be measured. This is normally done by stating three angles referring to three types of rotation a head can perform: Pitch, Roll and Yaw. Pitch is defined as the angle a person is looking up or down, roll is defined as the angle of the head leaning left or right (more precisely the rotation of the head according to an axis pointing orthogonally away from a person's face) and yaw is defined as the rotation of one's head left and right according to an axis pointing up vertically from a person's head. These three measures are enough to exactly locate the orientation of the head in space. The images shown in the VR are automatically aligned with the orientation of the head, but these measurements can also be used to analyse a person's visual exploration of the virtual scene. Additionally, since in many IVR applications people are able to move in the virtual surrounding, the position of a person is also tracked by the HMD device and aligned with the virtual environment presented to a person.

Using eye-tracking devices for on-screen tasks or combining HMD orientation and eye-tracking in IVR, leads to the detection and classification of different eye-movement features.

Eye-Movement Features. Several eye-movement features have been detected over the recent years. Since we do not focus on these features in this work, we briefly mention the important ones. Two prominent features to be distinguished are fixations and saccades. Fixations are detected when the gaze of a person stays located at a particular place on screen and no gaze movement is detected. When detecting fixations in IVR, we also need to consider the head movement. Therefore, a fixation is identified, if there is only little head movement of a person and additionally the velocity of the gaze vector stays below a certain threshold for at least a hundred milliseconds (Salvucci & Goldberg, 2000). Saccades

on the other hand are accurate ballistic eye-movements with a high velocity for a short period of time, identifying a shift in the visual field of view of a person or a repositioning of the fovea to another location (Singh & Singh, 2012). Certain visual behaviour has been classified depending on a person's intention or task in certain situations, which can be generalized beyond individual predispositions, for example patterns like visual searching (Nakayama & Martini, 2011). Therefore, eye-movement features are used in a variety of research questions in different scientific fields (cf. Kowler, 2011; Lai et al., 2013).

Despite the analysis of specific eye-movement features, another analysis can be done combining eye-tracking and head orientation information. So far the presented techniques of analysing a visual behaviour have not taken into account the actual location a person has looked at on the display or the object which has been observed by a person.

2.3.2 Gaze-Based Attention

Spatial Gaze Location. To locate where the measured gaze does hit the display, regardless of using an on-screen eye-tracking or a VR, a method called ray casting is used. Ray casting can be considered as using the gaze direction (gaze vector), calculated by the eye-tracking devices, as a laser pointer that invisibly points from people's eyes towards a certain location or object (Pietroszek, 2018). Therefore, we need to know the location of a person's eyes, their distance to the display and the direction of the gaze vector given as 3-D coordinates. Having these information allows us to calculate where the gaze hits the display and to track the 2-D gaze location on the display.

To know the location, where the gaze hits the display, can be interesting when we want to know which areas of the visual field are observed the most, or even to detect which object has been focussed by a user. This information can give valuable insights into a person's visual attention and has frequently been used to investigate different research topics (e.g. Cutrell et al., 2007; Kübler et al., 2017; Agtzidis et al., 2019).

When using eye-tracking for on-screen applications, it is relativity easy to assign a gaze location on screen to a certain object, since users only see a picture or movie, where objects appearing on screen can be labeled manually. When using

eye-tracking applications in an IVR, more variables need to be taken into consideration to extract which object is being observed during the virtual experience. The issue occurs since user in an IVR can explore a 360° environment and even if we know the 2-D location on the display, we do not know which part of the virtual environment the user has looked at. So one approach is to just record videos of the user's field of view, annotate the gaze location to video frame by frame and label the visual objects afterwards.

Another approach is to use information about HMD orientation and the user location in the virtual space to determine the visual field of a person at a specific time point. Here one important aspect is that if we use IVR that is created by a development environment (e.g. engines like Unreal or Unity3D) we not only present a 360° video but an actual virtual space subjects can move freely. This movement takes place in the virtual space according to a virtual coordinate system that locates the positions of all objects and the position of the user in the VR. Hence, every object in such a virtual environment has a spatial location according to a global coordinate system of the virtual environment.

As aforementioned, gaze direction is given as a gaze vector. Therefore, it is possible to combine the information about the location of a user in the virtual environment with the head orientation and gaze direction to ray cast the gaze vector throughout the virtual space and investigate at which location the gaze of a person hits certain objects in the virtual environment.

Gazed Object. The lengthened gaze vector is sent through the virtual environment, starting from the location of a person's eyes. The object hit by this vector in the virtual space is called the gazed object. Technical details about how to identify the gazed object in a virtual environment are given in Subsection 3.2. By extracting the gazed object during an experiment, it is possible to obtain gazed object information in each time step. Therefore, we are able to know which and how long an object has been focussed by the subjects.

This collection of the gazed object, does align with the idea of overt spatial attention. When analysing which object in the virtual space has been observed at a specific time during an experiment, we are measuring overt visual attention (Kübler et al., 2017). We assume that the gazed object, which is focussed directly by the subject, is the one they pay attention to. Regarding the duration spent on

a specific gazed object, we can investigate how much attention participants paid towards that object in comparison to other objects in the environment.

Analysing the Object of Interest. We are interested in the time spent on specific objects, we define as the objects of interest (OOI). We use information about the gazed object to analyse when and how long subjects have looked at an OOI. According to the different eye-movement features explained before, only looking at the total time spent on an object does not tell us if a participant has actually overtly attended towards this object. During a saccadic movement, a lot of objects could be gazed without really being realised by the participant. We need to assure that the participant spend at least some time on an object. Therefore, we introduced a minimum threshold that guarantees us that subject's spend at least a certain amount of time on the same object before we classify the time spent on an object as overt attention towards an object. With this approach, we were able to investigate visual behaviour over time, but also to measure similarities and differences in visual behaviour between different events, tasks and for different experimental conditions. For example, we are able to accumulate the time spent on specific OOI during the experiment to see if there is a significant time difference between certain experimental conditions. This gives us information about how subjects explore the virtual environment.

Another way to analyse gazed object data is to conduct a scanpath analysis. Usually, scanpaths are sequential representations of fixations (Noton & Stark, 1971). In our case, we build spatiotemporal sequences for the gazed objects, which show patterns of eye-movements between the OOIs. A visual scanpath can be either illustrated as a sequence of letters, where each letter corresponds to an OOI or presented in a transition matrix showing the number and direction of transitions from one object to another. Interpretation of letter-based sequences can be derived from the analysis of DNA sequences. Same as a DNA sequence of amino acids does not give us direct information about the utility and functionality of a cell, a visual scanpath does not reveal the cognitive state of a person directly. But it is possible to measure similarities and differences in such sequences. As we have argued before, that a person's eye-movement and vision gives us information about his or her cognitive and mental states, a visual scanpath can be seen as an indirect measurement of a person's underlying cognitive

model (McIntyre & Foulsham, 2018). We can translate the measurement techniques used in bio-informatics to characterize similarity of DNA sequences to analyse similarities and differences in a person's visual behaviour for a variety of questions. Slightly different interpretations can be derived from the transition matrix method, because in this representation we lose information about the temporal order of a sequence. But it is possible to store the number of transitions not only for pairs of transitions but also for longer series, e.g. attention to peer learners then to teacher and then back to the peers. These transition features can be used to analyse differences in visual behaviour, trying to classify for different experimental condition and to investigate important transition features. (cf Kübler et al., 2017). In this work we customized the ScanMatch algorithm (Cristino et al., 2010) as a string-based method and the SubsMatch algorithm (Kübler et al., 2014) as a matrix-based method to analyse gaze patterns from the gazed objects.

2.4 Research Questions

We argued, that IVR can be seen as a promising tool to investigate children's behaviour in a classroom situation and to measure potential educational classroom effects. Due to the immersion and presence of the participants in the IVR we can use IVR to analyse real life behaviour in a controlled experimental setting. As we have seen, only few studies have investigated IVR classrooms in the past and did not primarily focus on social dynamics or student teacher interactions.

In this study we investigated children's visual attention in an IVR classroom with different experimental conditions concerning the sitting position of the participant, the interaction level of the virtual peer learners and their virtual appearance. For each participant, we measure the gazed objects during the full run frame by frame. By selecting specific OOI, for example the teacher or the virtual peer learners, we are able to analyse overt spatial attention in the presented learning environment.

The gazed object data had to be collected separately. Therefore, we also present a method how this can be done in a virtual environment created in the

Unreal Game Engine[1]. This algorithm can be used to directly collect the gazed object during the experiment in real-time and might be a useful implementation for future experiments in the field.

Since the topic of overt visual attention in an IVR classroom is almost unexplored, this work can be seen as an exploratory research. Therefore, we are not able to formulate confirmatory hypothesis. We hope that interesting questions arise from our investigations, which can built the basis for future research. However, the following contributions can be made:

1. A first step towards a systematic understanding of children's overt visual attention in IVR classrooms with regard to visual behaviour and social dynamics.

2. Analysis of different IVR classroom manipulations considering future IVR classroom design choices for research and practice.

3. Application of different established methods, to detect possibilities and limits of analysing children's gaze-based attention in an IVR classroom.

[1]https://www.unrealengine.com/

3 Methods

3.1 Experiment

All data used in this analysis was collected from an IVR classroom experiment, in which children participated in a lesson about computational thinking. Specific manipulations were introduced to the classroom, resulting in a between design study with strict experimental control.

Sample, Procedure and Study Design. After approval from the ethics committee of the University of Tübingen and regional educational authorities, 381 children from sixth grade participated in the experiment. Due to incorrect eye-tracking calibration and hardware problems 31 participants had be excluded beforehand and another 61 participants data could not be used due to their low eye-tracking ratio (lower than 90%). Therefore, we used 289 data samples for our analysis from students (143 female, 146 male), whose age was between 10 and 13 ($M = 11.52$, $SD = 0.56$). We asked for their experience with VR (e.g. in video games) and 41% stated that they had no experience with VR before, 37.5% had used VR once and 21.5% had used VR from time to time.

The students participated in the experiment in groups of 10, for sessions with a length of 45 minutes on average. The experiment was conducted in a quiet room at the students' school. First, participants filled out a paper-based pre-test, including demographic and basic personal information about the student. In the second part, the participants experienced the IVR lesson by wearing the

(a) HMD headset HTC Vive Pro Eye[2].

(b) Virtual classroom overview

Figure 1: HMD headset and classroom design

[2]Picture from https://www.tobiipro.com/ de/produkte/vr-integration/

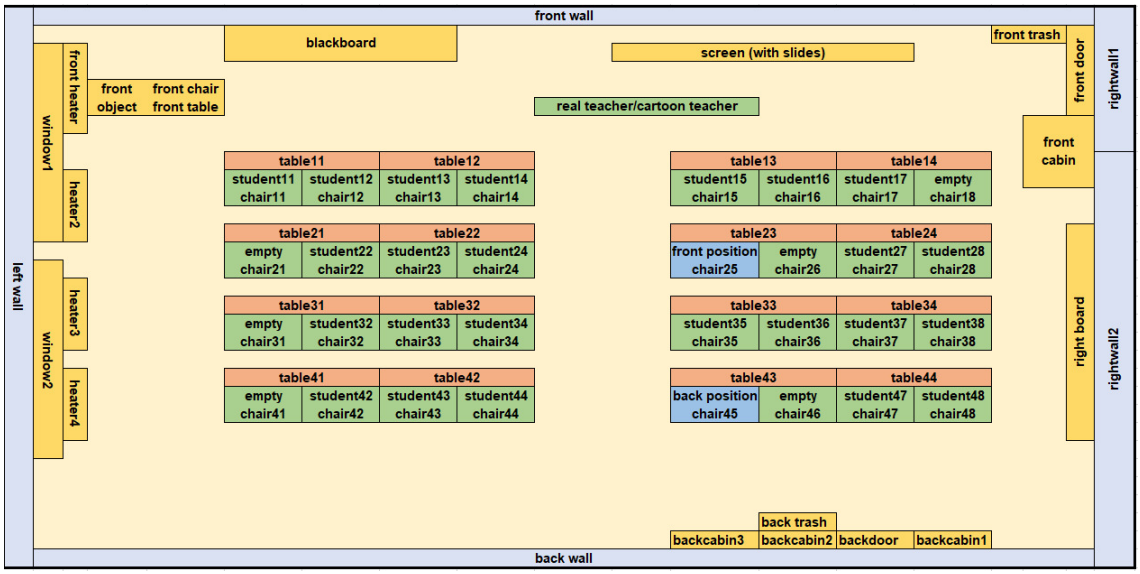


Figure 2: Arrangement of objects in the IVR classroom

HMD (HTC Vive Pro Eye, see Figure 1a). All students in one experiment session started the IVR part at the same time, after a successful calibration of the eye-tracking. The lesson started by pressing enter, which was also marked in the collected data sets. Research assistants, which also helped the students with technical issues regarding the use of the HMDs, introduced the virtual lesson as a learning experience. However, they did not mention any aspect regarding the design or social dynamics in the virtual classroom. After the virtual lesson, participants filled out a post-test questionnaire regarding different psychological measurements followed by a debriefing.

Classroom Design and Structure of the IVR Lesson. The virtual classroom, participants were seated in, was inspired by real classrooms (blackboard, screen, tables, chairs, windows, etc.) and consisted of four rows and two columns of tables, positioned in a way that all peer learners look straight forward to the front of the classroom (see Figure 1b). On the screen in front of the classroom a presentation was shown during the lecture and the virtual teacher was moving in front of the class. Beside other objects, like cabinets or trash cans, 24 virtual peer learners were distributed across the seats. The arrangement of the virtual objects and their names are depicted in in Figure 2. The virtual environment was designed and rendered with the Unreal Engine version 4.23.1. The movement of the virtual characters were designed to act naturally with animations like looking around, sliding on their chairs or dangling with their feet.

The whole virtual lesson took approximately 14 minutes and can be divided into four phases. In the first phase ($\approx$ 3 minutes), the female teacher was telling the participants to look around in the classroom and to take place. Then, she walked out of the classroom and back in after a short time. For the rest of the first phase, the teacher gave an *introduction* into the topic of computational thinking (Weintrop et al., 2016). In the second phase ($\approx$ 4.5 minutes), the teacher gave some *knowledge input* about the topic, but also interacted with the virtual peers via questions. The third phase ($\approx$ 5.5 minutes) is conducted as an *exercise* part, where the students had to choose the correct answer from four presented options. The teacher was checking the class performance and was asking the students for the correct option. In the last phase ($\approx$ 1.5 minutes), the teacher gave a quick *summary* and told the subjects at the end, that the lesson was now over. Social interaction between the teacher asking questions and peer learners raising their and answering in some cases, was present in the first three phases.

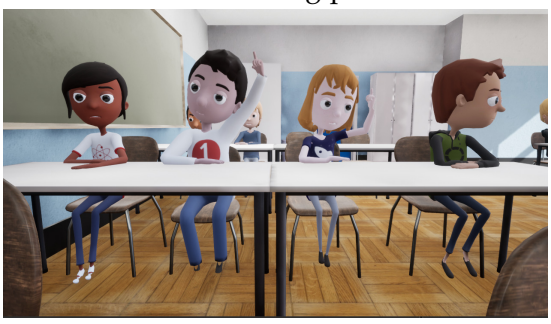
Experimental Conditions. Three different manipulations were implemented in the virtual classroom. First, the sitting position of the participants in the virtual classroom was manipulated. Subjects were either be placed in the second row (front) or in the fourth row (back) of the classroom. Second, the hand raising behaviour of the virtual peer learners was manipulated. After the virtual teacher asked a question a specific proportion of virtual peers raised their hand, to potentially answer the question. Despite that, the virtual peer that gave the answer was the same in all experimental conditions, only the number of virtual peers raising their hand was manipulated. Either 20%, 35%, 65% or 80% of the virtual peers raised their hand after a question from the teacher, simulating different level of engagement of the peers during the lesson. The third manipulation addressed the visual appearance of the virtual characters in the classroom. The avatars of teacher and peers were designed either more realistic or more cartoon-like, resulting in two additional manipulation regarding the avatar style. Different conditions can be seen in Figure 3.

This resulted in a $2x4x2$ between subject design, where each participants were randomly assigned to one of the 16 experimental conditions, leading to approximately 18 ($SD = 5.06$) participants per condition. A random number generator was used to select the experimental condition for each run and to guarantee a

(a) Back sitting position

(b) Front sitting position

(c) Realistic avatar style



(d) Cartoon avatar style

Figure 3: Classroom manipulations

random distribution of conditions within an across test groups. Additionally, students could freely decide which seat they take in the experiment room, without knowing the experimental condition.

Apparatus and Data Collection. For the experiment the HTC Vive Pro Eye was used with a refresh rate of 90 Hz and a field of view of 110°. The HMD consisted of two displays, one for each eye, which were set to 1440 × 1600 pixel. The eye-movement data was collected by the integrated Tobii eye-tracker with a 120 Hz sampling rate. All data was stored in separate data sets for each participant only labelled by an anonymous user ID.

Since valuable information about the gazed objects were not collected during the run, we carried out a separate data collection process after the experiment. Therefore, we used the already collected data, coded the necessary ray casting algorithm and implemented that into the existing virtual environment. We used the same Unreal Engine version and the original project to ensured that arrangement and structure of the lesson was the same as it was during the experiment.

3.2 Ray Casting for the Gazed Object

During the experiment, various data was collected regarding participants' location in the IVR, their head position or their gaze. Some information was not collected and needed some additional work afterwards. For instance, to extract information about the gazed object we implemented an algorithm , which we ran after the experiment. With this algorithm we were able to collect the gazed object, but also some other relevant information which were not collected beforehand (i.e. 2-D coordinates of gaze hit on screen, 3-D location of gaze hit in the virtual environment, distance to observed object).

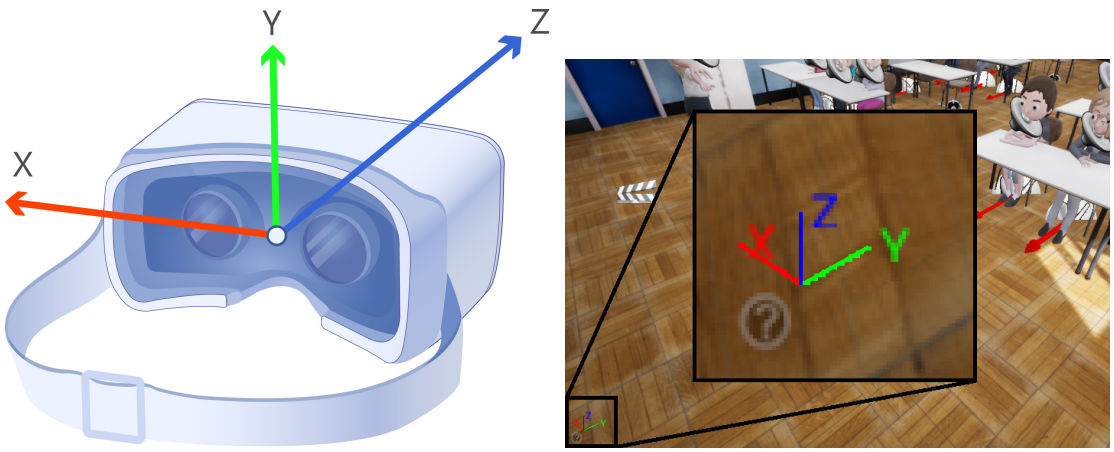
We used a method called *ray casting* (Roth, 1982), which is also used to calculate the gaze point from eye-tracking devices in the first place (see Tobii Tech, 2020). The idea is to forward a persons gaze vector, given in 3-D coordinates and to calculate which object the gaze hits. This can be imagined as an invisible beam following a straight line from a person's eye to a certain location in the environment (Soret et al., 2020). Therefore, it is necessary to know the location of the person's eyes in space, the location of the objects of interest in the environment and the direction of the gaze (gaze vector). A gaze hit on a specific object is given if

$$\{v_{location} + (v_{direction} \cdot k) \mid k \in \mathbb{R}\} \cap s_{object} \neq \emptyset \tag{1}$$

where $v_{location} \in \mathbb{R}^3$ are the coordinates of a person's eye location, $v_{direction} \in \mathbb{R}^3$ is the gaze direction and s_{object} the set of coordinates describing the surface of the object.

As gaze direction, we used the combination of the gazes from both eyes. The HTC Vive eye-tracker automatically calculates the combined gaze direction as the equidistant line between the gaze direction from the left and right eye. Since a fixation point in the environment is the direct intersection of the gaze vectors from both eyes, we use the combined gaze vector as the direction of one's gaze. Therefore, we used the normalized combined gaze direction from the experiment data as a person's gaze direction.

To prepare these data for further use, we needed to exclude missing gaze values in the gaze direction variables. Missing gaze information might occur due to the blinks or unsuccessful detection of the pupil. Since a blink produces a



(a) Local coordinates: HMD coordinate system[3].

(b) Global coordinates: Coordinate system of the Unreal Engine.

Figure 4: Local and global coordinate systems

frequent but only short period of missing gaze data, we were able to interpolate these periods. We assumed that during a blink, the gaze direction does not change rapidly. Therefore, we used a linear interpolation (polynomial interpolation with degree one). After the interpolation, all gaze values were between -1 and 1, linearly connecting the gaze values before and after the blink period. All other necessary variables did not contain any missing values, thus we used these data directly to perform the ray casting.

To do the ray casting as explained in equation 1, we made some adjustments. We placed a virtual camera actor into the virtual environment, simulating the participant during the experiment. Position and orientation of the camera were adjusted by the position and orientation variables from the HMD, collected from subjects during the run. This could be done with predefined functions (SetActorLocation and SetActorRotation). Furthermore, we could use some other preprogrammed functions, prepared for users working with the Unreal Engine Blueprint (Epic Games, 2020). First of all, it was possible to get the vector pointing out orthogonally away from the camera (GetRotationXVector). With this function, we got the x-axis (x vector) of the local coordinate system of the camera in world coordinates. A local coordinate system is the one that is adjusted to the rotation of an object (e.g. adjusted to the camera). It changes if the object rotates (see Figure 4a). World coordinates are relative to a fix coordinate system that is locating all objects in the virtual environment (see Figure 4b).

[3]Picture from http://developer. tobiipro.com/ commonconcepts/coordinatesystems.html

So far, this x vector, that was pointing away from the camera, only stated the direction of the camera but was not influenced by the gaze direction. Therefore, we calculated the angle between the gaze vector and the x vector and moved the x vector such that it was pointing in the right gaze direction. Since the normalized gaze vector was also given in local coordinates according to the HMD orientation, we calculated the angles (pitch and yaw) to rotate the x vector.

In general, for two vectors v_1,v_2 we can calculate an angle α in degree by

$$\alpha = \arccos\left(\frac{v_1 \cdot v_2}{|v_1| \cdot |v_2|}\right) \cdot \frac{180}{\pi}.$$

Having the gaze vector $g = (g_1\ g_2\ g_3)^T$ and the x vector in local coordinates $x = (1\ 0\ 0)^T$ we calculated the yaw rotation as the angle between x and $g_{flat} = (g_1\ g_2\ 0)^T$ and the pitch rotation as the angle between g and g_{flat}. Since an angle has no direction, we also needed to consider the direction of the rotation from the gaze vector (up or down and left or right).

After rotating the x vector by the two angles, it reflected the actual gaze direction of a participant in world coordinates and could be used as input for the ray casting. Once again, we used a predefined function from the Unreal Engine (LineTraceByChannel), which contained a variety of output information for the object that was hit by ray cast gaze vector. It was possible to output the name of the hit object, the distance between start location and hit of the gaze or the location of the hit object in the virtual environment.

One part that has been left out so far, is how the ray casting function gets information about the surface of an object s_{object}, as mentioned in Equation 1. The visual appearance of an object in a virtual environment is not the same as having a physical shape. Beside the visual shape, we added physical shapes to all object in the environment, also known as colliders. It is an invisible mesh grid that approximates the shape of an object and describes it's surface, which can be used to detect the gaze hit.

To obtain the gazed object and other variables for a full experiment session, we applied the ray casting frame by frame. Therefore, we pre-processed the experiment data for each participant, imported these information into the Unreal Engine and re-ran the whole lesson for each participant. To import and prepare the eye-tracking data, we programmed a C++ script, since the Unreal Engine is

built on that programming language. To ensure that we had exactly the same arrangement of the virtual classroom as in the experiment, we used the time tracked during the experiment and jumped to the exact same time point in the virtual lesson. This guaranteed a one to one mapping between the collected gaze information and the adjustment of the virtual scene. This was important because for example the teacher was moving during the lesson and we needed to make sure that the objects in the IVR are placed at the location where they were standing at a specific time point during the lesson. This ensured that the gaze, pointing in a certain direction, hit the object that was standing there at the exact same moment during the run.

3.3 Data Cleaning and Final Measures

Data Cleaning. After extracting the gazed object and other related information from the Unreal Engine, we cleaned the data and merged it with the already existing eye-tracking data from the original experiment according to the time variable. After merging the eye-tracking data with the processed object data, we had an average of 0.0028% of rows with missing object data. These missing values could occur due to rounding errors in the time variable processed in the Unreal Engine.

To determine the start of the experiment, a marker was set in the data set when enter was pressed to start the lesson. We used this as a starting point and removed all data before that maker. The end of the experiment was set at the time point when there were no gaze information available in the data from this point until the end of the run. Since the lesson ended after approximately 850 seconds, we cut all data after this time point. Then, the average time length for all session was 849.95 seconds with a standard deviation of 0.74 seconds. Thus, all trials have approximately the same length.

Measurements. The final data sets consist of measurements directly collected during the experiment or extracted afterwards from the ray casting.

- **Time:** This variable shows the time in seconds from the start of the virtual lesson calculated with the system time information collected during the experiment.

- **HMD position:** There are 3 variables describing the 3-D coordinates of the participant's head position in the virtual classroom, according to the world coordinate system of the virtual environment.
- **HMD orientation:** The variables pitch, yaw and roll are the angles of the head orientation and are measured in degree.
- **Gaze vector:** This is the normalized combined gaze direction given as 3 variables according to the local coordinate system of the HMD eye-tracker.
- **Gazed object:** The gazed objects are given as string variables showing the names of the observed object according to their object names in the virtual environment.

Gaze-Based Attention. From the gazed objects, collected frame by frame during the run, we did not consider all gaze on an object as attention. As aforementioned, we introduced a certain minimal threshold for time spent on an object. Only if subjects focussed on an object longer than this threshold, we considered that period as attention towards that object. The goal was to exclude gazed objects during saccadic movements, where subjects are not able to pay attention towards a specific object. There are different thresholds possible depending on different assumptions that can be made. For example, the threshold can be set to 100 milliseconds, because this often used as a lower bound for fixation detection. But also other threshold are plausible. Especially because we were interested in social information, a higher threshold would increase the possibility that we focus on periods where participants recognized the behaviour of others. Therefore, we investigated larger thresholds up to 1 second spent on an object. The algorithm that is selecting the time intervals on objects which are above this threshold, we will refer to as attention selection algorithm.

3.4 Analysis and Models

As aforementioned, we used different methods to detect possibilities and limits of analysing gaze-based attention. The analysis was mainly done in Python[1]. In this section we want to give a quick overview over the methods used in this work.

[1]Version 3.7.6 https://www.python.org/downloads/release/python-376/

3.4.1 Aligned Rank Transformation ANOVA

Data in human-computer interaction is often not suitable for established statistical method, because of its non-parametric distribution due to errors from devices or the experiment (Wobbrock et al., 2011). On the other hand, a necessary requirement to perform an analysis of variance (ANOVA) is that the data is independent, normal distributed and satisfies homoscedasticity (Sthle & Wold, 1989). Therefore Wobbrock et al. (2011) presented an approach to perform an Aligned Rank Transform (ART) on non-parametric data before applying the ANOVA. This method allows an accurate treatment of non normal distributed data for a full factorial ANOVA with main effects and interactions. They also provide a package[1] to perform the ART and the follow up ANOVA in R[2]. For main effect it is also allowed to perform a post-hoc t-test, if there are more than two conditions per category.

3.4.2 Visualisation with t-SNE

Experimental data is sometimes hard to visualize due to the large number of dimensions (large feature space) and the large number of samples. To get a first impression of the data and how it does cluster or separate one could use a visualization method that reduces dimension. In this work we used the t-distributed stochastic neighbour embedding (t-SNE) to visualize high dimensional data in a low dimensional space. The t-SNE algorithm transforms high dimensional Euclidean distances into conditional probabilities, were the similarity of two data points is the conditional probability (in a Gaussian distribution) that a neighbour point would be picked given the other one. A similar t-distribution is build for the low dimensional points. Then, t-SNE minimizes the sum of Kullback-Leibler divergences over all data points, to minimize the mismatch between the high and the low dimensional data points (Van der Maaten & Hinton, 2008).

3.4.3 Scanpath Analysis

Beside looking at the time spent on specific object of interest, we are also able to observe the visual transitions made between different OOI. For example, if a

[1] ARTool from https://cran.r-project.org/web/packages/ARTool/index.html
[2] Version 3.5.2 https://www.r-project.org/

persons attention goes from the teacher to the peer learner, we can count this as a transition between teacher and peer. There are tow ways to describe these transitions. We could either represent the transitions in a in a string, where each object gets a unique string, e.g. $teacher = t$ and $peer = p$. Then, a transition from teacher to peer to teacher would be coded as "tpt". For better illustration one can describe the transitions in a matrix, were we have one column and one row for each OOI and count the transitions from objects to object. In our example the transition matrix for our sequence would look like

from/to	peer	teacher
peer	0	1
teacher	1	0

.

Additionally, we are not only able to take transitions into consideration, but also to code longer fixation on the same object. Therefore we could either repeat the same string in the letter-based method, or fill up the diagonal entries in the matrix in the matrix-based method. For example (Cristino et al., 2010) suggested of to repeat strings for periods longer then $50ms$.

These two representations can be used for different analysis methods, because of the different properties they have. In the string-based representation the temporal order of the scanpath is preserved, while we loose temporal information in the matrix representation. On the other hand, when we analyse longer time sequences with a higher number of transitions, the sting representations increases, while the matrix representation remains the same size. It depends on the task and data, which representation is more suitable for further analysis.

ScanMatch Algorithm. Cristino et al. (2010) introduced a novel approach to compare fixation sequences. Their ScanMatch algorithm is a way to compare scanpath letter representations. It is build upon the Needlemann-Wunsch algorithm, which is used in bioinformatics to detect similarities between DNA or protein sequences. The basic idea is, that two strings are compared letter by letter and scored positively if the letters are matching or negatively if they mismatch. The sum over all scores gives a similarity score for the two stings (the higher the better). One question that occurs, is how to compare sequences with unequal length. Therefore, the Needleman-Wunsch algorithms includes a se-

quence alignment algorithm. This algorithm introduces gaps between the letters of the strings such that both sequences have the same length and align better. Since the gaps could be set at arbitrary positions, the algorithm searches for the alignment that maximizes the similarity score. After the alignment, we get a similarity score for each tuple of sequences.

The scoring can be manipulated by changing the scoring system. One could either define different values to score match or mismatch, but there is also a gap penalty introduced. If the gap penalty is positive in encourages the algorithm to introduce gaps for sequence alignment. if the gap penalty is negative the algorithm is optimizing towards an alignment with less gaps introduced.

Despite the equal length of two aligned sequences, the length between different sequences alignments could differ. The score that is given by the Needleman-Wunsch algorithm highly depends on the length of the sequences. To compare the scores between different sequences, we need to normalize for the length. When dividing each score by the length of the longer string of two compared sequences, it guarantees that the highest similarity score is always 1 and different scores can be compared.

For a set of strings the pairwise similarity score can be stored in a similarity matrix. This matrix can used to visualize the different scores in a heat map. Additionally, a k-Nearest Neighbour (kNN) classification can be performed by dividing the score sample into two classes (Guo et al., 2003). For example, we put the score of two sequences belonging to the same experimental condition in one class and the scores comparing sequences from different experimental conditions in the other class. Then, the kNN classifier searches for the k samples with similar scores and predicts the class label by majority vote.

This gives a convenient way to compare sequences from different experimental conditions. Beside that, the algorithm comes with some drawbacks. One problem is, that runtime increases quadratically when increasing the sequence length, which can be the case for scanpath sequences of longer time intervals. Another issues is that all sequences of one sample must be compared one by one leading to a total number $\frac{n \cdot (n-1)}{2}$ scores for n sequences. This can particularly be a problem, when we collect scanpaths from many different people.

Beside that the ScanMatch algorithm does not allow us the investigate which alignment of the sequence did lead to a certain similarity or dissimilarity. We

cannot determine which parts of the visual scanpath are important because they show distinctive visual behaviour. To investigate specific scanpath features we need conduct a different analysis using the SubsMatch algorithm.

SubsMatch Algorithm. SubsMatch is an algorithm that builds transition matrices from detected scanpaths, instead of creating strings. This means, that the features detected in a scanpath are all possible combinations of transitions between all OOI. This is equal to building all permutations for a set of objects. Each detected transition adds $+1$ to the respective feature value. By using this method, we loose information about order in which the transitions occurred during the scanpath sequence.

In the algorithm presented by Kübler et al. (2017), the OOI had to be extracted separately from the 2-D images. They identified fixation and saccades to calculate the transitions from one object to another. In our analysis the OOI was already given, because we extracted the gaze object with the ray cast beforehand. So, we used the gazed object information and calculated the transition matrices directly for our OOI. We did this according to their temporal order in the dataset also considering the given time variable. After applying the algorithm, we got a transition matrix for each scanpath with the total number of all transitions that occurred during the chosen time interval.

Since the SubsMatch does allow to identify all combinations of transitions, we also took time intervals spent on the same object into account. If an object is observed for a longer period of time, we counted more than one transition within the same object. Therefore, we basically looped over the same transition and filled up the diagonal entries of the transition matrix. We also implemented an additional parameter restricting the time between a transition from one object to another. We only counted transitions between two OOI if the transition time between them was smaller than a certain threshold. This helped to identify direct transitions between OOIs and to drop cases where subjects watched other objects in between the transitions from one OOI to another.

An advantage of counting individual transitions is, that we could not only take transitions between two objects but longer transition patterns into account. As in Kübler et al. (2014), we call a transition between two OOI a 2-gram feature. Longer transition patterns are called n-grams, with n beeing the length of the

pattern. Therefore, we could increase the number of features by collecting all combinations for transitions between three OOI (3-gram features). In theory, the matrix representation allows to collect up to n-gram features, with an arbitrary length of n. But in practice, by increasing the length of the n-grams the number of features increases almost quadratic. So for example, with two OOI and using only 2-grams, we get four transition features. If we additionally allow 3-grams, we would get four 2-gram features and eight 3-gram features, increasing the total number of features to 12. On the one hand, increasing the length of the n-grams gives access to more information from the scanpath. On the other hand, it also makes the interpretation of the results more difficult due to the high number of different features.

Another advantage in comparison to the string-based methods is, that there is no limitation for the length of one scanpath. Longer time intervals only result in higher values in the transition matrix but do not increase the number of dimensions. Therefore, we applied the SubsMatch algorithm to longer scanpath sequences. Then, we used the resulting transition matrices including all calculated feature values as data for further analysis. Differences in the scanpaths can be identified by using Support Vector Machine (SVM) classification on theses transition matrices.

Kernel Support Vector Machine. SVM is a classification algorithm that is trained supervised to separate two classes in a data set. The data is represented in a multidimensional feature space and the SVM algorithm tries to find the separating hyperplane with maximal margin between the two classes (classes labeled as $+1$ and -1). The algorithm optimizes for the perfect hyperplane such that the data points closest to the hyperplane (the support vectors) still have maximal distance to it. This allows accurate classification, since SVM only considers data points which are critical for correct classification. A practical way to deal with higher dimensional data is to use kernel functions, which compare the data samples pairwise. Kernel SVM algorithms use the so called kernel trick for fast processing high dimensional data and to perform non-linear classification (Scholkopf & Smola, 2001).

But kernel SVM can also be used for linear classification. In this project we used a linear kernel to identify the importance of specific features in the data

set. This was possible, by investigating the feature weights of the trained SVM classifier. In doing so, we identified which features were important for correct classification and which ones were not. This is basically the same as looking at regression coefficients (weights) in a linear regression model. By looking at the size of a feature weight one can determine the importance of that specific feature for classification. By looking at the sign of a feature weight one can determine the influence direction towards one or the other class. With this feature inspection one can find out which features play an important role in separating the two classes (Chang & Lin, 2008).

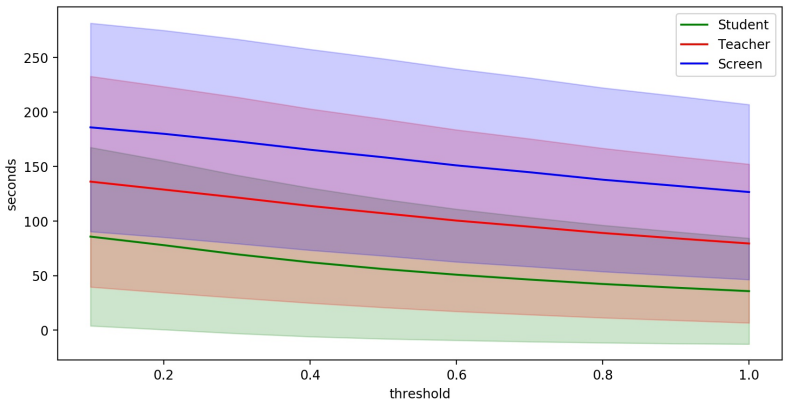
4 Evaluation and Results

4.1 Time spent on Objects of Interest

Attention on OOI. For the first analysis of total attention time spent on specific OOI, we distinguished different categories: peer learners (collision with peer learner objects), teacher (collision with the teacher object) and screen (collision with white board, where the slides were shown). We did not consider other objects in the classroom like tables, chairs, walls etc., since we were not interested in this information. We also did a more fine grained analysis of the time spent on specific peers.

Considering our three OOIs (peer learners, teacher, screen) we ran the attention selection algorithm with different thresholds from 0.1 to 1.0 seconds with intervals of 0.1. We stated the mean attention time on these three objects for all runs and also stated their standard deviation. The results can be seen in Figure 5. The overall time spent on the OOI linearly decreased when we increased the threshold. However, the relative time differences between the OOIs remained. Taking the standard deviation as a measure for the between subject differences, we found out that the most changes can be seen in time on peers from $SD = 81.86$ seconds to $SD = 48.59$ seconds, when increasing the threshold.

For further analysis we picked out an attention threshold of 0.5 seconds to not over or underestimate the time spent on the OOIs. For a threshold of 0.5 seconds, the participants paid most attention (time in seconds for the full ex-



Notes. Mean time and standard deviation spent on three OOI for different thresholds from 0.1 to 1.0 seconds.

Figure 5: Attention time on objects for different thresholds

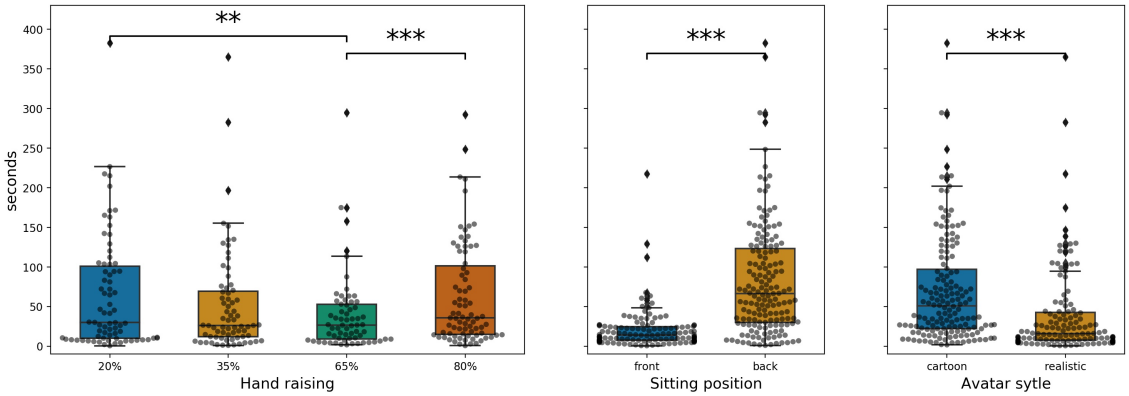


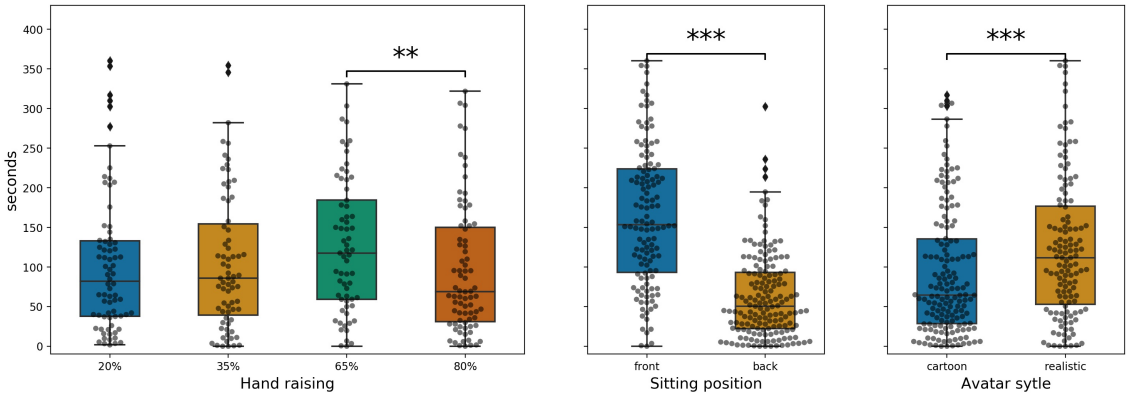
Figure 6: Attention time on peer learners for all conditions

periment) to the screen ($M = 158.57$, $SD = 90.57$), less time on the teacher ($M = 107.16$, $SD = 86.55$), but at least about one minute towards the peers ($M = 56.02$, $SD = 64.14$).

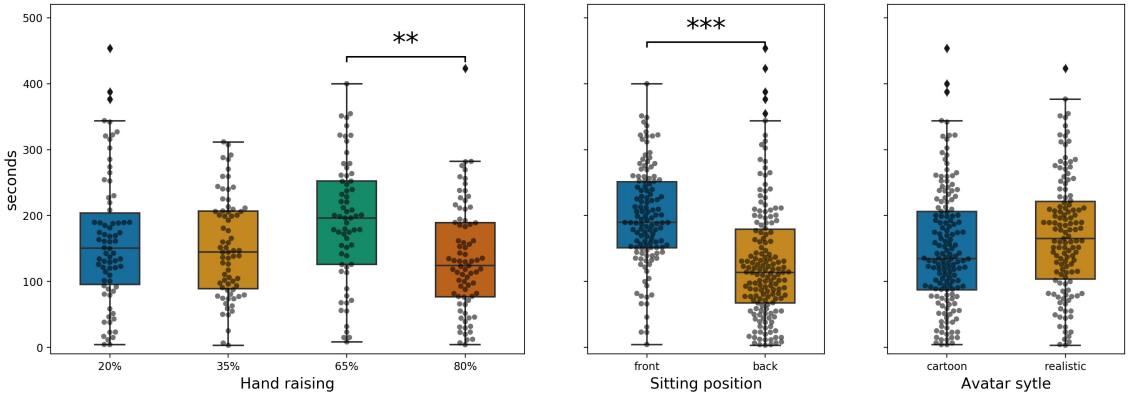
Analysis of Variance. We investigated the time spent on the OOIs for the different experimental conditions. We tested for normal distribution in the data for all condition groups to perform a 3-way-full-factorial ANOVA. Unfortunately, some of the samples were not normally distributed. To test for normal distribution we used the Kolmogorov-Smirnov test and found out that for on average 4 out of the 16 condition the null hypothesis of normal distribution could be rejected (with $p < 0.05$) for all three OOIs. Therefore, we had to perform the ART on the data before applying the full factorial ANOVA.

First we answered the question whether the subjects' time spent on the virtual peer learners differed between the experimental condition. There were significant differences between the total times in the hand raising condition, between the avatar styles and the sitting position (see Figure 6). Subjects spent significantly more time on the peers when sitting in the back ($M = 83.26$, $SD = 71.26$) then in the front ($M = 20.78$, $SD = 26.02$) with $F(273) = 158.38$, $p = .000$. Subjects also spent more time on the peers when they were portrayed in the cartoon style ($M = 70.48$, $SD = 67.51$) then in the realistic style ($M = 38.59$, $SD = 55.20$) with $F(273) = 54.42$, $p = .000$. Variation in the hand raising condition also showed a significant effect on attention time ($F(273) = 7.60$, $p = .0001$).

The post-hoc t-test (Tukey adjustment) for the hand raising condition showed a significant difference between the 20% ($M = 64.49$, $SD = 72.17$) and the 65% ($M = 40.91$, $SD = 50.61$) condition ($t(273) = 3.519$, $p = .00028$). And a signifi-



(a) Attention time on teacher for all conditions



(b) Attention time on screen for all conditions

Figure 7: Boxplots for attention time on teacher and screen

cant negative difference between the 65% and the 80% ($M = 63.56$, $SD = 63.99$) condition ($t(273) = -4.513$, $p = .0001$).

We also investigated the time spent on the teacher. The results for time spent on teacher point in the opposite direction (see Figure 7a). Conditions with less time on peer learners show more time on the teacher and vice versa. Participants spend significant more time on the teacher, when they were seated at the front of the classroom ($M = 163.62$, $SD = 88.03$) as compared to the back ($M = 63.52$, $SD = 54.23$) with $F(273) = 150.78$, $p = .0000$. They spent significantly more time on the teacher with the realistic avatars ($M = 123.53$, $SD = 89.87$) in comparison to the cartoon avatars ($M = 93.59$, $SD = 81.51$), with $F(273) = 13.69$, $p = .00026$. For the hand raising condition we found a significant difference between the condition of 65% hand raising ($M = 126.89$, $SD = 85.28$) and 80% hand raising ($M = 95.81$, $SD = 82.69$) with $t(273) = 3.295$,$p = .0061$.

For the time spent on screen we found almost the same effects as for the time spent on teacher with the same effect size and direction. The only difference was, that there was no significant effect between the realistic and the cartoon

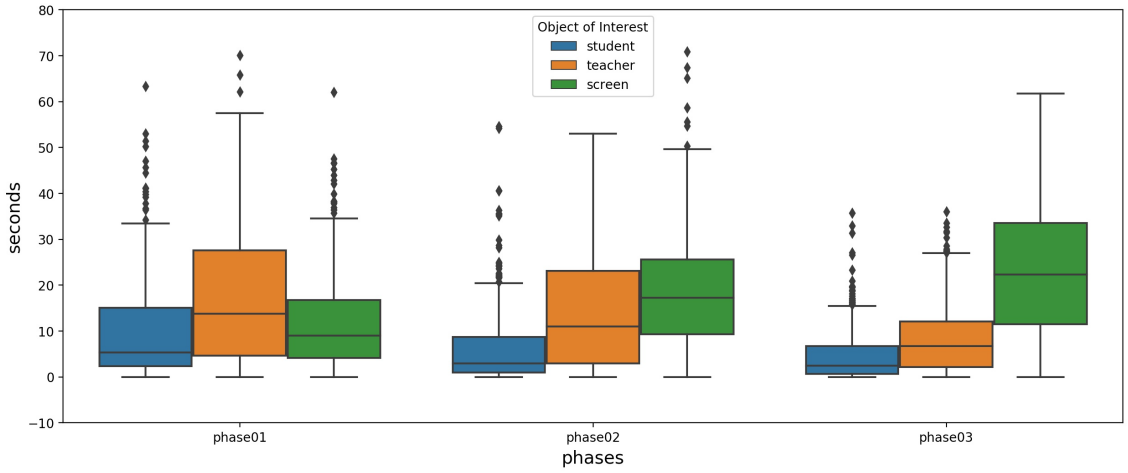
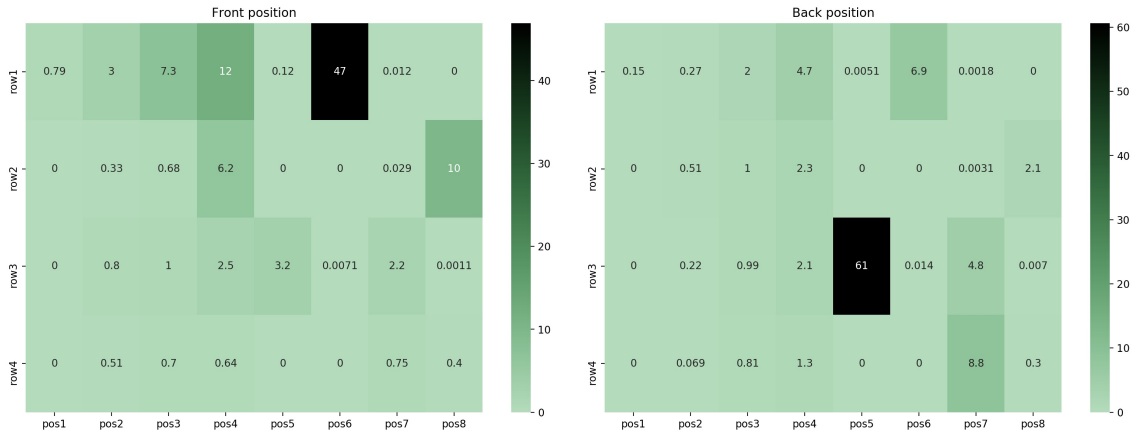


Figure 8: Attention time on OOI for different phases of the lesson

condition for the time spent on screen, as we can see in Figure 7b. All tables for the different analysis can be found in the appendix 5.3.

Within the Lesson. We investigated the time spent on the three OOIs for the first three phases of the lesson, to establish a preliminary idea, at which part of the lesson the participants looked at the different objects of interest. We did not conduct another statistical analysis, because the phases have different length and different periods of social interaction or frontal teaching. An overview for the three lessons with the OOI can be seen in Figure 8. At the beginning, participants spent more time on the teacher ($M = 18.24$, $SD = 17.19$) and less time on the screen ($M = 11.84$, $SD = 10.93$). Over the course of the lecture these values change. At the end, most time is spent on the screen ($M = 22.70$, $SD = 13.50$) and only little time on the teacher ($M = 8.62$, $SD = 8.03$). Also the overall variance of time spent on teacher decreased over time. Time spent on peers was always the lowest value in all three phases, but at the beginning it had the highest value ($M = 10.21$, $SD = 11.49$) and also decreased over time (phase 3, $M = 4.94$, $SD = 6.28$). It is worth mentioning, that the lengths of the phases were not similar. The first phase was the shortest. Phase 2 and 3 had roughly the same length.

Observed Peer Learners. As a second step, we wanted to investigate the frequency, in which the different peer learners had been looked at. Therefore, we collected the time (by frames) spent on the different peers. A visual rep-



(a) Observed peers for condition front **(b)** Observed peers for condition back

Figure 9: Percentage of observed peer learners according to their sitting position

resentation of the $4x8$ sitting positions of the peers in the classroom and their observation frequency can be seen in Figure 9. The average frequency on the peer learners for participants seated in the front (Figure 9a) and seated in the back (Figure 9b) showed different results. Participants sitting in the front were seated in in the second row, fifth position. For this condition, the most frequent observed peer learner was the one in the first row between the subject an the screen. This peer learner (row 1, position 6) was focused on 47% of the time, when subjects spent time on peers. A similar result was observed for the participants seated in the back (row 4, position 5). Here, the most frequently observed peer learner was also the one sitting directly in front of the participant (row 3, position 5). This peer was also in the direct line of sight between the subject and the screen or the teacher at the front and was focused on 61% of the time, when subjects spent time on peers. For the other conditions, we did not found significant patterns in the observation behaviour.

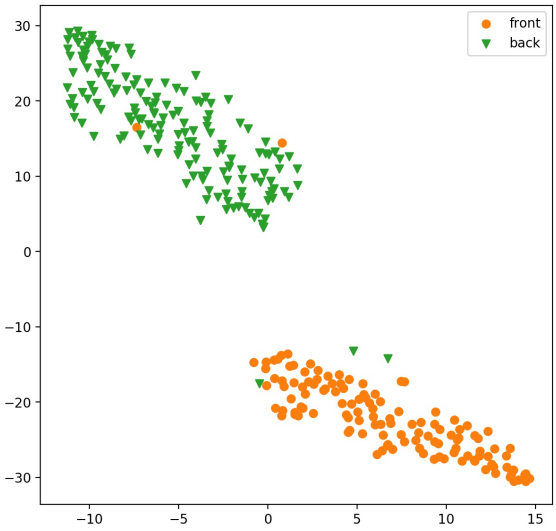


Figure 10: t-SNE with sitting position labels

As an additional step, we calculated the classroom matrix for each participant separately and used the t-SNE to visualize all matrices in a 2-dimensional space. We also plotted the labels for the sitting position, since this condition almost perfectly matched with the cluster, detected by the t-SNE. The t-SNE visualisation can be seen in Figure 10. We also calculated the average number of observed peer learners and found out that for all conditions nearly the same number of peers were observed (e.g. $M_{back} = 16.46$, $M_{front} = 15.36$).

4.2 Scanpath Analysis

String-Based ScanMatch. To perform a string-based scanpath analysis we wrote an algorithm which is similar to the one suggested in ScanMatch (Cristino et al., 2010). Instead of comparing specific regions of interest, we used the OOIs to detect transitions. First, we transformed the consecutive OOI into a sequence of letters. We also included the time spent on the same OOI by repeating the same letter multiple times. Each $50ms$ time interval on the same object is represented as a repeating letter. Furthermore, all transitions between two OOI are represented in the string, independent of the time the transition took (see appendix 9).

Since the Needleman-Wunsch algorithm is not suitable for long sequences with different length, we focused only on the parts of the lesson where we assumed most social interaction. Therefore, we selected five time periods at points of question and answer during the lesson. We only took the first five $Q\&A$

sessions at the beginning of the lecture, because we have seen in Figure 8 that the time spent on our three OOI is almost equal in phase one. There, we expected most transitions, compared to the other phases, where we observed more screen time and probably less transitions (time intervals in seconds: I= $[52-62]$, II= $[73-83]$, III= $[94-104]$, IV= $[108-118]$, V= $[121-131]$).

We considered teacher, screen, peer learners on the left side of the class and peer learners on the right side of the class as our OOI categories. We distinguished the two groups of peer learners, because participants had to show more visual behaviour to look at peers on the left side of the classroom, because of their own sitting position on the right. Therefore, we created the sting patterns with four different letters ($T = teacher$, $S = screen$, $L = left\,peers$, $R = right\,peers$).

For each time interval we calculated the similarity score between all samples by using the a preprogrammed version of the Needleman-Wunsch algorithm[4] and normalized the similarity matrix afterwards. Then, we took the mean values over all five similarity matrices entry-wise. To get an idea of clusters or groups in the sample and to get a first impression about the similarity scores, we printed a heat map with all similarity scores in one matrix (see Figure 11). After that, we applied a k-Nearest Neighbour classification for each sample in the matrix

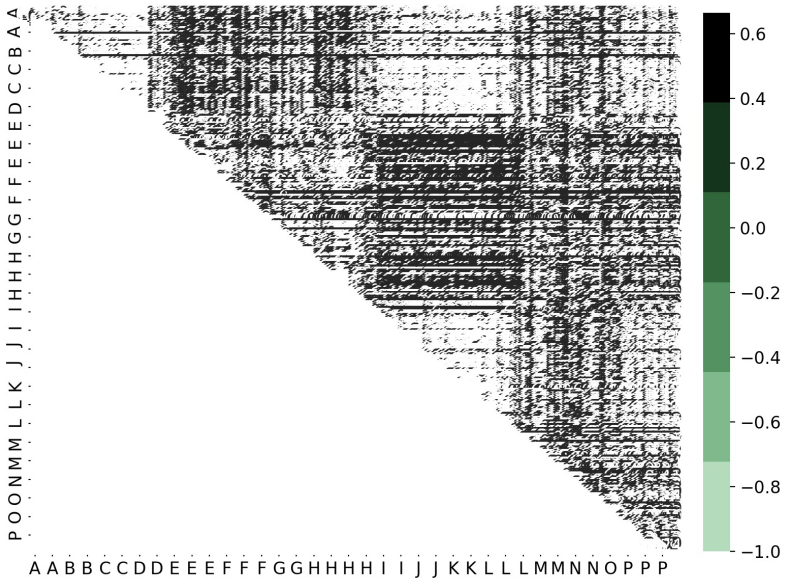
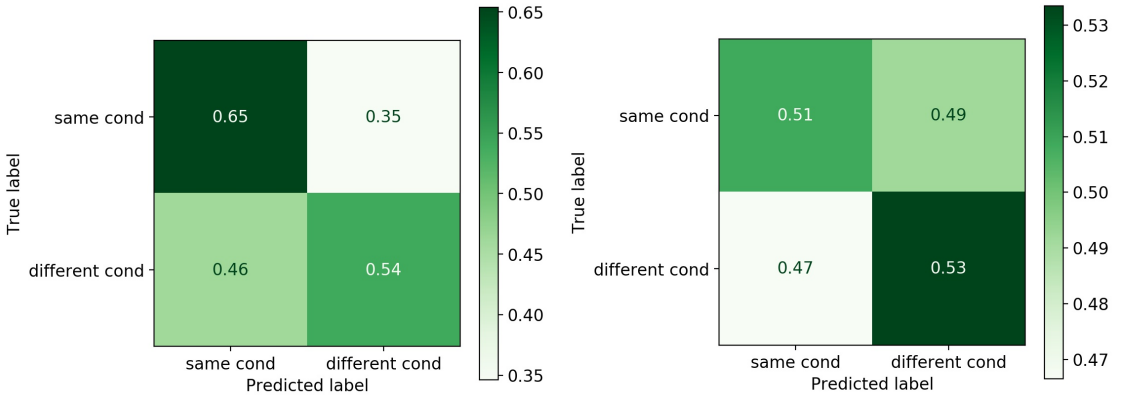


Figure 11: Similarity scores of the Needleman-Wunsch algorithm for all samples according to all 16 conditions (see appendix 8).

[4]minineedle from https://github.com/scastlara/minineedle



(a) Sitting position **(b)** Avatar style

Figure 12: Confusion matrices of the kNN classification for different conditions

($n = 220$ found by rough grid search). Therefore, we labeled the scores as either 1 if both compared sequences belong to the experimental condition or 0 if they belong to different conditions. This results in two classes either being in the same condition or in different conditions. We split the data into 80% train set and 20% test set and performed a 10 fold cross validation. For classifying the sitting position or the avatar style the overall classification accuracy was in both conditions only about average (sitting 58.9%, avatar 51.5%). For the sitting conditions the correct prediction for the same condition class was slightly better (65%) then for different conditions class(54%). For for the avatar style condition both accuracies were only at $\approx$ 52%. (Confusion matrices in Figures 12a and 12b).

For the hand raising condition we had to deal with unequal sample size (1 : 3), which influences the results of a kNN classifier. Because with four hand raising conditions a sample is only 25% of the time in the same condition class. Therefore, we randomly dropped two third of samples from the different conditions class and performed the same kNN classification multiple times. The total classification accuracy was lower than average (49.80%) and also $\approx$ 50% for all accuracies in the confusion matrix (see Figure 16).

Matrix-Based SubsMatch. The SubsMatch algorithm creates transition matrices from given scanpaths. To apply this to our data we wrote an algorithm that worked similar to the one used in the string-based methods (see appendix 10). But instead of building strings, we count the number of transitions in a scanpath matrix. Here we ensured that the time between the transitions from one OOI to

another is smaller than 2 seconds. We counted time intervals larger than $50ms$ on the same object as 'transitions' within the object, to integrate the duration spent on an object as additional information.

Using the matrix representation instead of strings, allows the analysis of longer longer sequences, without longer runtime or loss of precision. Thus, we treated the full experiment of a person as one sequence and calculated transition matrices for full lessons.

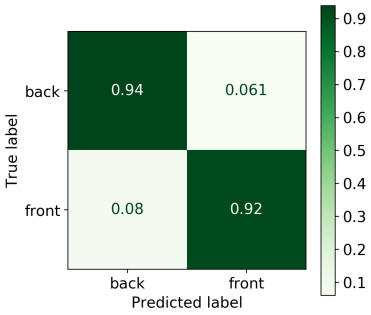
First, we analysed 2-grams, by taking only transitions between two objects into account. As before, we used the teacher and the screen as two OOI. For the virtual peer learner we decided to distinguish specific groups. Since we wanted to investigate if the subjects actually looked at virtual peers, which raised their hand during a question, we divided the peers in different groups. We could not just divide the peers into the ones who raised their hand and the ones who did not, because then the group size would be different for different hand raising conditions. To ensure that we have the same groups for all conditions we divided the peers into these four groups:

- *Peer-group 1:* All peers that raised their hand in the 20% hand raising condition
- *Peer-group 2:* All peers that raised their hand in the 35% condition, except the ones in Peer-group 1
- *Peer-group 3:* All peers that raised their hand in the 65% condition, except the ones in group 1 and 2
- *Peer-group 4:* The rest of the remaining peers which are not in group 1, 2 or 3.

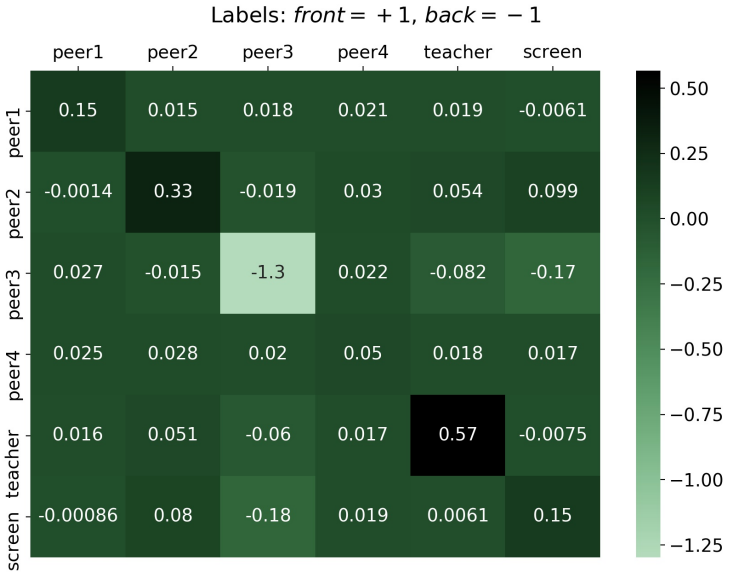
So, we were able to investigate if peers, that raised their hand, influenced the scanpath patterns, without having different OOI group sizes for the different conditions. Note, that virtual peer learners in peer-group 1 show hand raising behaviour in all conditions. Peer-group 2 shows hand raising behaviour in condition 35% to 80%, and so on. As before, we also focussed on the teacher and the screen as additional OOI. With 6. OOI and 2-grams we got a feature space of 36 features (OOIs are labeled as: $teacher = t$, $screen = s$ $peer - group\,1 = p1$, etc.). We randomly split our data set, consisting of the scanpath matrices of

all participants, into train set (80%) and test set (20%). Then, we applied the SVM algorithm with a linear kernel and a regularisation parameter $C = 0.1$. For SVM classification we used the *sklearn* library[5]. We conducted the analysis three times, separately for the different conditions (sitting position, avatar style and hand raising).

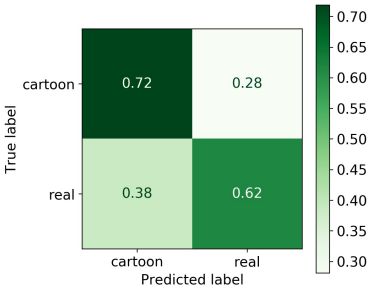
The prediction of the correct sitting position of the participant was accurate for both conditions (see Figure 13a). The front position was classified correctly in 92% of the trials and the back position in 94%. Looking at the different feature



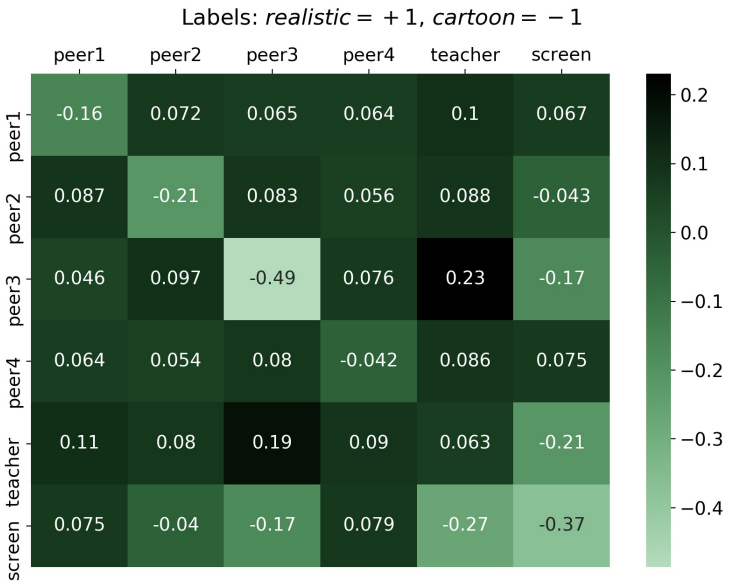
(a) CM sitting position



(b) FM sitting position



(c) CM avatar style



(d) FM avatar style

Figure 13: Confusion matrices (CM) and feature matrices (FM) of SVM classification with 2-gram transitions

[5]https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

weights, we found out that the most important features that support classification for the back condition ($back = -1$) are transitions within peer-group 3 ($w_{p3p3} = -1.3$). Most important features to support the front sitting position are transitions within the teacher ($w_{tt} = 0.57$), but also within peer-group 1 and 2 ($w_{p1p1} = 0.15$, $w_{p2p2} = 0.33$) and within the screen ($w_{ss} = 0.15$, see Figure 13b). For the avatar style conditions, we got an accuracy of 72% for correct classification in the cartoon condition and an accuracy of 62% for the realistic condition (see Figure 13c) . The most important features for classifying the cartoon condition are transitions within peer-group 3 but also the time spent on screen. Important feature weights were found for transitions between peer-group 3 and the teacher ($w_{p3t} = 0.23$, $w_{tp3} = 0.19$), which support classification for the realistic condition ($realistic = +1$). Feature weights that support classification towards the cartoon condition were transitions within the peer-groups 1, 2 and 3 and within the screen ($w_{p1p1} = -0.16$, $w_{p2p2} = -0.21$, $w_{p3p3} = -0.49$, $w_{ss} = -0.37$, see Figure 13d).

We were not able to classify for the correct hand raising condition (see Figure 18). The highest correct classification was reached for the 80% hand raising condition with an accuracy of 38%. But there was a frequent misclassification between the 35% and the 65% hand raising condition. It could also be the case that the sample size is too small ($n = 289$) to perform SVM with four classes.

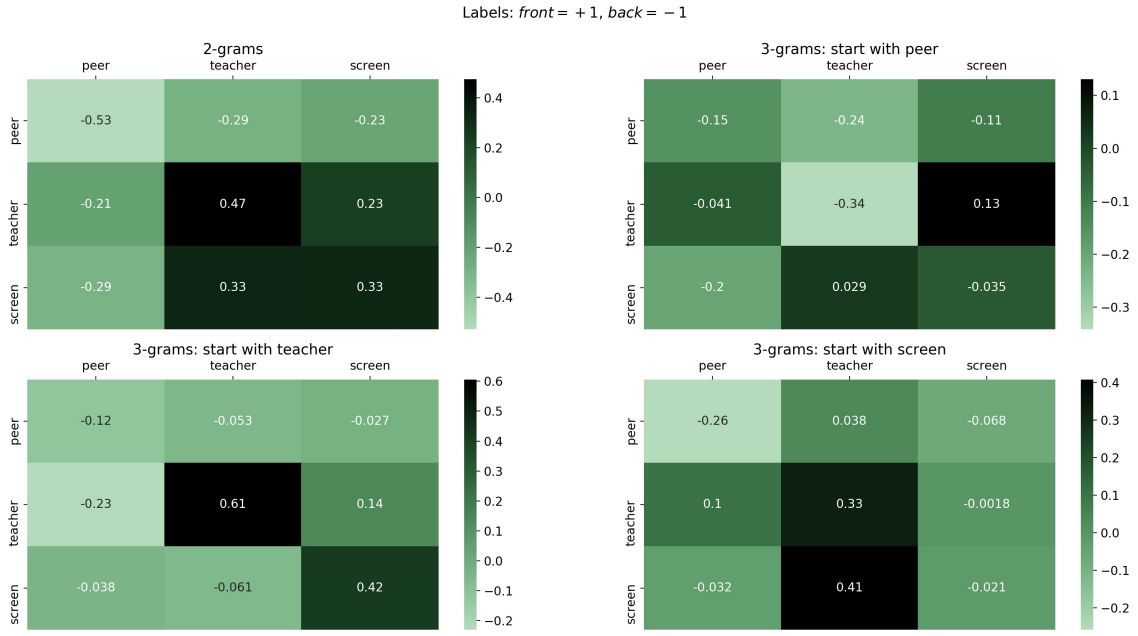


Figure 14: Feature matrices for sitting condition from 3-gram SVM classification

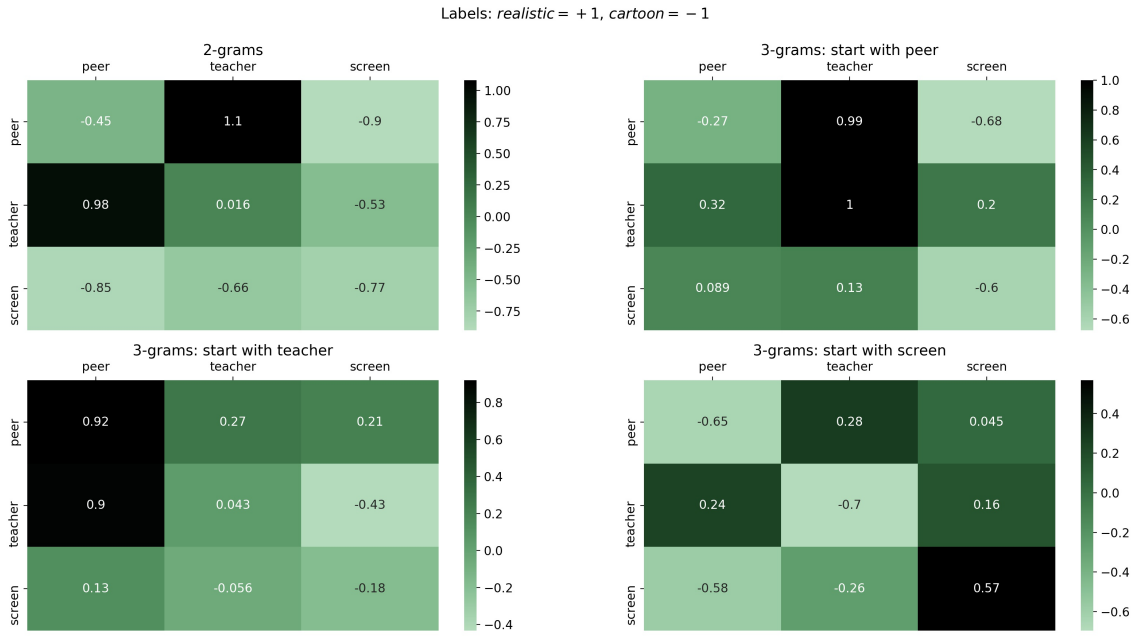


Figure 15: Feature matrices for avatar condition from 3-gram SVM classification

In a second step, we performed the analysis with 3-grams investigating transition combinations up to 2 transitions. Here we only used three OOI (peer learner p , teacher t, screen s), which led also to feature space of 36 features. The SVM classification was done in the same way as in the analysis before. The sitting position could be classified correctly with an accuracy of 84.5% (front 79%, back 90%). The avatar style could be classified correctly in 67% of the times (cartoon 62%, realistic 72%). For the hand raising condition the SVM was not able to predict the classes correctly in most of the times. There was a strong tendency to assign most samples to the same class. We found that the 65% hand raising condition is predicted for most of the samples from the 20% and the 35% hand raising conditions. The class that had most correct classifications was the 80% hand raising condition with an accuracy of 46% (Confusion matrices can be seen in the appendix 19).

Looking at the feature weights , the most important features for classifying the sitting positions were the transitions within the teacher and between teacher and screen ($w_{tt} = 0.47$, $w_{ttt} = 0.61$, $w_{tst} = 0.42$, $w_{sst} = 0.41$, $w_{stt} = 0.33$). All these features influence the decision function towards the front condition ($front = +1$). Important feature influencing the decision function to classify the back condition were transitions within the peer learners and between screen or

teacher and peer learners ($w_{pp} = -0.53$, $w_{ptt} = -0.34$, $w_{ps} = -0.29$, $w_{spp} = -0.26$; see Figure 14).

For the feature importance in the avatar style condition, we found that the cartoon condition is supported by features of transitions between peers and the screen ($w_{ps} = -0.9$,$w_{sp} = -0.85$,$w_{pps} = -0.68$, $w_{pss} = -0.66$). Features supporting the realistic condition were found mostly regarding the transitions between teacher and peers ($w_{pt} = 1.1$, $w_{tp} = 0.98$, $w_{ppt} = 0.99$, $w_{ptt} = 1$, $w_{tpp} = 0.92$, $w_{ttp} = 0.9$).

In some cases, the features weights differed regarding the 2-grams and 3-grams (e.g. $w_{ss} = -0.77$, $w_{sss} = 0.57$,) not allowing a clear interpretation of the importance of all features for one or the other class (see Figure 15).

Since the classification was not accurate for the hand raising condition, we did not investigate the features weights for this classification.

5 Discussion

In this work we focussed on visual behaviour of children in an IVR learning environment. We analysed overt visual attention in the virtual classroom and measured children's attention towards the lecture content but also towards the virtual surrounding in the classroom. We investigated, whether our participants paid attention towards their virtual peer learners to see if they recognize social dynamics in the classroom.

Therefore, we focussed on differences in the visual behaviour for different experimental conditions. The classroom manipulations (subject's sitting positions, avatar style of the virtual characters and different hand raising behaviour of the peer learners) allowed us to analyse group similarities and differences in visual attention between different participants. The acquired insights give a first understanding of the influence of character design choices, of different social dynamics and the importance of the placement of participants in a virtual classroom.

We also presented different established methods to analyse visual attention in IVR with regard to the gazed objects. Information about the gazed objects allowed us the apply different scanpath analysis, which could be used to investigate overt attention patterns. This can be seen as a first step towards a more detailed understanding of visual attention in a virtual classrooms. As we have seen before, most studies which analysed visual attention in an IVR classroom, did not even analyse gaze information. Hence, we wanted to evaluate the potential of using information about the gazed objects to discuss possibilities and limits of analysing this data. Being able to understand visual attention in IVR with regard to a classroom learning environment might be a major contribution for research and practice.

5.1 Visual Attention

Using eye-tracking and HMD information turned out to be an accurate source to perform ray casting in the virtual environment. Even though the colliders of the OOIs in the virtual environment were relativity small, we were able to detect significant gaze time on them.

One critical parameter was the attention threshold. In previous studies, de-

tected fixations (100*ms*) were interpreted as overt attention towards an object (cf. Kübler et al., 2017). In our analysis we were able to define different thresholds for visual attention towards an object. By increasing the threshold, which means that only longer durations on an object are identified as attention, we also increased the chance that relevant information about the object are actually recognized. Since we wanted to ensure that participants are able to encode social information, we decided to set the threshold to half a second. It turned out that with a higher threshold we still observed attention time on the OOIs, which indicates that the participants also spent longer durations on these objects.

Overall participants paid most attention towards the screen, but also towards the teacher. This indicates that, in addition to the information given on the screen, the virtual teacher is a relevant feature in a virtual learning environment. Despite that, participants spent on average almost one minute on the virtual peer learners. These results show that they not only paid attention towards the screen and the teacher, but also a significant amount of time towards social dynamics in the classroom. But we also realised that the most frequent observed peer learners were the ones in the direct line of sight between the participants and the teacher or the screen. One the one hand, this result seems reasonable, since these peer learners are the most salient ones in the visual field with orientation towards the front of the classroom. On the other hand, there is a chance that the gaze only stayed on these peer learners for longer time, but participants did not pay attention towards them. For example, during periods where participants think about the lecture content, the gaze could be directed towards the virtual peers, without paying attention. From the given data we are not able to distinguish these two possibilities.

Investigating the different phases of the lesson, we found out that children spent more time on the teacher at the beginning and more time on the screen at the end of the lesson. A reason could be that at the beginning there is no information shown on the screen and the teacher is the most important character, because she gives instructions and guides the participant. We also found that the variance of the total time on peer learners decreased during lesson. One possible interpretation could be that after some time participants did not need to observe the virtual peers any more, since their behaviour (e.g. the hand raising) was always the same for the whole lesson. It would be interesting to explore, if

children at some point felt that they could anticipate the behaviour of the virtual peers and felt no need the pay further attention towards them.

5.2 Classroom Manipulations

Overall, we found that visual attention was distributed differently between the participants. For each experimental condition there was always a large variances within the same experiment group regarding the time spent on specific OOIs. This indicates that visual attention is driven by individual factors or factors we did not take into consideration for the experiment, too. Despite that, we found some indicators for shared visual attention behaviour. The average number of observed peer learners during the lesson was similar in all experimental conditions. But we did not check if most of them are only observed at the beginning of the lesson, when participants are asked to look around in the classroom.

Furthermore, we found that visual behaviour was also influenced by the different classroom manipulations introduced for different experimental conditions.

5.2.1 Sitting Positions

The most different visual attention behaviour was found for the different sitting positions of the participants. The differences in time spent on peer learners, teacher and screen were highly significant comparing both sitting conditions. When seated in the front, participants showed significantly more screen time and time on the teacher. When seated in the back, they spend more time on their virtual peer learners. Additionally, the t-SNE visualisation showed two clear clusters according to the sitting position when using frequency information about the observed peer learners.

Another interesting result was, that the variance for time spent on peer learners was larger for the back condition then for the front condition. This indicates that subjects show different attention behaviour with regard to the peer learners when seated in the back. In the front condition the variance for time on peers was very small, indicating that participants seated in the front behave more similar with regard to attention towards their peers. In contrast, however, the time on

the teacher showed more variance for participants in the front. Since the visual distance between teacher and screen is larger in the front condition, participants had to show more visual behaviour to pay attention to the screen but also to the teacher. These results suggest that we not only encounter a trade-off between concentrating on peer learners or on the content of the lesson dependent on the sitting position. On top of that, the visual distance between objects may be an important factor when distinguishing different attention behaviour in an IVR.

The SVM feature analysis, supported the results from the ANOVA. We found that the model identified peer-group 3 as a support for the back condition. This group contained the peer leaner that was sitting in front of the participant, when he or she was seated in the back. Peer-group 2 supported classification for the front condition, which contained the peer learner observed the most by participants seated in the front. Additionally, more time on the teacher was identified as a strong indicator for the front condition. Transition patterns within the teacher and between teacher and screen were also strong indicators for the front condition, while interaction between peers and teacher showed behaviour that was prominent for participants in the back. These results indicate, that positioning participants in the back of the classroom encourages them to pay attention towards the peer learners, while a sitting position in the front forces them to look at the screen and the teacher and mostly ignore their classroom surrounding.

5.2.2 Avatar Styles

The different avatar styles also influenced the time spent on peers and teacher. Participants paid more attention towards the virtual peers, when they appeared as cartoon characters. This result indicates that the virtual presentation of the peer learners influences their attention behaviour.

In the cartoon condition the heads of the peers are bigger and therefore more recognizable. Longer transitions within the virtual peer groups were identified to support the cartoon condition, when looking at the feature weights of the SVM classification. This indicates that participants spent longer durations on the virtual peers, when they are presented as cartoon characters. More transitions between the teacher and the peers supported the realistic condition. This indicates that participants showed more interactions between the teacher and the

peers, when they were presented in a realistic avatar style. The feature weights also indicated that more transitions between peer learners and screen support classifying the cartoon condition. This could be the case, because the bigger size of the virtual peers in cartoon style could interfere with participants need to recognize information on the screen. Since there is always a peer learner in front of the participant, free vision towards the board is blocked.

In contrast, however, participants spent less time on the teacher in the cartoon condition. This finding also aligns with the general trade-off we observed between looking at the teacher and looking at the peers. No significant difference was found for the time spent on screen. We can see this as a control for the avatar style condition, since nothing on the screen was changed for this manipulation. Participants attention on the lecture content was not influenced by the appearance of the virtual characters.

5.2.3 Hand Raising

The hand raising condition can be seen as the most complex classroom manipulation, since it requires the participants to recognize social dynamics. In contrast to the other classroom manipulations which were very salient stimuli in the classroom, the different hand raising behaviour of the virtual peer learners was a rather small manipulation that occurred only a few times during the virtual lesson. We can argue that finding differences for this manipulation, would be a strong indicator that participants process social information in a virtual classroom.

Looking at the overall time spent on the OOI in the hand raising manipulation, we observed an unexpected result. Participants paid most attention to the virtual peer learners for the 20% and the 80% condition and less to the medium hand raising of 35% and 65%. This indicates that the hand raising attracted participants attention, when the behaviour of the peer learners tend to the extreme cases. The lowest attention towards the peer learners was paid for the 65% hand raising condition. One possible interpretation of this result could be that this condition is experienced as the most natural one and participants felt no need to encode social information. In this condition, they experienced the classroom as a common learning environment and generally focussed more on the lecture

content. The significant differences in time spent on peer learners between the extreme hand raising conditions and the 65% hand raising condition supports that interpretation. The time spent on teacher and screen for the different conditions also pointing in the same direction. The results show a reverse pattern for the time on these two OOI in contrast to the time spent on the peer learners. This means that when people paid more attention towards their peer learners, the paid less attention towards the content of the lesson.

Unfortunately, we could not analyse the transition patterns from the SVM classification, since the algorithm was not able to accurately classify the hand raising conditions from the data. In the way we conducted the scanpath analysis it could be possible that potential changes during the hand raising periods did not influence the overall attention pattern of the participants for the whole lesson. Moreover, a sample size of $n = 289$ could to small to perform a SVM classification with four classes. It would be interesting to see if the classification accuracy increases, when we use only two larger classes for the hand raising conditions. Since we investigated that the time spent on the OOI is more similar between 20% and 80% and between 35% and 65%, we would suggest to concatenate the four classes into these two.

5.3 Possibilities, Limits and Future Research

In general, we can argue that using an IVR classroom to investigate different visual attention behaviour can be a promising alternative to research conducted in a real classroom. The introduced manipulations showed a variety of different effects. The sitting position of the participants in the virtual environment had a great influence on their visual attention. In an experimental setting, where one wants the participants to recognize social dynamics in a virtual classroom, we suggest using a placement further in the back. Also the virtual appearance of the peer learners influenced participants' visual behaviour and may help to understand effects of presence or immersion in an IVR classrooms to create virtual learning environments that are experienced as more realistic.

Participants visual behaviour also indicates that they process social information in the virtual classroom and that they anticipate habituated classroom behaviour. We were able to find differences for different behaviour of the virtual

peer learners, even though participants only spent a short period of time in the virtual classroom. These results give evidence for the potential of using IVR to investigate social classroom effects.

In addition to the analysis for total time spent on OOIs, classification with the SubsMatch algorithm produced some interesting result. Some additional information could be provided by looking at the feature weights to detect important transition patterns. The scanpath results from the SubsMatch algorithm pointed in the same direction as the results from the ANOVA and provided some useful insights about group specific visual attention behaviour.

Unfortunately, the kNN algorithm used to classify string similarities did not produce accurate results for the ScanMatch algorithm. But if we look at the heatmap of the similarity matrix we are able to identify some cluster or patterns. Further investigations are necessary to analyse these structures in the similarity matrix. On the one hand, the Needleman-Wunsch algorithm allows some additional fine tuning within its scoring system, for example with regard to additional information about the similarities of certain OOIs (Cristino et al., 2010). On the other hand, we can use a more sophisticated clustering algorithm to see if we are able to detect some unique properties of potential clusters.

Apart from improvements, that can be made to analyse the gazed object data, future virtual classroom experiment should consider a clear spatial separation between the peer learners, the virtual teacher and the screen. If we want to make sure, that the time spent on a specific OOIs is motivated by the need for information about them, they should not overlap in the visual field of the participant. If the OOIs have a greater visual distance in the visual field of the participants, attention towards them requires more visual behaviour and allows us to identify stronger visual behavioural cues. Thereby, we can also ensure that gaze inaccuracies do not lead to misclassification of the gazed object.

Additionally, with our method of ray casting we only tracked one object for each frame. We did not collect information about objects close to the gazed object. If objects in the virtual environment are close to each other it could also be possible that attention is distributed and not directed to the most salient object or the object the hit by the gaze vector. Investigations in the field of covert attention have shown that attention is not always equivalent to the spatial gaze location (Carrasco, 2011).

Another aspect, with regard to visual attention in general is, that we only used a fix attention threshold. For further investigations it would be interesting to threat this threshold as a free parameter. Information about the overall time spent on the different OOIs for specific attention thresholds could be used as prior information in a probabilistic approach of analysing group differences. Since we do not know which threshold is the correct one for analysing visual attention, it could be promising to consider a Bayesian approach (e.g. Bayesian ANOVA, Cleophas & Zwinderman, 2018).

Despite the mentioned limitations, we were able to find significant patterns of visual attention behaviour in IVR and were able to show the effect of classroom manipulations. However, with regard to children's visual attention, there is more research needed to be done to fully understand the effects which influence attention behaviour in IVR classrooms.

References

Adams, R., Finn, P., Moes, E., Flannery, K., & Rizzo, A. S. (2009). Distractibility in attention deficit hyperactivity disorder (adhd): The virtual reality classroom. *Child Neuropsychology*, 15(2), 120–135. https://doi.org/10.1080/09297040802169077

Agtzidis, I., Startsev, M., & Dorr, M. (2019). 360-degree video gaze behaviour: A ground-truth data set and a classification algorithm for eye movements. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, p. 1007–1015, New York. Association for Computing Machinery. https://doi.org/10.1145/3343031.3350947

Awh, E., Vogel, E. K., & Oh, S. (2006). Interactions between attention and working memory. *Neuroscience*, 139(1), 201–208. https://doi.org/10.1016/j.neuroscience.2005.08.023

Bailenson, J. N., Yee, N., Blascovich, J., Beall, A. C., Lundblad, N., & Jin, M. (2008). The use of immersive virtual reality in the learning sciences: Digital transformations of teachers, students, and social context. *Journal of the Learning Sciences*, 17(1), 102–141. https://doi.org/10.1080/10508400701793141

Bailey, J. O. & Bailenson, J. N. (2017). Considering virtual reality in children's lives. *Journal of Children and Media*, 11(1), 107–113. https://doi.org/10.1080/17482798.2016.1268779

Beck, D. M. & Kastner, S. (2009). Top-down and bottom-up mechanisms in biasing competition in the human brain. *Vision Research*, 49(10), 1154 – 1165. https://doi.org/10.1016/j.visres.2008.07.012

Billingsley, G., Smith, S., Smith, S., & Meritt, J. (2019). A systematic literature review of using immersive virtual reality technology in teacher education. *Journal of Interactive Learning Research*, 30(1), 65–90. https://www.learntechlib.org/p/176261

Bioulac, S., Lallemand, S., Rizzo, A., Philip, P., Fabrigoule, C., & Bouvard, M. (2012). Impact of time on task on adhd patient's performances in a

virtual classroom. *European Journal of Paediatric Neurology*, 16 5, 514–21. https://doi.org/10.1016/j.ejpn.2012.01.006

Blascovich, J., Beall, A., Swinth, K., Hoyt, C., & Bailenson, J. (2002). Immersive virtual environment technology as a methodological tool for social psychology. *Psychol. Inq*, 13. https://doi.org/10.1207/S15327965PLI1302 01

Blume, F., Göllner, R., Moeller, K., Dresler, T., Ehlis, A., & Gawrilow, C. (2019). Do students learn better when seated close to the teacher? a virtual classroom study considering individual levels of inattention and hyperactivity-impulsivity. *Learning and Instruction*, 61, 138–147. https://doi.org/10.1016/j.learninstruc.2018.10.004

Bozkir, E., Geisler, D., & Kasneci, E. (2019). Person independent, privacy preserving, and real time assessment of cognitive load using eye tracking in a virtual reality setup. pp. 1834–1837. https://doi.org/10.1109/VR.2019.8797758

Brünken, R. & Seufert, T. (2006). Aufmerksamkeit, Lernen, Lernstrategien. In H. Mandl & H. F. Friedrich (Hrsg.), *Handbuch Lernstrategien*, pp. 27–37. Hogrefe Verlag.

Broadbent, D. (1958). *Perception and Communication*. Pergamon Press. https://doi.org/10.1037/10037-000

Carrasco, M. (2011). Visual attention: The past 25 years. *Vision Research*, 51(13), 1484 – 1525. https://doi.org/10.1016/j.visres.2011.04.012

Cavanagh, P. & Alvarez, G. A. (2005). Tracking multiple targets with multifocal attention. *Trends in Cognitive Sciences*, 9(7), 349 – 354. https://doi.org/10.1016/j.tics.2005.05.009

Chang, Y.-W. & Lin, C.-J. (2008). Feature ranking using linear svm. In Guyon, I., Aliferis, C., Cooper, G., Elisseeff, A., Pellet, J.-P., Spirtes, P., & Statnikov, A. (Hrsg.), *JMLR: Workshop and Conference Proceedings*, volume 3 of *Proceedings of Machine Learning Research*, pp. 53–64, Hong Kong. PMLR.

Cleophas, T. J. & Zwinderman, A. H. (2018). Bayesian analysis of variance (anova). In *Modern Bayesian Statistics in Clinical Research*, pp. 83–89. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-92747-3 8

Connor, C. E., Egeth, H. E., & Yantis, S. (2004). Visual attention: Bottom-up versus top-down. *Current Biology*, 14(19), 850–852. https://doi.org/10.1016/j.cub.2004.09.041

Cristino, F., Mathôt, S., Theeuwes, J., & Gilchrist, I. D. (2010). Scanmatch: A novel method for comparing fixation sequences. *Behavior Research Methods*, 42(3), 692–700. https://doi.org/10.3758/BRM.42.3.692

Cutrell, E., Guan, Z., & Cutrell, E. (2007). What are you looking for? An eye-tracking study of information usage in web search. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 407–416. Association for Computing Machinery, Inc. https://doi.org/10.1145/1240624.1240690

Díaz-Orueta, U., Garcia-López, C., Crespo-Eguílaz, N., Sánchez-Carpintero, R., Climent, G., & Narbona, J. (2014). Aula virtual reality test as an attention measure: Convergent validity with conners' continuous performance test. *Child Neuropsychology*, 20(3), 328–342. https://doi.org/10.1080/09297049.2013.792332

Driver, J. (2001). A selective review of selective attention research from the past century. *British Journal of Psychology*, 92(1), 53–78. https://doi.org/10.1348/000712601162103

Emmer, E. T. & Stough, L. M. (2001). Classroom management: A critical part of educational psychology, with implications for teacher education. *Educational Psychologist*, 36(2), 103–112. https://doi.org/10.1207/S15326985EP3602 5

Epic Games (2020). Blueprints visual scripting. *https://docs.unrealengine.com/en-US/Engine/Blueprints/index.html*.

Goldberg, P., Sümer, Ö., Stürmer, K., Wagner, W., Göllner, R., Gerjets, P., Kasneci, E., & Trautwein, U. (2019). Attentive or not? toward a machine learning approach to assessing students' visible engagement in classroom instruction. *Educational Psychology Review*. https://doi.org/10.1007/s10648-019-09514-z

Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). Knn model-based approach in classification. In Meersman, R., Tari, Z., & Schmidt, D. C. (Hrsg.), *On The*

Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE, pp. 986–996, Berlin, Heidelberg. Springer Berlin Heidelberg.

Guo, Z., Chen, R., Zhang, K., Pan, Y., & Wu, J. (2016). The impairing effect of mental fatigue on visual sustained attention under monotonous multi-object visual attention task in long durations: An event-related potential based study. *PLoS One*, 11(9). https://doi.org/10.1371/journal.pone.0163360

Hamre, B. & Pianta, R. C. (2010). Classroom environments and developmental processes: Conceptualization and measurement. *Handbook of Research on Schools, Schooling and Human Development*, pp. 25–41.

Hasenbein, L., Trautwein, U., Hahn, J.-U., Soller, S., & Göllner, R. (2020). Does a 15-Minute Exposure to Strong Classmates Affect Students' Self-Concept? An Experimental Test of the Big-Fish-Little-Pond-Effect Using an Immersive Virtual Reality Classroom. *in revision*. unpublished.

Hutmacher, F. (2019). Why is there so much more research on vision than on any other sensory modality? *Frontiers in Psychology*, 10, 2246. https://doi.org/10.3389/fpsyg.2019.02246

Jacobs, R. J. (1979). Visual resolution and contour interaction in the fovea and periphery. *Vision Research*, 19(11), 1187 – 1195. https://doi.org/10.1016/0042-6989(79)90183-4

Johnston, E., Olivas, G., Steele, P., Smith, C., & Bailey, L. (2018). Exploring pedagogical foundations of existing virtual reality educational applications: A content analysis study. *Journal of Educational Technology Systems*, 46(4), 414–439. https://doi.org/10.1177/0047239517745560

Kamińska, D., Sapiński, T., Wiak, S., Tikk, T., Haamer, R., Avots, E., Helmi, A., Ozcinar, C., & Anbarjafari, G. (2019). Virtual reality and its applications in education: Survey. *Information (Switzerland)*, 10, 318. https://doi.org/10.3390/info10100318

Kübler, T. C., Rothe, C., Schiefer, U., Rosenstiel, W., & Kasneci, E. (2017). Subsmatch 2.0: Scanpath comparison and classification based on subsequence frequencies. *Behavior Research Methods*, 49(3), 1048–1064. https://doi.org/10.3758/s13428-016-0765-6

Kowler, E. (2011). Eye movements: The past 25years. *Vision Research*, 51(13), 1457–1483. https://doi.org/10.1016/j.visres.2010.12.014

Kübler, T. C., Kasneci, E., & Rosenstiel, W. (2014). Subsmatch: Scanpath similarity in dynamic scenes based on subsequence frequencies. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, ETRA '14, p. 319–322, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/2578153.2578206

Lai, M.-L., Tsai, M.-J., Yang, F.-Y., Hsu, C.-Y., Liu, T.-C., Lee, S. W.-Y., Lee, M.-H., Chiou, G.-L., Liang, J.-C., & Tsai, C.-C. (2013). A review of using eye-tracking technology in exploring learning from 2000 to 2012. *Educational Research Review*, 10, 90 – 115. https://doi.org/10.1016/j.edurev.2013.10.001

Lodge, J. & Harrison, W. (2019). The role of attention in learning in the digital age. *The Yale Journal of Biology and Medicine*, 92, 21–28.

Lundqvist, D. & Ohman, A. (2005). Emotion regulates attention: The relation between facial configurations, facial emotion, and visual attention. *Visual Cognition*, 12(1), 51–84. https://doi.org/10.1080/13506280444000085

MacAulay, D. J. (1990). Classroom environment: a literature review. *Educational Psychology*, 10(3), 239–253. https://doi.org/10.1080/0144341900100305

Mangalmurti, A., Kistler, W. D., Quarrie, B., Sharp, W., Persky, S., & Shaw, P. (2020). Using virtual reality to define the mechanisms linking symptoms with cognitive deficits in attention deficit hyperactivity disorder. *Scientific Reports*, 10(529). https://doi.org/10.1038/s41598-019-56936-4

Marsh, H. W. (2005). Big-fish-little-pond effect on academic self-concept. *Zeitschrift für Pädagogische Psychologie*, 19(3), 119–129. https://doi.org/10.1024/1010-0652.19.3.119

McCallum, W. C. (2015). Attention. *Encyclopædia Britannica*. https://www.britannica.com/science/attention

McIntyre, N. A. & Foulsham, T. (2018). Scanpath analysis of expertise and culture in teacher gaze in real-world classrooms. *Instructional Science*, 46(3), 435–455. https://doi.org/10.1007/s11251-017-9445-x

Nakayama, K. & Martini, P. (2011). Situating visual search. *Vision Research*, 51(13), 1526 – 1537. https://doi.org/10.1016/j.visres.2010.09.003

Nolin, P., Stipanicic, A., Henry, M., Lachapelle, Y., Lussier-Desrochers, D., Rizzo, A., & Allain, P. (2016). Clinicavr: Classroom-cpt: A virtual reality tool for assessing attention and inhibition in children and adolescents. *Computers in Human Behavior*, 59, 327–333. https://doi.org/10.1016/j.chb.2016.02.023

Noton, D. & Stark, L. (1971). Scanpaths in eye movements during pattern perception. *Science*, 171(3968), 308–311. https:// https://doi.org/10.1126/science.171.3968.308

Pietroszek, K. (2018). Raycasting in virtual reality. In Lee, N. (Hrsg.), *Encyclopedia of Computer Graphics and Games*, pp. 1–3. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-08234-9 180-1

Piontkowski D., C. R. (1979). Attention in the classroom. In Hale G.A., L. M. (Hrsg.), *Attention and Cognitive Development*. Springer, Boston, MA. https://doi.org/10.1007/978-1-4613-2985-5 11

Rizzo, A., Bowerly, T., Buckwalter, J., Klimchuk, D., Mitura, R., & Parsons, T. (2006). A virtual reality scenario for all seasons: The virtual classroom. *CNS spectrums*, 11, 35–44. https://doi.org/10.1017/S1092852900024196

Rizzo, A., Buckwalter, J., Bowerly, T., van der Zaag, C., Humphrey, L., Neumann, U., Chua, C., Kyriakakis, C., Rooyen, A., & Sisemore, D. (2001). The virtual classroom: A virtual reality environment for the assessment and rehabilitation of attention deficits. *CyberPsychology & Behavior*, 3. https://doi.org/10.1089/10949310050078940

Roth, S. D. (1982). Ray casting for modeling solids. *Computer Graphics and Image Processing*, 18(2), 109 – 144.

Rouinfar, A., Agra, E., Larson, A. M., Rebello, N. S., & Loschky, L. C. (2014). Linking attentional processes and conceptual problem solving: visual cues facilitate the automaticity of extracting relevant information from diagrams. *Frontiers in Psychology*, 5, 1094. https://doi.org/10.3389/fpsyg.2014.01094

Salvucci, D. D. & Goldberg, J. H. (2000). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications*, ETRA '00, p. 71–78, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/355017.355028

Scholkopf, B. & Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA.

Seidel, T. & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research*, 77(4), 454–499. https://doi.org/10.3102/0034654307310317

Seo, S.-H., Kim, E., Mundy, P., Heo, J., & Kim, K. K. (2019). Joint attention virtual classroom: A preliminary study. *Psychiatry Investigation*, 16(4), 292–299. https://doi.org/10.30773/pi.2019.02.08

Sezer, A., İnel, Y., Seçkin, A., & Uluçınar, U. (2017). The relationship between attention levels and class participation of first-year students in classroom teaching departments. *International Journ al of Instruction*, 10, 55–68.

Singh, H. & Singh, J. (2012). Human eye tracking and related issues: A review. *International Journal of Scientific and Research Publications*, 2, 1–9.

Slater, M. & Sanchez-Vives, M. V. (2016). Enhancing our lives with immersive virtual reality. *Frontiers in Robotics and AI*, 3, 74. https://doi.org/10.3389/frobt.2016.00074

Soret, R., Charras, P., Khazar, I., Hurter, C., & Peysakhovich, V. (2020). Eye-tracking and virtual reality in 360-degrees: Exploring two ways to assess attentional orienting in rear space. In *ACM Symposium on Eye Tracking Research and Applications*, ETRA '20 Adjunct, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/3379157.3391418

Sthle, L. & Wold, S. (1989). Analysis of variance (anova). *Chemometrics and Intelligent Laboratory Systems*, 6(4), 259 – 272. https://doi.org/10.1016/0169-7439(89)80095-4

Tobii Tech (2020). What is eye tracking? *https://tech.tobii.com/technology/what-is-eye-tracking/*.

Tomasello, M. (1995). Joint attention as social cognition. In Moore, C. & Dunham, P. J. (Hrsg.), *Joint Attention: Its Origins and Role in Development*, pp. 103–130. Lawrence Erlbaum Associates, Inc.

Treisman, A. M. & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12(1), 97 – 136. https://doi.org/10.1016/0010-0285(80)90005-5

Van der Maaten, L. & Hinton, G. E. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9, 2579–2605.

Weintrop, D., Beheshti, E., Horn, M., Orton, K., Jona, K., Trouille, L., & Wilensky, U. (2016). Defining computational thinking for mathematics and science classrooms. *Journal of Science Education and Technology*, 25(1), 127–147. https://doi.org/10.1007/s10956-015-9581-5

Wilson, K. & Korn, J. H. (2007). Attention during lectures: Beyond ten minutes. *Teaching of Psychology*, 34(2), 85–89. https://doi.org/10.1080/00986280701291291

Witmer, B. G. & Singer, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence: Teleoper. Virtual Environ.*, 7(3), 225–240. https://doi.org/10.1162/105474698565686

Wobbrock, J. O., Findlater, L., Gergle, D., & Higgins, J. J. (2011). The aligned rank transform for nonparametric factorialanalyses using only anova procedures. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/1978942.1978963

Appendix

ANOVA results

Condition	Df	Df.res	F value	Pr(>F)	
hand	3	273	7.60	0.0001	***
sitting	1	273	158.38	0.0000	***
avatar	1	273	54.42	0.0000	***
hand:sitting	3	273	4.60	0.0037	**
hand:avatar	3	273	3.63	0.0136	*
sitting:avatar	1	273	13.85	0.0002	***
hand:sitting:avatar	3	273	1.85	0.1388	

Signif. codes: *** 0.001 ** 0.01 * 0.05

Table 2: ART-ANOVA results for time spent on peer learners

contrast	estimate	SE	df	t.ratio	p.value	
H20 - H35	19.2	13.8	273	1.389	0.5074	
H20 - H65	49.3	14.0	273	3.519	0.0028	**
H20 - H80	−13.7	13.8	273	−0.996	0.7519	
H35 - H65	30.2	14.0	273	2.157	0.1381	
H35 - H80	−32.9	13.7	273	−2.392	0.0810	
H65 - H80	−63.1	14.0	273	−4.513	0.0001	***

Signif. codes: *** 0.001 ** 0.01 * 0.05

Table 3: Post-hoc t-test with hand raising for time spent on peer learners

Condition	Df	Df.res	F value	Pr(>F)	
hand	3	273	3.8183	0.01048913	*
sitting	1	273	150.7832	$< 2.22e-16$	***
cartoon	1	273	13.6909	0.00026059	***
hand:sitting	3	273	1.2895	0.27830720	
hand:cartoon	3	273	0.5537	0.64606921	
sitting:cartoon	1	273	3.1260	0.07816888	
hand:sitting:cartoon	3	273	3.1410	0.02578130	*

Signif. codes: *** 0.001 ** 0.01 * 0.05

Table 4: ART-ANOVA results for time spent on teacher

contrast	estimate	SE	df	t.ratio	p.value	
H20 - H35	−0.794	14.3	273	−0.056	0.9999	
H20 - H65	−32.969	14.5	273	−2.269	0.1080	
H20 - H80	14.736	14.3	273	1.033	0.7300	
H35 - H65	−32.175	14.5	273	−2.218	0.1209	
H35 - H80	15.530	14.2	273	1.091	0.6953	
H65 - H80	47.705	14.5	273	3.295	0.0061	**

Signif. codes: *** 0.001 ** 0.01 * 0.05

Table 5: Post-hoc t-test with hand raising for time spent on teacher

Condition	Df	Df.res	F value	Pr(>F)	
hand	3	273	4.62123	0.00359	**
sitting	1	273	53.90726	$2.4235e-12$	***
cartoon	1	273	2.69008	0.10213	
hand:sitting	3	273	0.10669	0.95614	
hand:cartoon	3	273	0.58133	0.62772	
sitting:cartoon	1	273	2.46982	0.11721	
hand:sitting:cartoon	3	273	0.44074	0.72405	

Signif. codes: *** 0.001 ** 0.01 * 0.05

Table 6: ART-ANOVA results for time spent on screen

contrast	estimate	SE	df	t.ratio	p.value	
H20 - H35	4.26	14.2	273	0.300	0.9906	
H20 - H65	−29.77	14.4	273	−2.065	0.1674	
H20 - H80	23.16	14.1	273	1.637	0.3596	
H35 - H65	−34.03	14.4	273	−2.364	0.0866	
H35 - H80	18.90	14.1	273	1.338	0.5394	
H65 - H80	52.94	14.4	273	3.685	0.0016	**

Signif. codes: *** 0.001 ** 0.01 * 0.05

Table 7: Post-hoc t-test with hand raising for time spent on screen

Labels for different experimental conditions

Condition	Hand raising proportion	Sitting position	Graphical representation
A	20%	front	cartoon
B	35%	front	cartoon
C	65%	front	cartoon
D	80%	front	cartoon
E	20%	back	cartoon
F	35%	back	cartoon
G	65%	back	cartoon
H	80%	back	cartoon
I	20%	front	realistic
J	35%	front	realistic
K	65%	front	realistic
L	80%	front	realistic
M	20%	back	realistic
N	35%	back	realistic
O	65%	back	realistic
P	80%	back	realistic

Table 8: Labels for all 16 experimental conditions

Scanpath analysis - additional figures

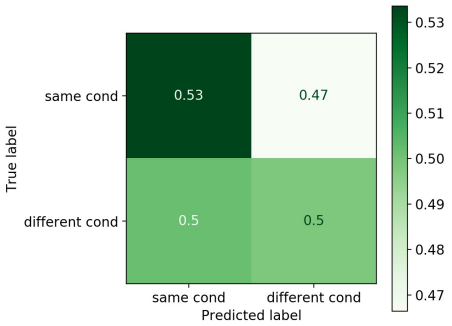


Figure 16: ScanMatch confusion matrix for hand raising condition

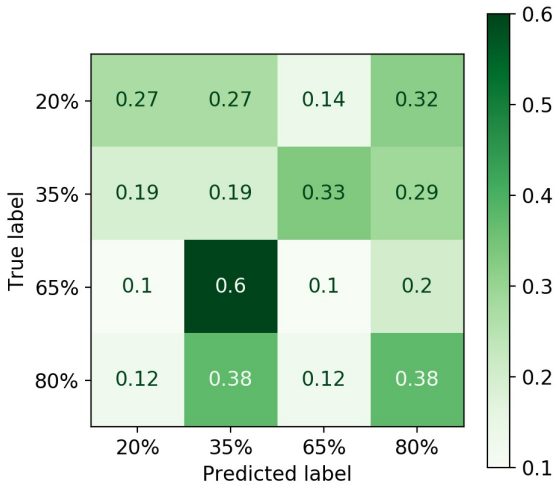


Figure 17: Confusion matrix for hand raising condition from 2-gram SVM classification

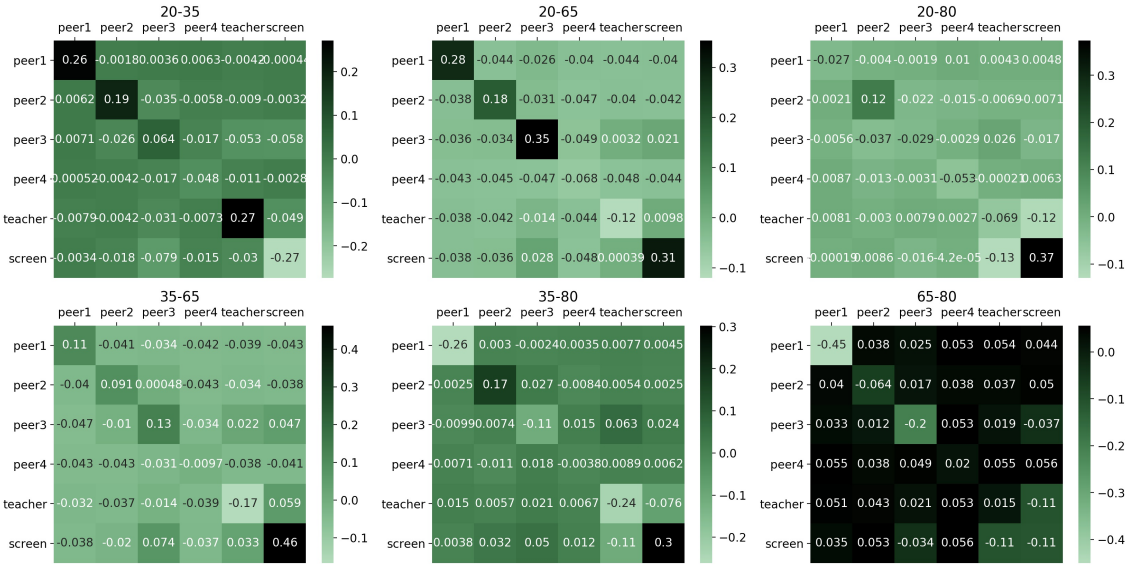
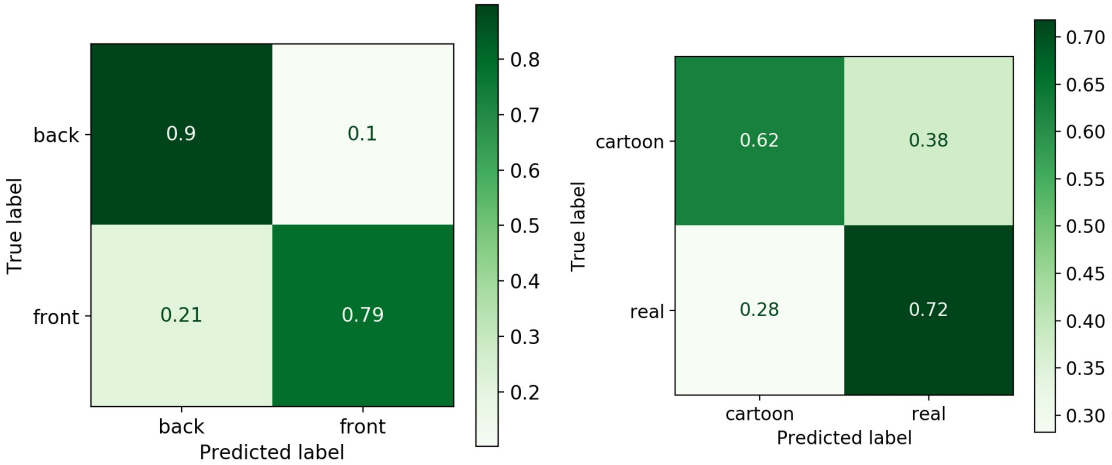
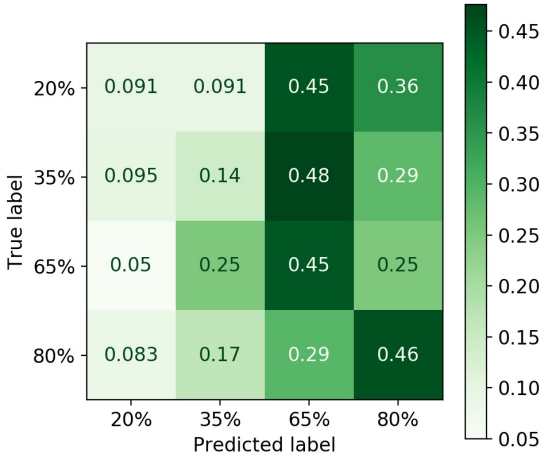


Figure 18: Feature matrices for hand raising condition from 2-gram SVM classification



(a) Sitting position

(b) Avatar style



(c) Hand raising

Figure 19: Confusion matrix (CM) of SVM classification with 3-gram transitions

Functions for ScanMatch and SubsMatch algorithm in Python

```
def string_lst(dat, start, end, fixation_time, interval, ooi_lst):
    """
    dat = pandas dataframe with variables 'time' and 'object'
    start and end point according to the time variable
    fixation_time = minimal time spent on an object to append it to the list
    interval  = time until repeating the same letter
    ooi_lst = list of letters representing the OOI

    return: string sequence of OOI in temporal order
    """

    #Determine start and end and reset index
    cond = np.logical_and(dat['time']>=start,dat['time']<=end)
    dat = dat[cond].reset_index()

    #delete all rows with no OOI
    cond = dat['object'].isin(ooi_lst)
    dat = dat[cond]


    string_lst = [] #list with the index of the ooi from ooi_lst
    t=0
    while t<len(dat):
        start_obj = dat['object'][t]
        current_t = dat['time'][t]
        while t+1 < len(dat) and dat['object'][t] == start_obj:
            t+=1
            if dat['time'][t]-current_t >=interval: #add repeating letters
                string_lst.append(dat['object'][t])
                current_t = dat['time'][t]

        #add the current ooi if the fixation was long enough
        if (dat['time'][t]-current_t)>=fixation_time:
            string_lst.append(start_obj)

        t+=1
    return string_lst ,len(string_lst)
```

Table 9: Code for creating string sequences (according to ScanMatch)

```
def trans_matrix(dat, start, end, transition_time, fixation_time, interval, ooi_lst, ngrams=2):
    """
    dat = pandas dataframe with variables: 'time', 'object'
    start and end point according to the time variable
    transition_time = the max time length before introducing a gap transiton
    fixation_time = minimal time spent on an object to append it to the list
    interval  = time until repeating the same letter
    ooi_lst = list of letters representing the OOI
    ngrams = length of contiguous sequences of items considered for the analysis (default 2)

    return: list of transition matrices one for each ngram-dimensions
    """

    #Determine start and end and reset index
    cond = np.logical_and(dat['time']>=start,dat['time']<=end)
    dat = dat[cond].reset_index()
    #append gap to ooi_lst
    ooi_lst.append('gap')

    scan_lst = [] #list with the index of the ooi from ooi_lst
    t=0
    while t<len(dat):
        start_obj = dat['object'][t]
        start_t = dat['time'][t]
        current_t = dat['time'][t]
        while t+1 < len(dat) and dat['object'][t] == start_obj:
            t+=1
            if dat['time'][t]-current_t >=interval: #add repeating letters
                scan_lst.append(ooi_lst.index(dat['object'][t]))
                current_t = dat['time'][t]
        #add the current ooi if the fixation was long enough
        if (dat['time'][t]-current_t)>=fixation_time:
            scan_lst.append(ooi_lst.index(start_obj))

        if t+1 < len(dat):
            next_t = dat['time'][t+1]
        else:
            next_t = dat['time'][t]

        if next_t-dat['time'][t]>=transition_time: #add the gap if the transition was to long
            scan_lst.append((len(ooi_lst)-1))
        t+=1

    #build transition matrices with correct dimensions and length
    matrix_lst = []
    for ngram in range(2,ngrams+1):
        s = []
        for i in range(ngram):
            s.append(len(ooi_lst))
        transition_matrix = np.zeros(tuple(s))
        matrix_lst.append(transition_matrix)

    #fill the matrix lists with info from scan_lst if index list has filled up to ngram size
    for t in range(len(scan_lst)):
        for m in range(0,ngrams-1):
            if t+m+2<=len(scan_lst):
                index_lst = scan_lst[t:t+m+2]
                lst = list(np.expand_dims(index_lst, axis=1))
                matrix = matrix_lst[m]
                matrix[lst]+=1
                matrix_lst[m] = matrix

    return matrix_lst
```

Table 10: Code for creating transition matrices (according to SubsMatch)