Virtual Reality
https://doi.org/10.1007/s10055-023-00804-0

ORIGINAL ARTICLE

MAC: multimodal, attention-based cybersickness prediction modeling in virtual reality

Dayoung Jeong[1] · Seungwon Paik[2] · YoungTae Noh[3] · Kyungsik Han[4]

Received: 7 March 2022 / Accepted: 4 May 2023

Abstract
Cybersickness is one of the greatest barriers to the adoption of virtual reality. A growing body of research has focused on identifying the characteristics of cybersickness and finding ways to mitigate it through the utilization of data from VR content, physiological signals, and body movement, along with artificial intelligence techniques. In this work, we extend prior research on cybersickness prediction by considering the role of different data modalities. We propose a novel deep learning model named multimodal, attention-based cybersickness (MAC), which learns temporal sequences and characteristics of video flows, eye movement, head movement, and electrodermal activity. Based on data collected from 27 participants, we demonstrate the effectiveness of MAC, showing an F1-score of 0.87. Our experimental results further show not only the influences of gender and prior VR experience but also the effectiveness of the attention mechanism on model performance, emphasizing the importance of considering the characteristics of data types and users in cybersickness modeling.

Keywords Virtual reality · Cybersickness · Deep learning · User characteristics

1 Introduction

Cybersickness is one of the key problems that must be solved through technological advancements in virtual reality (VR) in order to provide a good user experience and a safe virtual environment (Davis et al. 2014; Keshavarz et al. 2015; Rebenitsch and Owen 2016; Weech et al. 2019). Cybersickness has symptoms similar to motion sickness that may cause eye fatigue, nausea, and disorientation. Most users experiencing cybersickness find it difficult to continue using VR, and in some cases, dizziness may persist even after VR use is halted.

Much research has been conducted to determine the characteristics of cybersickness and find ways to mitigate it. Some research has focused on identifying the relationship between visual features (e.g., speed of content, color changes, visual objects, field of view, and latency) that can be extracted from VR content and the levels of cybersickness, as measured by questionnaires, such as the Simulator Sickness Questionnaire (SSQ) (Kennedy et al. 1993) and the Fast Motion Sickness Scale (FMS) (Keshavarz and Hecht 2011). Other research has collected information, such as heart rate (HR), breathing rate (BR), skin temperature (SKT), electrodermal activity (EDA), electrocardiogram (ECG), and electroencephalography (EEG), using various physiological sensor equipment, to investigate those factors' relationships with cybersickness (Draper et al. 2001; Groth et al. 2021; Dennison et al. 2016). Recently, a growing body of research has employed machine or deep learning (Jeong et al. 2019; Wang et al. 2019; Islam et al. 2020; Kim et al. 2018, 2019, 2020; Bos et al. 2008; Balasubramanian and Soundararajan 2019), aiming to build a model that learns the characteristics of VR content or sensor data

✉ YoungTae Noh
ytnoh@kentech.ac.kr

✉ Kyungsik Han
kyungsikhan@hanyang.ac.kr

Dayoung Jeong
dayoungjeong@hanyang.ac.kr

Seungwon Paik
seungwon.paik@lge.com

[1] Department of Artificial Intelligence, Hanyang University, Seoul, Republic of Korea

[2] LG Electronics, Seoul, Republic of Korea

[3] School of Energy Engineering, Korea Institute of Energy Technology, Naju, Republic of Korea

[4] Department of Data Science, Hanyang University, Seoul, Republic of Korea

Published online: 24 May 2023

when cybersickness occurs. Research has also proposed a time series-based model that learns the temporal changes of data (Jin et al. 2018; Wang et al. 2019; Lee et al. 2019; Islam et al. 2020). Although some of the learning components in modeling (e.g., datasets, feature types, time window sizes, and cybersickness levels) were slightly different between experiments, such research has demonstrated the possibility of predicting cybersickness (Jeong et al. 2019; Kim et al. 2018, 2019; Martin et al. 2020; Islam et al. 2020, 2021).

The objective of our research was to expand the present understanding of cybersickness by developing cybersickness prediction model. Through a literature review, we ascertained that various data modalities had not yet been fully explored in cybersickness modeling, leaving the role of each data modality unknown. This represented an important research gap since, for a model with multiple modalities, it is important to consider the different influences of each modality on cybersickness. Each modality needs to be prioritized to a different degree depending on its relationship with other modalities and time. We also found that, although the level of cybersickness often varies across individuals, research is somewhat lacking on how we should understand user characteristics in cybersickness modeling, along with a discussion of the application directions of cybersickness models.

In this paper, we propose a multimodal, attention-based cybersickness (MAC) prediction model that learns the temporal sequence of the data collected by four different types of data modalities (i.e., video flow, eye movement, head movement, and electrodermal activity) and weights a dynamic representation of features to capture the context of the inputs (You et al. 2016). To that end, MAC employs two attention subnetworks—individual convolutional and bidirectional long short-term memory (BiLSTM)—that separately characterize individual data modalities and temporal sequences. The term "attention" in MAC (which is an artificial neural network) refers to mechanisms designed to mimic cognitive attention. In humans, attention is limited and, therefore, distributed to necessary tasks and multimodal sensory signals as required (Lindsay 2020). The attention mechanism in our current machine learning model reflects the contribution of information from each modality (relative to the others) to the experience of cybersickness as an attention weight. This is a novel development since an attention mechanism for cybersickness prediction has not been applied in prior research. The model is intended to be applied to predict a possible occurrence of cybersickness so that the VR system can take proactive action to limit or prevent cybersickness. To develop the model, we collected datasets (i.e., multimodal data and a level of cybersickness) through a user study with 27 participants in which they watched the 360-degree VR videos.

The experimental results show that MAC yielded the best performance (an F1-score of 0.87) compared with other widely used algorithms for sensor data (e.g., support vector machine (SVM), convolutional neural network (CNN), BiLSTM, and CNN-BiLSTM). Among the modalities, eye movement received the highest attention weight. Additionally, when the participant was male or had prior VR experience, the model achieved F1-scores of 0.91 and 0.83, respectively. These results highlight the importance of considering user characteristics in cybersickness prediction.

In summary, this paper makes the following contributions.

- We present a novel deep learning model named multimodal, attention-based cybersickness (MAC) that not only accounts for various sensor data modalities but also prioritizes the importance of data modalities through an attention mechanism. The application of attention for cybersickness prediction has not previously been researched.
- We demonstrate the effectiveness of MAC and articulate the role of data modalities in cybersickness prediction via an in-depth comparative analysis.
- We show the influence of demographic characteristics on cybersickness prediction and the different application of attention on each data modality. We discuss approaches to model improvement and application.

2 Related work

Many studies have been conducted on cybersickness in order to improve the user experience of VR. Researchers are using various methods to identify ways to understand cybersickness, such as the collection and analysis of sensor data, the investigation of visual factors, and the development of prediction models based on machine/deep learning. In this section, we first introduce cybersickness theories and explain the theoretical rationale of utilizing body signals to understand cybersickness. Then, we present two primary directions of data-driven research on cybersickness, as summarized in Table 1. Considering those directions, we propose a novel cybersickness prediction model that not only accounts for various sensor data modalities but also prioritizes the importance of data modalities through an attention mechanism. We demonstrate the effectiveness of our approach through an in-depth comparative analysis.

2.1 Summary of cybersickness theories

The perception of self-motion involves the integration of multisensory information; however, there are cases in which the sensory feedback received from these different sources is conflicting. Motion sickness is a common consequence of sensory mismatch and has been explained by several sensory conflict theories (e.g., sensory rearrangement theory,

Table 1 Summary of cybersickness prediction modeling research

Literature	VR exposure	Sensor data	# of Participants	Model	Predicted	Performance
Dennison et al. (2016)	Active	E/H/P	20	Regression	SSQ score	0.29 (adjusted R^2 score)
Jeong et al. (2017)	Passive	V	28	Visual comfort assesment	Comfort score	0.90 (PLCC)
Kim et al. (2017)	Passive	V	15	Convolutional autoencoder	SSQ score	0.92 (PLCC)
Jin et al. (2018)	Active	V/H	24	LSTM	SSQ score	0.86 (R^2 score)
Padmanaban et al. (2018)	Passive	V	96	Decision Tree	SSQ score	12.00 (RMS)
Lee et al. (2019)	Passive	V	No info	3D-CNN	Degree of motion sickness	8.49 (RMSE)
Jeong et al. (2019)	Passive	P	25	CNN/DNN	CS level (2 classes)	0.98(CNN)/0.98(DNN) (Accuracy)
Balasubramanian and Soundararajan (2019)	Passive	V	43	Ridge Regression	Discomfort score	4.37 (RMSE)
Kim et al. (2018)	Active	V	20	Deep generative	SSQ score	0.88 (PLCC)
Lee et al. (2019)	Passive	V/P	20	Fully connected layers	SSQ score	0.83 (PLCC)
Kim et al. (2019)	Passive	V/P	202	CNN-RNN	CS level (5 classes)	0.89 (Accuracy)
Martin et al. (2020)	Active	P	103	Random Forest	CS level (3 classes)	0.87 (Accuracy)
Islam et al. (2020)	Passive	P	31	CNN-LSTM	CS level (3 classes)	0.98 (Accuracy)
Kim et al. (2020)	Passive	V	154	CNN	VIMS score (5 classes)	0.86 (PLCC)
Palmisano et al. (2020)	Passive	H	26	Differences in virtual and physical head pose (DVP)	CS level	0.0001 (p-value) (sig. relationship with cybersickness)
Islam et al. (2021)	Active	V/E/H	26	CNN-LSTM	CS level (4 classes)	0.87 (Accuracy)
Chang et al. (2021)	Passive	E	26	Regression	SSQ score	0.34 (Total variance)
Islam et al. (2021)	Passive	P	23	DNN	CS level (11 classes)	2.47 (RMSE)
Oh and Kim (2021)	Passive	P	20	Deep ensemble	CS level (3 classes)	0.96 (Accuracy)
Kundu et al. (2022)	Passive	P	31	Explainable boosting machine	CS level (2 classes)	0.99 (Accuracy)
Qu et al. (2022)	Active	P	9	LSTM-Attention	CS level (2 classes)	0.96 (Accuracy)

"VR exposure" refers to the type of VR content (Active means participants move the virtual environment, and Passive means participants watch VR videos at a fixed position). "Sensor data" refers to the modality used in the study

V, VR video content; E, Eye movement; H, Head movement; P, Physiological data. Note that four types of modalities are not exclusive but are identified based on our literature review. "Predicted" means target (dependent) variable; CS: cybersickness

subjective vertical conflict theory, and vection conflict theory). Cybersickness is considered a subtype of motion sickness (as they share similar symptoms, such as nausea, sweating, dizziness, and fatigue), and theories of motion sickness have been applied to understand cybersickness characteristics.

Sensory conflict theory is the most widely accepted theory of motion sickness and cybersickness (Reason and Brand 1975). This theory is based on the premise that discrepancies between the senses that provide information about the body's orientation and motion movement cause a perceptual conflict that the body cannot handle. The disconnect often occurs between the eyes and the vestibular system, which controls the functioning of the inner ear, overall balance, and the person's orientation in a physical space. Beyond that, in this section, we present several other key theories of motion sickness/cybersickness.

Sensory rearrangement theory by (Reason 1978) suggests that motion sickness is caused by sensory conflict (e.g., when the person's visual information is inconsistent with their available inner ear stimulation). This theory argues that sensory conflict only triggers cybersickness when it results in a neural mismatch (e.g., when the person's currently sensed motion is different from what they were expecting based on past experience). Sensory rearrangement theory was subsequently modeled by Oman (1982). The subjective vertical conflict (SV-conflict) theory by Bles et al. (1998) is a later variant of Reason's sensory rearrangement theory, stating that only neural mismatches that involve the subjective vertical will trigger motion sickness. Hence, according to SV-conflict theory, cybersickness should only be caused by conflicts between the currently "sensed vertical" (based on integrated information from the sensory organs) and the "expected vertical" (estimated based on prior experience and expectations). Meanwhile, another conflict theory is vection conflict theory (Hettinger et al. 1990), which proposes that sensory conflict only triggers motion sickness when the motion stimulation generates vection (an illusion of self-motion). The DVP (i.e., differences in one's virtual and physical head pose) hypothesis is the most recent variant of these sensory conflict theories (Palmisano et al. 2020). It attempts to predict cybersickness based on the amount of sensory conflict presented to the observer's sense organs. Unlike the other sensory conflict theories, the DVP hypothesis does not attempt to model the subsequent neural mismatches generated by these sensory input conflicts.

However, it should be noted that there are other explanations for motion sickness beyond sensory conflict, posing that sensory conflict is hypothetical rather than a fact. Riccio and Stoffregen (1991), who disagree entirely with the concept of sensory conflict, instead argue that motion sickness is caused by postural instability. Similarly, Ebenholtz et al. (1994), Ebenholtz (1992) stated that motion sickness is caused by excessive eye muscle traction, not by sensory conflict. Another account of motion sickness is that it is an automatic response to perceived poisoning, and therefore the symptoms of this sickness are actually the result of defensive hypothermia (Nalivaiko et al. 2014; Treisman 1977).

As these theories suggest, it appears that several factors influence motion sickness and cybersickness. Many studies have investigated the effect of sensory conflict between visual, vestibular, and body cues on the perceived timing of visual motion and the relationship between sensory conflict and sensory reweighting, to study specific characteristics of cybersickness and find ways to reduce cybersickness.

In that spirit, we aimed to build a cybersickness prediction model that could reflect multisensory characteristics based on the body signals that occur when experiencing VR content and potentially cause cybersickness. Body signals when experiencing VR can be recorded using a standard VR head mounted display (HMD) or additional sensor devices attached to the body (e.g., an Empatica E4 wristband, as used in our study). Many off-the-shelf sensor data collection devices can be used in VR; thus, we reviewed the types of sensor data used in previous studies to investigate cybersickness, which provided a basis for our selection of sensor data.

2.2 Correlation between cybersickness and sensor data

Researchers have evaluated a user's state by identifying the relationship between cybersickness and data collected during the VR experience, such as VR video content, eye movement, head movement, and physiological data (Dennison et al. 2016). One such evaluation approach is to identify the relationship between cybersickness caused by watching a video screen and sensor data collected through the HMD. For example, Groth et al. (2021) identified the user's eye gaze using an eye tracker built into the HMD, minimized the movement of the field of view (FOV) and reduced cybersickness by applying visual techniques, such as blurring or opaque occlusion around the user's FOV. Bala et al. (2018) suggested that an independent background and restricted FOV in VR content are technical factors that reduce sickness and showed the relationship between head movement and FOV size. Jeong et al. (2019) discussed common patterns that cause cybersickness based on a characteristic analysis of videos. Chang et al. (2021) reported that the cybersickness level can vary depending on the characteristics of VR content and observed a unique eye movement associated with cybersickness. Similarly, Lopes et al. (2020) investigated the relationship of pupil position and eye blinking patterns with cybersickness. In their evaluation of two groups of participants, they suggested that the group with high sickness had a higher blink frequency per minute than the group

without sickness. Nalivaiko et al. (2014) focused on sweating as one of the symptoms of motion sickness. EDA is a sensitive approach to quantitatively evaluate sweating, and the researchers showed that EDA data can be used to detect motion sickness, which manifests as the EDA data value increases. Palmisano et al. (2022) suggested that DVP is a major cause of cybersickness. To test that, 22 participants rotated their heads (roll, pitch, and yaw) while viewing a virtual room with a display delay deliberately set by the researcher. When the DVP data collected were compared with the participants' cybersickness levels, the levels were found to have increased continuously according to the amplitude and variability of the DVP. Taking a similar approach, Kim et al. (2020) discussed head orientation in relation to cybersickness, based on a study with 30 participants. They observed that when the difference in head orientation between the virtual environment and the real physical environment was large, posture became unstable and cybersickness was amplified, which the researchers proposed to reflect both sensory conflict and postural instability theories.

Similarly, studies have shown the correlation between cybersickness and physiological data, such as EDA, ECG, HR, and BR (Jeong et al. 2019; Lee et al. 2019; Kim et al. 2019; Martin et al. 2020; Islam et al. 2020; Dennison et al. 2016). For example, Bosser et al. (2006) showed a high correlation between vasovagal syncope (a condition that leads to fainting) and motion sickness. Islam et al. (2020) found a significant relationship between the FMS score and the change in the mean percentage of physiological data. Jeong et al. (2019) highlighted that EEG also pertains to cognitive activity, and its spectrogram was used as a data feature to classify cybersickness. Dennison et al. (2016) examined whether changes in physiological signals caused by HMD use could be applied to predict cybersickness, finding that a combination of neurophysiological and non-physiological measures could be effective in that regard. McHugh et al. (2019) suggested that real-time recording using a physical dial, a Surface Dial device by Microsoft, can provide more accurate cybersickness levels. Jung et al. (2021) hypothesized that cybersickness could be reduced by delivering motion on real ground, similar to that in VR driving simulations. HR and EDA were collected from 22 participants using E4 wristband and pupil diameter was recorded using HTC VIVE Pro Eye; only EDA was found to be associated with cybersickness. Magaki and Vallance (2020) confirmed that physiological signals can be used to detect cybersickness when performing tasks in VR. After measuring several types of signals (e.g., EDA, SKT, HR, blood volume pulse (BVP), and accelerometer) from 16 participants using E4 wristband, peak EDA and skin conductance response (SCR) were found to be indicators of cybersickness. Gavgani et al. (2017) collected physiological signals (HR, BR, and EDA) from 14 participants who had performed 15 min of VR rollercoaster riding per day for three days. They confirmed that each signal was correlated with cybersickness and that the gradual change in EDA could objectively quantify the sickness.

Based on the insights/findings of prior studies on the relationship between sensor data modalities and cybersickness, we decided to use four types of data modalities (i.e., video, eye movement, head movement, and electrodermal activity) and investigate their association with and influence on cybersickness.

2.3 Cybersickness prediction

As an extension of the research in the previous section (Sect. 2.2), a growing body of research has proposed cybersickness prediction models using machine learning or deep learning. Through such a model, it is expected that a possible occurrence of cybersickness can be predicted in advance and that the VR system can then take proactive measures to prevent cybersickness. For example, Islam et al. (2020) proposed a CNN-LSTM model using physiological data (HR, EDA, and BR) and achieved a mean accuracy of 0.98 (three levels of cybersickness). Jeong et al. (2019) extracted cybersickness-related features through EEG analysis and demonstrated the effectiveness of a CNN/deep neural network (DNN) model with an accuracy of 0.98/0.98 (two levels). Martin et al. (2020) built a machine learning model based on BVP and EDA collected during a 30 min VR experience, showing it to have an accuracy of 0.85 (three levels). Islam et al. (2021) presented Cybersense, a framework that collects physiological signals and transmits information after a DNN model predicts the cybersickness level ranging between 0 and 10. The DNN model was trained on physiological signals collected from 23 participants, and the root mean square error (RMSE) was 2.47. Oh and Kim (2021) investigated whether physiological signals (heart rate variability (HRV) and BR) could be used to classify cybersickness. The authors proposed a deep ensemble model that stacks SVM, k-Nearest Neighbors, Random Forest, and AdaBoost and classifies cybersickness into three levels (neutral, non-cybersickness, and cybersickness). The authors collected data from 20 participants, and the ensemble model showed an accuracy of 0.96. Kundu et al. (2022) predicted cybersickness (two levels) using a dataset of physiological signals (HR, BR, EDA, and HRV). One of the machine learning models, the Explainable Boosting Machine (EBM), showed an accuracy of 0.99. Qu et al. (2022) predicted cybersickness (two levels) using a dataset of physiological signals (HR, BR, EDA, and HRV), finding that the EBM again showed an accuracy of 0.99.

Similar to the research on modeling cybersickness using physiological sensor data, some studies have applied features extracted from VR videos. For instance, Kim et al.

(2018) proposed VRSA Net, a cybersickness prediction model trained on motion patterns for 360-degree VR videos similar to everyday life with non-exceptional motions (e.g., normal walking) and exceptional motions (e.g., racing and roller coaster ride). Their model achieved an accuracy of 0.88. Jeong et al. (2017) proposed a deep learning model utilizing the features of left view, right view, saliency absolute disparity, and saliency absolute differential disparity of VR video. Their proposed model showed a high Pearson's linear correlation coefficient (PLCC) of 0.90 with visual discomfort in stereoscopic viewing. Balasubramanian and Soundararajan (2019) built a database of 100 videos and 4000 responses to use for the automatic assessment of cybersickness. A regression model was proposed based on features related to camera movements (e.g., shake, depth, and velocity) and the participants' response scores, with the highest correlation (0.84 coefficient) was shown in the case of ridge regression. Kim et al. (2017) used a convolutional auto-encoder and decoder for the exceptional motion that induces cybersickness, which showed a PLCC of 0.92. More recently, Kim et al. (2020) proposed a model called DeepVIMSP, which they developed by combining the VR video features extracted from temporal and spatial sequences using CNN with a neurological mechanism, achieving an accuracy of 0.86 (five levels of cybersickness). Lee et al. (2019) developed a three-dimensional (3D) CNN prediction model using optical flow, disparity, and saliency of the video, showing both cybersickness and a PLCC of 0.84. Finally, Padmanaban et al. (2018) used single frames, disparities, and optical flow of VR videos to predict the degree of sickness induced by VR content. Sensor data were collected and the SSQ was completed by 96 participants. The Decision Tree regressor showed an RMSE of 12.00.

Taking another approach, some studies have utilized data from multiple modalities to model cybersickness. Jin et al. (2018) collected video and head movement, using those to measure the level of cybersickness, and thus built three machine learning models (CNN, Long-Short Term Memory Recurrent Neural Networks (LSTM-RNN), and Support Vector Regression (SVR)) and compared their performance. They found that the LSTM-RNN model performed best with an R^2 of 0.86. Lee et al. (2019) presented a deep learning framework of a ConvLSTM structure, to be used for individual cybersickness prediction using visual and physiological features (e.g., EEG, ECG, and EDA). The model result was a PLCC of 0.85 when both kinds of features were applied, showing a high association with cybersickness. Islam et al. (2021) developed a cybersickness prediction model based on a 3D-CNN and CNN-LSTM using eye movement, head movement, video, optical flow, and disparity data. When testing their model, they found that using only video, optical flow, and disparity achieved an accuracy of 0.57, while using eye and head movements achieved the best performance with an accuracy of 0.87 (for four levels of cybersickness). In another study, Dennison et al. (2016) recorded eye and head movements and physiological signals from 20 participants, linking those to the difference in the level of cybersickness between the experience through the display and that through the HMD. When wearing the HMD, the regression model that predicted SSQ scores showed an adjusted R^2 score of 0.29. Kim et al. (2019) estimated the cognitive state from VR content using EEG data that represented information about brain activity. The authors collected EEG data and VR content from 202 participants and predicted cybersickness in five classes (extreme, strong, neutral, mild, and comfortable). The model showed an accuracy of 0.89. Palmisano et al. (2020) attempted to objectively estimate/quantify the DVP during HMD VR exposure by using time series data. To do so, they predicted levels of cybersickness by correlating the DVP with head frequency conditions. Chang et al. (2021) proposed a regression model to predict subjective cybersickness levels based on eye movement. The total variance of the regression model in predicting SSQ scores (ranging between 0 and 3) was 0.34.

Recently, a technical approach that is increasingly being used in deep learning is attention. In neuroscience and psychology, attention is an approach to understanding how limited resources are focused. Attention can be distributed across modalities to perform tasks that require the integration of multiple sensory signals (Lindsay 2020). An attention mechanism in an artificial neural network effectively represents variables' relative importance, as indicated by the iterative reweighting of vectors of the input data during the training of the neural network model, dynamically highlighting different components of a pre-processed input as they are needed for output generation. Bahdanau et al. (2014) first proposed an attention mechanism for artificial neural networks. They claimed that information about each data element can be reflected in the dataset as a whole and that data can be selectively retrieved to generate an output. Deep learning models that do not use an attention mechanism also learn the relationship between sensors according to the target variable. However, when inputs of variable length, size, and structure become long, large, or complicated, there can be a loss of information as the information is aggregated into a fixed-size vector. Applying an attention mechanism to a deep learning model allows it to focus on which data are important while minimizing information loss with limited resources.

Notably, attention in machine learning does not always adhere to biological attention; however, attention mechanisms hold the potential to improve the training performance of neural networks on various tasks with multimodal inputs. Nonetheless, in our literature review, we noted that the cybersickness prediction models presented in prior studies did not fully exploit the potential advantages of attention

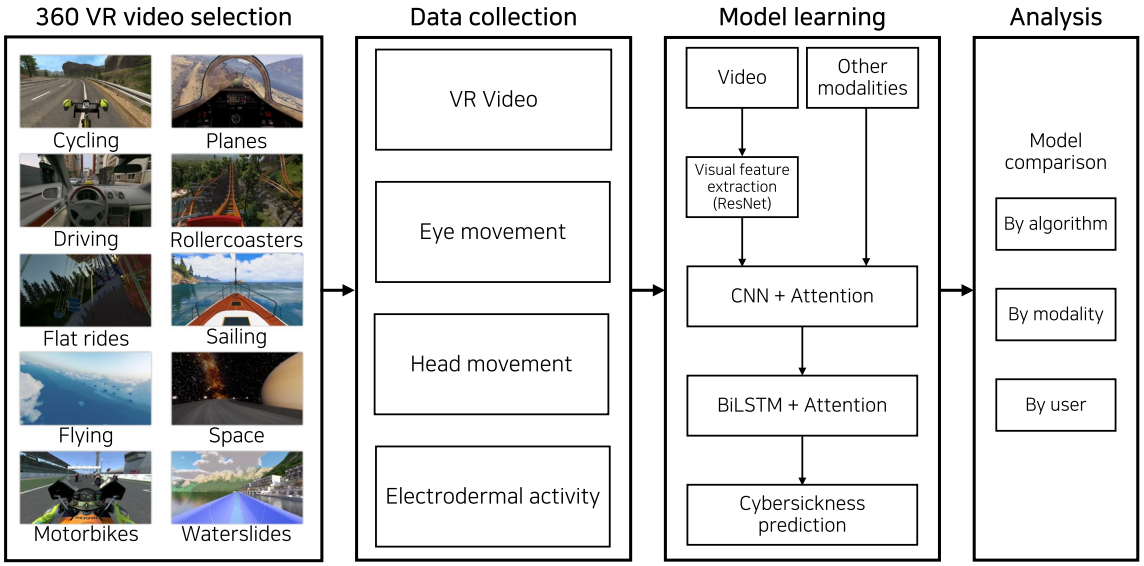

Fig. 1 The overall research procedure. We selected two videos for each of 10 video topics. Participants watched a total of 20 videos during the experiment. Video screen (viewed by a user through the HMD), eye movement, head movement, and electrodermal activity were collected. We verified our model for cybersickness prediction

mechanisms. Most prior studies used traditional machine learning algorithms such as Random Forest and SVM, and some recent studies applied deep learning algorithms such as CNN and LSTM. While the proposed model structures were reasonable, there are opportunities to advance those and thereby achieve better performance.

In summary, many researchers believe cybersickness is caused by sensory conflict (Reason and Brand 1975; Reason 1978; Bles et al. 1998; Hettinger et al. 1990; Palmisano et al. 2020), which occurs when multiple sensory signals collide, or when the input signal does not match the information learned pre-attentively. As we explained in Sect. 2.1, while we also noted the evidence for other theories (e.g., postural instability as a necessary precursor to motion sickness (Riccio and Stoffregen 1991), we chiefly elected to base the development of our model on sensory conflict theory. After exposure to sensory conflict, the neural memory reweights the information about cybersickness. In similar way, data weights are repeatedly updated as a model is trained, and when an attention mechanism is applied to a deep learning network, that reweighting can be flexibly trained by assigning attention weights. Hence, we approach deep learning based on a grounding in cybersickness theories, which we applied to develop a cybersickness prediction model with an attention mechanism.

Table 2 Summary of 27 participants' demographic information

Age	Gender		VR experience		MSSQ
26.2 ±3.3	Male	16	Yes	19	10.0 ± 8.8
	Female	11	No	8	

VR experience is about whether a participant experienced VR before. MSSQ is an abbreviation of Motion Sickness Susceptibility Questionnaire (Golding 1998) that measures the degree of motion sickness that a participant feels in his/her daily life

3 Study procedure

The overall procedure of our study is illustrated in Fig. 1. Our study was reviewed and approved by the internal institutional review board at the authors' university (IRB#202105-HS-001).

3.1 360-degree VR video selection

We used 360-degree VR videos in the experiment to provide participants with an immersive experience (Anwar et al. 2020; Shahid Anwar et al. 2020). We prepared 10 topics of videos (i.e., cycling, driving, flat rides, flying, motorbikes, planes, roller coasters, sailing, space, and water slides) to consider visual factors from various content and to train the model accordingly. Each video topic has two videos, resulting in a total of 20 videos for the experiment. The video

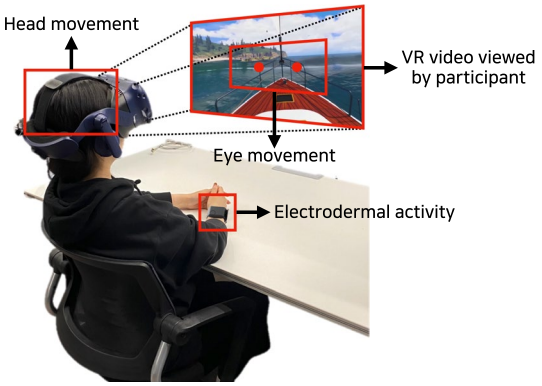

Fig. 2 Study participation and types of data collection

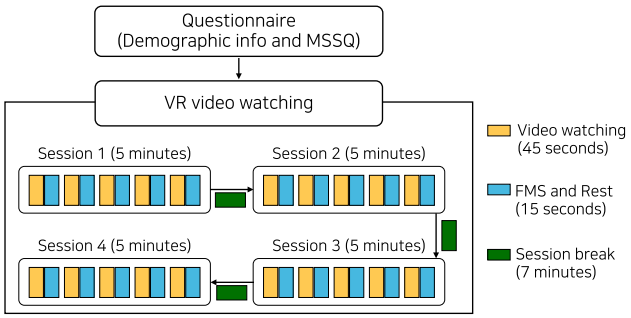

Fig. 3 The data collection procedure. During the video-watching phase, the participants were asked to watch videos and answer the FMS. We ran four sessions to give the participants enough time to rest before watching the next video. The study took approximately 46 min on average

content was set to 4K resolution and 30 frames per second (FPS) .[1]

3.2 Data collection and feature extraction

The HTC VIVE Pro Eye [2] was used to collect the video flow, eye movement, and head movement. The Empatica E4 wristband [3] was used to collect electrodermal activity. It received regulatory compliance from Europe, the USA, and Japan and has been used as a valid sensing device.[4] Research in cybersickness has also started to use the E4 wristband and proved the association between the sensor signals from it and cybersickness (Jung et al. 2021; Magaki and Vallance 2020). We ran the VR videos on a Unity 3D over a Windows 10 PC with Intel Core i7 and GeForce RTX 2070.

We recruited 27 participants via university bulletin boards and word-of-mouth (Table 2). The experiment consisted of two phases: (1) survey answering and (2) VR video watching. Before starting the experiment, we explained the goal and procedure of our study to each participant. We explained that they could opt-out of the study anytime. We then obtained informed consent from each participant.

First, each participant was asked to answer demographic questions (age, gender, and prior VR experience) and to complete the MSSQ before starting the experiment. The average MSSQ score of the participants was 10.01 (SD = 8.82), which is found to be similar to that mentioned in previous studies (Golding 1998; Bala et al. 2018). The participants were instructed to describe their condition when they experienced severe cybersickness and were assured that they could halt their participation at any time. We gave the participants enough time to become familiar with the HMD and the E4 wristband.

Second, the participants were asked to sit on a chair and to watch the VR video in their comfortable positions (Fig. 2) (Litleskare and Calogiuri 2019). The video-watching phase consisted of four sessions. Five VR videos were played during each session (Fig. 3). We provided the participants with a break by referring to the designs of previous studies. For example, Kim et al. (2019) included 14 videos (30 s each) in one session and provided a 3-min break for each session. Jeong et al. (2019) watched six videos of about 1–5 min, with a 3-min break between each video. Thus, to minimize the effect of their experience in the previous video before watching the next video, the participants were given a 15-s break at the end of each video and a 7-min break at the end of each session. The participants were offered additional time to rest between videos and sessions. For each VR video, the VR screen viewed by the participants was recorded at 30FPS through Unity.

The participants who completed the experiment received a $10 gift card for their time and participation. The participants may have had different degrees of inherent motion sickness; thus, the level of cybersickness that each participant experienced while watching the videos during the experiment might also have varied. Thus, we asked the participants to indicate their level of cybersickness via the FMS after watching each VR video. The collected raw FMS data was in the form of imbalanced data. When building a prediction model using imbalanced data, problems such as overfitting are likely to occur. Thus, the FMS data was processed to solve this problem. We followed the same method in prior research (Islam et al. 2020, 2021) to label the level of cybersickness based on the FMS data from all participants. In our data, the first quartile of FMS distribution (Q_1) was 1.0, the second quartile (Q_2) was 4.0, and the third quartile (Q_3) was 7.0. As a result, the quartile converted FMS data consisted of reasonably balanced data with 7,335, 5,850, 5,265, and 5,850 samples for each class. The data was labeled as follows:

[1] The videos are available at http://tiny.cc/q04luz.

[2] https://www.vive.com/us/product/vive-pro-eye/overview/.

[3] https://www.empatica.com/research/e4/.

[4] https://www.empatica.com/en-int/research/e4/.

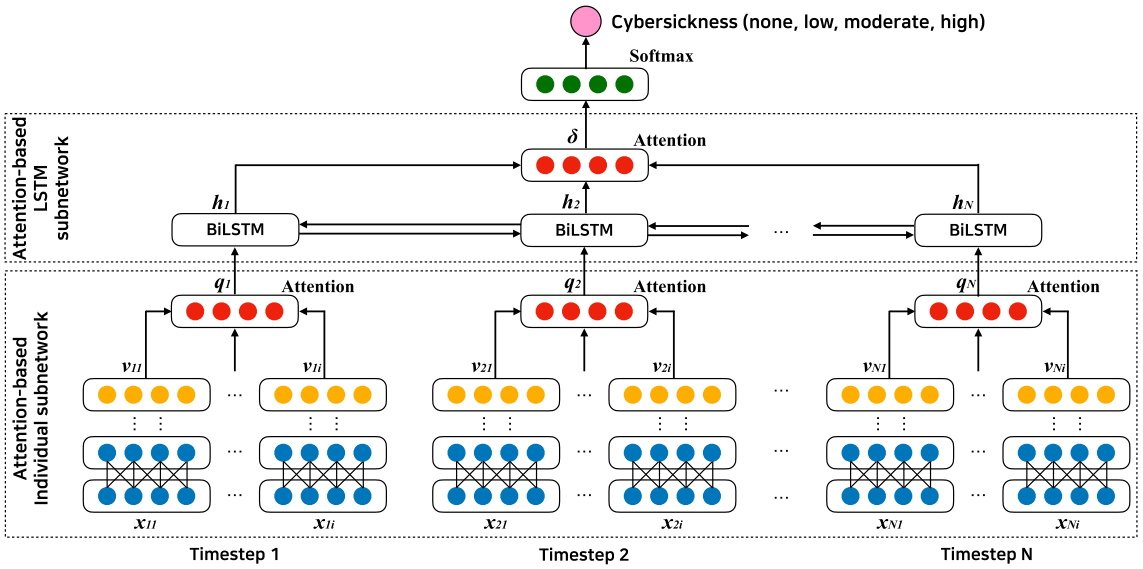

Fig. 4 The architecture of a multimodal, attention-based cybersickness (MAC) prediction model. The model consists of an attention-based individual convolution subnetwork, an attention-based BiLSTM subnetwork, and an output layer

$$Cybersickness\ level = \begin{cases} \text{None} & \text{if,} \quad 0 \le \text{FMS} \le Q_1 \\ \text{Low} & \text{if,} \quad Q_1 < \text{FMS} \le Q_2 \\ \text{Moderate} & \text{if,} \quad Q_2 < \text{FMS} \le Q_3 \\ \text{High} & \text{if,} \quad Q_3 < \text{FMS} \le 20 \end{cases} \tag{1}$$

For feature extraction, we resized each image frame to $3 \times 512 \times 512$ and applied a Gaussian filter for noise removal (Cai et al. 2017; Lee et al. 2009). The eye movement data consisted of 25 features, including gaze direction (x, y, z) for both eyes as well as gaze direction (x, y, z), gaze origin (x, y, z), pupil diameter, pupil position (x, y), number of blinks and eye openness for each eye (left and right). The head movement data consist of six features, including position (x, y, z) and rotation data (roll, pitch, and yaw). We considered a single electrodermal activity feature as a physiological datum that is frequently applied in VR research (Martin et al. 2020; Islam et al. 2020; Dennison et al. 2016), as highlighted in Table 1.

3.3 Model development

MAC, which employs various sensor signals and the attention mechanism for cybersickness prediction, was constructed grounded in theoretical perspectives, as explained in Sect. 2. MAC was built to learn repetitive sensory conflict by reweighting the relationship with various sensor data types through an attention mechanism. MAC consists of three key components as follows: (1) an attention-based individual convolutional subnetwork that considers the relative importance of each data modality to fuse modality-specific features, (2) an attention-based BiLSTM subnetwork that extracts the importance of timestep and fuses the hidden state of the BiLSTM, and (3) an output layer that uses a softmax function to obtain the probabilities for activity recognition. Figure 4 illustrates the detailed architecture of MAC.

3.3.1 Attention-based individual convolutional subnetwork

Attention-based individual convolutional subnetwork consists of a convolutional network and an attention network. The convolutional network was used to extract features from each data modality and consisted of two stacked convolutional layers and pooling layers. A batch normalization layer was applied at each layer to reduce internal covariate shift. The frequency representation of the ith sensor at time t, x_{ti} was passed to the convolutional network. Then, a feature vector v_{ti} was generated and used as the input to the attention network.

We employed an attention network to prioritize the importance of data modalities. The network takes the feature vectors of data modality $[v_{t1}, v_{t2}, ..., v_{ti}]$ as input and generates an attention weight for each modality. The hidden representation of v_{ti} was computed to get μ_{ti} with a sensor-level context vector w_1.

$$\mu_{ti} = tanh(W_1 v_{ti} + b_1) \tag{2}$$

Then a normalized weight α_{ti} was computed through a softmax function.

$$\alpha_{ti} = \frac{exp((\mu_{ti})^T w_1)}{\sum_i exp((\mu_{ti})^T w_1)} \tag{3}$$

where W_1, b_1, w_1 are parameters of the attention network. They are randomly initialized and jointly learned through the training phase. Then the vectors of all data modalities are fused by using their attention scores as weights in order to make a uniform feature representation vector q_t.

$$q_t = \sum_i \alpha_{ti} v_{ti} \tag{4}$$

3.3.2 Attention-based BiLSTM subnetwork

The output $[q_1, q_2, ..., q_N]$ is passed to a stacked LSTM structure (Greff et al. 2016). LSTM is a RNN architecture that remembers values over arbitrary intervals and deals with the vanishing gradient problem that can be encountered when training traditional RNNs.

$$f_t = \sigma(W_f \cdot [h_{t-1}, q_t] + b_f) \tag{5}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, q_t] + b_i) \tag{6}$$

$$\widetilde{C}_t = tanh(W_C \cdot [h_{t-1}, q_t] + b_C) \tag{7}$$

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \tag{8}$$

$$o_t = \sigma(W_o[h_{t-1}, q_t] + b_o) \tag{9}$$

$$h_t = o_t * tanh(C_t) \tag{10}$$

where f_t is a forget gate, i_t is an input gate, and o_t is an output gate. C_t is a cell state. A hidden state h_t is generated at each timestep. Standard RNN (including LSTM) uses the last timestep as a single representation for the whole input sequence. This generally leads to less consideration of the front part of the sequence for classification. Because the hidden state at each timestep may show a different level of impact on classification (in our case, occurrence of cybersickness), we applied the attention mechanism again to calculate the weighted average sum of all hidden states.

Given all hidden states $H = [h_1, h_2, ..., h_N]$ (h_t refers to a hidden state at timestep t), the attention for LSTM can be formalized as follows:

$$\gamma_t = tanh(W_2 h_t + b_2) \tag{11}$$

$$\beta_t = \frac{exp((\gamma_t)^T w_2)}{\sum_t exp((\gamma_t)^T w_2)} \tag{12}$$

$$\delta = \sum_t \beta_t h_t \tag{13}$$

where w_2 is a time-level context vector, β_t is a normalized weight through a softmax function, and δ is the uniform representation of the whole sequence computed based on the sum of all hidden states. Each hidden state is updated by its attention weights. W_2, b_2, w_2 are the parameters of the attention-based BiLSTM subnetwork which are randomly initialized and jointly learned during the training phase. We constructed a BiLSTM model that better learns the temporal characteristics of the data. BiLSTM has been found to be more efficient than unidirectional LSTM because it considers both past and future data through an interactive network (Huang et al. 2015).

3.3.3 Output layer

The output of attention-based BiLSTM subnetwork is calculated through an output layer using a fully connected layer and a softmax function to predict cybersickness.

$$\text{prediction} = \underset{a \in A}{\text{argmax}}(\text{softmax}(W_3 \cdot \delta + b_3)) \tag{14}$$

where A is the set of all data modalities. δ is transformed to the probability of each modality, and the prediction result is determined by searching modality with maximum probability.

3.4 Experiment setup

We implemented our model in Pytorch and trained it on a server with GeForce RTX 2070.

First, we extracted visual features for each image frame (30FPS) through a ResNet18 (He et al. 2016), a CNN that is 18 layers deep and has been widely used and demonstrated its effectiveness. The data consists of a set S of four data modalities in the form of time series data $S_t = \{V_t, E_t, H_t, P_t\}$, where V, E, H, and P refer to video, eye movement, head movement, and electrodermal activity, respectively. Each item in S_t is divided into a set of r time windows $W_t^a = \{w_{t1}^a, w_{t2}^a, ..., w_{tr}^a\}$ of a fixed length of T_w seconds (we set $r = 30$ and $T_w = 1$). S_t is then split into the training set, the validation set, and the test set with the ratio of 7:1:2 by chronological order.

Second, for model training, we used cross entropy for loss function and Adam for optimizer. The model was trained up to 500 epochs, and an early stop strategy was used with 20 times of patience to avoid overfitting. The best parameters of the model was selected through parameter tuning with the validation set (as a result, batch size = 64 and learning rate = 0.001). We used accuracy,

Table 3 The performance of models by model architecture. MAC achieved the highest performance, especially demonstrating the effectiveness of the attention mechanism in learning cybersickness

Model architecture	Accuracy	Precision	Recall	F1-score
SVM	0.67	0.67	0.67	0.67
CNN	0.59	0.68	0.63	0.63
BiLSTM	0.71	0.75	0.74	0.74
CNN-BiLSTM	0.75	0.75	0.74	0.74
A-INV (CNN-Attention-BiLSTM)	0.83	0.83	0.84	0.84
A-BiLSTM (CNN-BiLSTM-Attention)	0.82	0.80	0.82	0.81
MAC (CNN-Attention-BiLSTM-Attention)	**0.87**	**0.88**	**0.89**	**0.87**

Bold values indicate the model architecture with the highest performance

Table 4 The performance of the models by data modality (MAC was used). As a result, the model that uses all the modalities shows the highest performance

Data modality	Accuracy	Precision	Recall	F1-score
Video view	0.30	0.10	0.25	0.14
Eye movement	0.63	0.64	0.66	0.64
Head movement	0.51	0.46	0.46	0.45
Electrodermal activity	0.39	0.33	0.41	0.35
All modalities	**0.87**	**0.88**	**0.89**	**0.87**

Bold values indicate the highest performance achieved when applying all modalities

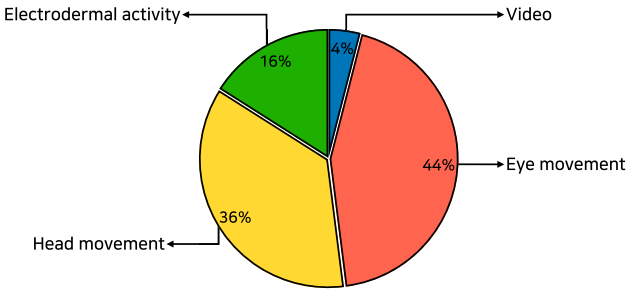

Fig. 5 The attention weight of MAC. The values refer to the percentage of the weights

precision, recall, and F1-score as the metrics for the performance of the model on the testing dataset. We used 5-fold cross validation.

We compared our model with the following algorithms.

- SVM (Cortes and Vapnik 1995): One of the traditional machine learning algorithms that has been used extensively for learning characteristics of sensor data, and demonstrated its effectiveness (Jahangiri and Rakha 2015).
- CNN (Schmidhuber 2015): A single CNN model with two convolutional layers, a pooling layer, and a fully connected layer. CNN has been widely used over sensor data (Um et al. 2017; Uddin and Hassan 2018).
- BiLSTM (Huang et al. 2015): A LSTM model that consists of forward and backward layers, which has shown its effectiveness compared with a single LSTM model.
- CNN-BiLSTM: Each data modality was processed with CNN and is passed to a BiLSTM layer, which has used in many studies that dealt with sensor data (Jeong et al. 2020).

We also considered two variants of our model by attention to see the impact of the attention in different conditions.

Table 5 The performance of models by user characteristics

Factor	Accuracy	Precision	Recall	F1-score
Gender				
Male	0.88	0.90	0.93	0.91
Female	0.87	0.85	0.87	0.86
Prior VR experience				
Yes	0.86	0.86	0.83	0.83
No	0.80	0.77	0.72	0.74
MSSQ				
High	0.83	0.72	0.85	0.83
Low	0.85	0.83	0.83	0.83

- A-INV: This model removes the attention layer in the BiLSTM subnetwork and instead uses the last hidden layer of BiLSTM (same as CNN-Attention-BiLSTM).
- A-BiLSTM: This model removes the attention layer in the individual subnetwork and instead uses naïve concatenation to fuse feature vectors of different data modalities (same as CNN-BiLSTM-Attention).

4 Results

4.1 Performance of models by architecture

Table 3 shows the performance of the models considered in the experiment. MAC yielded the highest performance among the models (an F1-score of 0.87). In particular, adding the attention layers played an important role in improving the model performance. More specifically, compared with the performance of CNN-BiLSTM, the performance of A-INV and that of A-BiLSTM increased by 10% and 7% (F1-scores), respectively. Interestingly, A-INV and MAC differed by 3% (F1-score), indicating that the attention layer in the individual convolutional subnetwork had a greater role in improving model performance. Overall, these results confirmed the effectiveness of prioritizing data modalities and learning their temporal characteristics on learning cybersickness.

4.2 Performance of models by modality

Table 4 shows the performance of MAC with different data modalities. When a single modality was applied to model training, the performance of the model that used eye movement was the highest (an F1-score of 0.64). The models that utilized head movement and electrodermal activity achieved F1-scores of 0.45 and 0.35, respectively. Additionally, the model that employed all the modalities yielded the highest performance (an F1-score of 0.87). Lastly, we analyzed the attention weights by the data modality to evaluate the impact of each modality on learning cybersickness. Figure 5 illustrates the results. The eye movement was the highest for the ratio of 44%. The head movement and electrodermal activity accounted for 36% and 16%, respectively. We also noted that the attention weight on the video modality was the lowest. These results indicate that, while each modality yielded a different influence on cybersickness, close attention should be paid to eye movement.

4.3 Performance of models by user characteristics

Table 5 shows the performance of MAC with different user characteristics. Previous studies have reported an association between demographic factors and cybersickness (e.g., older users and women tend to be more susceptible to cybersickness) (Davis et al. 2014; Weech et al. 2019; MacArthur et al. 2011). Based on this finding, we examined whether a group of users with similar demographic characteristics affects the performance of the model. We classified the participants into two groups for each demographic attribute (gender, prior VR experience, and MSSQ score). We obtained two groups of participants by gender (male: 16 and female: 11); two groups by prior VR experience (yes: 19 and no: 8); and two groups by MSSQ scores (low: 17 and high: 10, based on the median value). We did not consider age because most participants are in their 20s (mean: 26.2).

As a result, we found that the F1-score of the model using only male participants' data was 0.91, which is about 5% higher than that using only female participants' data. In terms of VR experience, the F1-score of the model using only the data of participants who previously had VR experience was 0.83, which is 9% higher than that using only the data of participants without prior VR experience. The low group and high group of MSSQ yielded the same performance (F1-score of 0.83).

5 Discussion

In this work, we showed the possibility of predicting cybersickness reasonably well through the development of a deep learning model. We demonstrated that the consideration of (1) data from multiple modalities that may have a different association with cybersickness, (2) the attention mechanism that assigns different weights to each data modality, and (3) the BiLSTM method that learns temporal sequences of the data, is effective in characterizing one's degree of cybersickness.

5.1 Improvement of cybersickness modeling

Although the role of the attention mechanism was proved to be useful to improve model performance, we observed a slightly smaller influence of the attention placed in the BiLSTM subnetwork than that in the individual convolutional subnetwork. Assigning different weights on the modalities was highly effective to learn a degree of cybersickness. Since such an important role of attention has been proved in our study, we can consider ways to improve the model by applying more advanced approaches. For example, the fast Fourier transform (Alsheikh et al. 2016) (presenting the changes in the energy content of a signal) applied to each data modality may improve characterization of the data and produce a better representation of data through attention. Video image data can also be better vectorized to improve its association with cybersickness using other state-of-the-art techniques, such as the vision transformer (Dosovitskiy et al. 2020), which has shown impressive results compared to state-of-the-art convolutional networks while requiring substantially fewer computational resources to train.

5.2 Consideration of user characteristics in modeling

One of the findings was the influence of demographic factors on cybersickness.

First, after the study, we checked the level of cybersickness (since each participant watched 20 videos, we averaged the level of cybersickness for each participant) according to gender. For men, none, low, moderate, and high were distributed at 29%, 21%, 22%, and 28%, respectively, and for women, 34%, 26%, 22%, and 18% were distributed, respectively. Although the percentage of women who answered that their cybersickness level was slightly higher (10%) than that of men, it can be considered that cybersickness levels are reasonably well distributed for both men and women. In this sense, it is somewhat difficult to conclude that women are more sensitive to cybersickness than men.

Second, regarding the distribution of cybersickness level according to prior VR experience, the participants who had VR experience showed distributions of none, low, moderate, and high as 24%, 20%, 25%, and 31%, respectively. For those who had no VR experience, the distributions of cybersickness were 48%, 31%, 14%, and 8%, respectively. From these results, we can see that the responses of 80% of participants without VR experience were distributed in a low level of cybersickness. From the perspective of model performance, the uniformity of the distribution of data was somewhat lower in the case of no VR experience, but the fact that it performed more than 80% indicates that the model had learned those data reasonably well.

Lastly, before beginning the VR experience, we asked the participants to answer a MSSQ. Based on the MSSQ responses, we divided the participants into either a low or a high group. For the low group, the percentages of the cybersickness levels of none, low, moderate, and high were 35%, 26%, 21%, and 18%, respectively. Those for the high group were 24%, 19%, 23%, and 34%, respectively. More than 60% of the participants in the low group showed none or low cybersickness levels. More than 55% of the participants in the high group showed moderate or high cybersickness levels. These results show that people who tend to be less vulnerable to motion sickness are less likely to be affected by cybersickness, and the opposite also holds.

Especially the influence of gender, although many studies have demonstrated that women are more sensitive to cybersickness (Davis et al. 2014; Weech et al. 2019; MacArthur et al. 2011; Munafo et al. 2017), it should also be noted that other studies have questioned the validity of women's self-reporting. For example, Jokerst et al. (1999) recorded gastric myoelectric activity to quantitatively measure nausea but did not find significant differences by gender. Stanney et al. (2003) demonstrated that overall women reported higher levels of cybersickness than men but did not actually show more nausea than men. These studies indicate that the relationship between gender and cybersickness is not always the same across studies but could vary depending on the condition/environment of the study and the participants (Sharples et al. 2008).

Considering these three results with Table 5, one interesting implication is that the distribution of cybersickness levels influences the model performance of cybersickness. The model with the data of male participants and that with the data of participants who had VR experience showed a better performance of cybersickness prediction than the model with the opposite conditions.

In addition, it is worth noting that there are research findings that cybersickness decreases as the user's VR experience increases. For example, Hill and Howarth (2000) confirmed that cybersickness was remarkably low on the fifth day when watching VR content for about 20 min every day for five days. Howarth and Hodder (2008) measured the cybersickness level before and after a VR racing game for seven days and measured the level of malaise (i.e., a general feeling of being ill or having no energy) at 1-min intervals. All participants reported that the levels of cybersickness and malaise decreased over time. The characteristics of users who have repeated VR experiences are called habituation. In addition, as explained in other responses, considering habituation in VR may change the findings or insights of our study. Hence, in a later study, we will check the correlation between cybersickness symptoms and sensor data expressed in users with characteristics of habituation.

Lastly, one possible application of our study results is to prepare and use the model that could be more effective to certain groups of users. For example, for users who are male or have VR experience, it may be more effective to use a model trained on data generated from a specific group with same demographic factors (i.e., specialized model) than the one trained on data from all users (i.e., population model). On the other hand, for female or user without VR experience, it may be effective to use a population model. Furthermore, we observed high performance in both the population model and the specialized model. Both ways of model development can be proceeded by continuously retraining each model, which is expected to lead a more effective detection of cybersickness in various user conditions and contexts.

These findings show the relationship between demographic factors and cybersickness. However, we acknowledge that our study participants may not represent all VR users. Thus, our study findings, insights, and speculations need to be further examined and verified through an additional study with more participants, which will be done in our future work.

5.3 Application of the model

It is also important to study the utilization of the cybersickness prediction model. Since our study showed the possibility of building a model with reasonably high performance by comprehensively considering user demographics and the multi-modalities that pertain to cybersickness, we expect that our model can be used in various VR scenarios. For example, previous research has verified the effect of reducing cybersickness by blurring peripheral vision (Groth et al. 2021). When cybersickness is predicted through our model, the VR system can adjust the peripheral vision of the VR content. This technology has the advantage of preventing cybersickness before it worsens. It can be expected that users will not be interrupted by cybersickness or may be more immersed in the content, thus enhancing the overall user experience. We can then measure one's perceived effectiveness of the model from the standpoint of model applicability. We expect that there will be more techniques to mitigate cybersickness and support the user experience, and the application of our model is expected to be expanded.

We are aware that the preparation of the sensor data and a level of cybersickness is needed to build a prediction model. One can consider collecting the data from scratch, but this would require a great amount of time and effort. Another consideration can be a reuse of existing models (e.g., our model) through transfer learning (Pan and Yang 2009), a method in which a model developed for a task is reused as the starting point for a model on a second task. Based on this idea, a new dataset can be first collected and used as a training data for the population model. Then we can obtain and use the model that is more tailored to the corresponding VR context. Researchers or practitioners who are interested in developing a learning model but do not have enough samples to train the model could use our model as a pre-trained model and apply their relatively smaller samples to make the model better represent their own samples. Additionally, our model was built with data collected from a passive virtual environment, similar to the method used in previous studies. However, since VR gameplay takes place in an active virtual environment (e.g., controller-based teleportation and walking-in-place locomotion), we believe that it is necessary to experiment with model application in other types of virtual environments. In terms of the utilization of the model, we think that such a research direction should be considered.

Compared to previous cybersickness prediction models, MAC can learn the relative relevance of each modality to cybersickness and the sensory conflict between temporal data through the attention mechanism applied in the individual convolutional and the BiLSTM subnetworks. This learning process can be viewed as a form of iterative reweighting (Lindsay 2020). MAC also reweights the relationship with sensor data, reducing training loss, and learns the association between cybersickness and sensory conflict more accurately.

6 Conclusion

In this work, we aimed to propose a way to predict cybersickness through the utilization of multimodal data and deep learning. Our study demonstrated the effectiveness of MAC and highlighted the importance of the attention mechanism and the eye movement. We proposed that the model can be improved to be more robust and applicable to other VR systems or domains. We hope that our methodologies of the model and our study results provide useful insights to other researchers, developers, and practitioners who are interested in solving the problem of cybersickness and in providing better user experience in VR.

Author contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by DJ and SP. The first draft of the manuscript was written by DJ and KH, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This research was supported by the National Research Foundation (2021M3A9E4080780) and Institute for Information & Communication Technology Planning & Evaluation (IITP-2020-0-01373, IITP-2023-2018-0-01431).

Data availability The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Code availability The code generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare that they have no conflict of interests.

Ethics approval The study was reviewed and approved by the internal institutional review board at the authors' institution (No. 02105-HS-001).

Consent to participate Informed consent was obtained from all individual participants included in the study.

Consent for publication The participant has consented to the submission of the case report to the journal.

References

Alsheikh M.A, Selim A, Niyato D, Doyle L, Lin S, Tan H.-P (2016) Deep activity recognition models with triaxial accelerometers. In: Workshops at the AAAI conference on artificial intelligence

Anwar MS, Wang J, Khan W, Ullah A, Ahmad S, Fei Z (2020) Subjective qoe of 360-degree virtual reality videos and machine learning predictions. IEEE Access 8:148084–148099

Bahdanau D, Cho K, Bengio Y (2014) Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473

Bala P, Dionísio D, Nisi V, Nunes N (2018) Visually induced motion sickness in 360°videos: comparing and combining visual optimization techniques. In: 2018 IEEE International symposium on mixed and augmented reality adjunct (ISMAR-Adjunct), pp. 244–249 . IEEE

Balasubramanian S, Soundararajan R (2019) Prediction of discomfort due to egomotion in immersive videos for virtual reality. In: 2019 IEEE International symposium on mixed and augmented reality (ISMAR), pp. 169–177. IEEE

Bles W, Bos JE, De Graaf B, Groen E, Wertheim AH (1998) Motion sickness: only one provocative conflict? Brain Res Bull 47(5):481–487

Bos JE, Bles W, Groen EL (2008) A theory on visually induced motion sickness. Displays 29(2):47–57

Bosser G, Caillet G, Gauchard G, Marçon F, Perrin P (2006) Relation between motion sickness susceptibility and vasovagal syncope susceptibility. Brain Res Bull 68(4):217–226

Cai K, Yang R, Chen H, Li L, Zhou J, Ou S, Liu F (2017) A framework combining window width-level adjustment and gaussian filter-based multi-resolution for automatic whole heart segmentation. Neurocomputing 220:138–150

Chang E, Kim HT, Yoo B (2021) Predicting cybersickness based on user's gaze behaviors in hmd-based virtual reality. J Comput Des Eng 8(2):728–739

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297

Davis S, Nesbitt K, Nalivaiko E (2014) A systematic review of cybersickness. In: Proceedings of the 2014 conference on interactive entertainment, pp. 1–9

Dennison MS, Wisti AZ, D'Zmura M (2016) Use of physiological signals to predict cybersickness. Displays 44:42–52

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929

Draper MH, Viirre ES, Furness TA, Gawron VJ (2001) Effects of image scale and system time delay on simulator sickness within head-coupled virtual environments. Hum Factors 43(1):129–146

Ebenholtz SM (1992) Motion sickness and oculomotor systems in virtual environments. Presence Teleoperators Virtual Environ 1(3):302–305

Ebenholtz SM, Cohen MM, Linder BJ (1994) The possible role of nystagmus in motion sickness: a hypothesis. Aviat Space Environ Med 65(11):1032–1035

Gavgani AM, Nesbitt KV, Blackmore KL, Nalivaiko E (2017) Profiling subjective symptoms and autonomic changes associated with cybersickness. Auton Neurosci 203:41–50

Golding JF (1998) Motion sickness susceptibility questionnaire revised and its relationship to other forms of sickness. Brain Res Bull 47(5):507–516

Greff K, Srivastava RK, Koutník J, Steunebrink BR, Schmidhuber J (2016) Lstm: a search space odyssey. IEEE Trans Neural Netw Learn Syst 28(10):2222–2232

Groth C, Tauscher J.-P, Heesen N, Grogorick S, Castillo S, Magnor M (2021) Mitigation of cybersickness in immersive 360 videos. In: 2021 IEEE conference on virtual reality and 3d user interfaces abstracts and workshops (VRW), pp. 169–177. IEEE

Hettinger LJ, Berbaum KS, Kennedy RS, Dunlap WP, Nolan MD (1990) Vection and simulator sickness. Mil Psychol. 2(3):171–181

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778

Hill KJ, Howarth PA (2000) Habituation to the side effects of immersion in a virtual environment. Displays 21(1):25–30

Howarth PA, Hodder SG (2008) Characteristics of habituation to motion in a virtual environment. Displays 29(2):117–123

Huang Z, Xu W, Yu K (2015) Bidirectional lstm-crf models for sequence tagging. arXiv preprint arXiv:1508.01991

Islam R, Ang S, Quarles J (2021) Cybersense: A closed-loop framework to detect cybersickness severity and adaptively apply reduction techniques. In: 2021 IEEE Conference on virtual reality and 3d user interfaces abstracts and workshops (VRW), pp. 148–155. IEEE

Islam R, Desai K, Quarles J (2021) Cybersickness prediction from integrated hmd's sensors: A multimodal deep fusion approach using eye-tracking and head-tracking data. In: 2021 IEEE International symposium on mixed and augmented reality (ISMAR), pp. 31–40. IEEE

Islam R, Lee Y, Jaloli M, Muhammad I, Zhu D, Rad P, Huang Y, Quarles J (2020) Automatic detection and prediction of cybersickness severity using deep neural networks from user's physiological signals. In: 2020 IEEE international symposium on mixed and augmented reality (ISMAR), pp. 400–411 . IEEE

Jahangiri A, Rakha HA (2015) Applying machine learning techniques to transportation mode recognition using mobile phone sensor data. IEEE Trans Intell Transp Syst 16(5):2406–2417

Jeong J-H, Shim K-H, Kim D-J, Lee S-W (2020) Brain-controlled robotic arm system based on multi-directional cnn-bilstm network using eeg signals. IEEE Trans Neural Syst Rehabil Eng 28(5):1226–1238

Jeong H, Kim H.G, Ro Y.M (2017) Visual comfort assessment of stereoscopic images using deep visual and disparity features based on human attention. In: 2017 IEEE international conference on image processing (ICIP), pp. 715–719. IEEE

Jeong D, Yoo S, Yun J (2019) Cybersickness analysis with eeg using deep learning algorithms. In: 2019 IEEE conference on virtual reality and 3D user interfaces (VR), pp. 827–835 . IEEE

Jin W, Fan J, Gromala D, Pasquier P(2018) Automatic prediction of cybersickness for virtual reality games. In: 2018 IEEE Games, entertainment, media conference (GEM), pp. 1–9. IEEE

Jokerst M, Gatto M, Fazio R, Gianaros P.J, Stern R.M, Koch K.L (1999) Effects of gender of subjects and experimenter on susceptibility to motion sickness. Aviation, space, and environmental medicine

Jung S, Li R, McKee R, Whitton MC, Lindeman RW (2021) Floor-vibration vr: mitigating cybersickness using whole-body tactile stimuli in highly realistic vehicle driving experiences. IEEE Trans Vis Comput Graphics 27(05):2669–2680

Kennedy RS, Lane NE, Berbaum KS, Lilienthal MG (1993) Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. Int J Aviat Psychol 3(3):203–220

Keshavarz B, Hecht H (2011) Validating an efficient method to quantify motion sickness. Hum Factors 53(4):415–426

Keshavarz B, Riecke BE, Hettinger LJ, Campos JL (2015) Vection and visually induced motion sickness: how are they related? Front Psychol 6:472

Kim HG, Lim H-T, Lee S, Ro YM (2018) Vrsa net: Vr sickness assessment considering exceptional motion for 360 vr video. IEEE Trans Image Process 28(4):1646–1660

Kim J, Oh H, Kim W, Choi S, Son W, Lee S (2020) A deep motion sickness predictor induced by visual stimuli in virtual reality. IEEE Trans Neural Netw Learn Syst 33:554

Kim J, Luu W, Palmisano S (2020) Multisensory integration and the experience of scene instability, presence and cybersickness in virtual environments. Comput Hum Behav 113:106484

Kim H.G, Baddar W.J, Lim H.-t, Jeong H, Ro Y.M (2017) Measurement of exceptional motion in vr video contents for vr sickness assessment using deep convolutional autoencoder. In: Proceedings of the 23rd ACM symposium on virtual reality software and technology, pp. 1–7

Kim J, Kim W, Oh H, Lee S, Lee S (2019) A deep cybersickness predictor based on brain signal analysis for virtual reality contents. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 10580–10589

Kundu R.K, Islam R, Calyam P, Hoque K.A (2022) Truvr: Trustworthy cybersickness detection using explainable machine learning. arXiv preprint arXiv:2209.05257

Lee S, Kim GJ, Choi S (2009) Real-time depth-of-field rendering using anisotropically filtered mipmap interpolation. IEEE Trans Visual Comput Graphics 15(3):453–464

Lee TM, Yoon J-C, Lee I-K (2019) Motion sickness prediction in stereoscopic videos using 3d convolutional neural networks. IEEE Trans Visual Comput Graphics 25(5):1919–1927

Lee S, Kim S, Kim H.G, Kim M.S, Yun S, Jeong B, Ro Y.M (2019) Physiological fusion net: Quantifying individual vr sickness with content stimulus and physiological response. In: 2019 IEEE International conference on image processing (ICIP), pp. 440–444. IEEE

Lindsay GW (2020) Attention in psychology, neuroscience, and machine learning. Front Comput Neurosci 14:29

Litleskare S, Calogiuri G (2019) Camera stabilization in 360 videos and its impact on cyber sickness, environmental perceptions, and psychophysiological responses to a simulated nature walk: a single-blinded randomized trial. Front sychol 10:2436

Lopes P, Tian N, Boulic R (2020) Eye thought you were sick! exploring eye behaviors for cybersickness detection in vr. In: Motion, Interaction and Games, pp. 1–10

MacArthur C, Grinberg A, Harley D, Hancock M (2021) You-re making me sick: a systematic review of how virtual reality research considers gender & cybersickness. In: Proceedings of the 2021 CHI conference on human factors in computing systems, pp. 1–15

Magaki T, Vallance M (2020) Seeking accessible physiological metrics to detect cybersickness in vr. Int J Virtual Augmented Real 4(1):1–18

Martin N, Mathieu N, Pallamin N, Ragot M, Diverrez J.-M (2020) Virtual reality sickness detection: an approach based on physiological signals and machine learning. In: 2020 IEEE international symposium on mixed and augmented reality (ISMAR), pp. 387–399. IEEE

McHugh N, Jung S, Hoermann S, Lindeman R.W (2019) Investigating a physical dial as a measurement tool for cybersickness in virtual reality. In: 25th ACM symposium on virtual reality software and technology, pp. 1–5

Munafo J, Diedrick M, Stoffregen TA (2017) The virtual reality head-mounted display oculus rift induces motion sickness and is sexist in its effects. Exp Brain Res 235(3):889–901

Nalivaiko E, Rudd JA, So RH (2014) Motion sickness, nausea and thermoregulation: The toxic hypothesis. Temperature 1(3):164–171

Oh S, Kim D-K (2021) Machine-deep-ensemble learning model for classifying cybersickness caused by virtual reality immersion. Cyberpsychol Behav Soc Netw 24(11):729–736

Oman CM (1982) A heuristic mathematical model for the dynamics of sensory conflict and motion sickness. Acta Otolaryngol 94(sup392):4–44

Padmanaban N, Ruban T, Sitzmann V, Norcia AM, Wetzstein G (2018) Towards a machine-learning approach for sickness prediction in 360 stereoscopic videos. IEEE Trans Visual Comput Graphics 24(4):1594–1603

Palmisano S, Allison RS, Kim J (2020) Cybersickness in head-mounted displays is caused by differences in the user's virtual and physical head pose. Front Virtual Real 1:587698

Palmisano S, Allison RS, Teixeira J, Kim J (2022) Differences in virtual and physical head orientation predict sickness during active head-mounted display-based virtual reality. Virtual Real. https://doi.org/10.1007/s10055-022-00732-5

Pan SJ, Yang Q (2009) A survey on transfer learning. IEEE Trans Knowl Data Eng 22(10):1345–1359

Qu C, Che X, Ma S, Zhu S (2022) Bio-physiological-signals-based vr cybersickness detection. CCF Trans Pervasive Comput Interact 4:268

Reason JT (1978) Motion sickness adaptation: a neural mismatch model. J R Soc Med 71(11):819–829

Reason JT, Brand JJ (1975) Motion sickness. Academic press

Rebenitsch L, Owen C (2016) Review on cybersickness in applications and visual displays. Virtual Real 20(2):101–125

Riccio GE, Stoffregen TA (1991) An ecological theory of motion sickness and postural instability. Ecol Psychol 3(3):195–240

Schmidhuber J (2015) Deep learning in neural networks: an overview. Neural Netw 61:85–117

Shahid Anwar M, Wang J, Ahmad S, Ullah A, Khan W, Fei Z (2020) Evaluating the factors affecting qoe of 360-degree videos and cybersickness levels predictions in virtual reality. Electronics 9(9):1530

Sharples S, Cobb S, Moody A, Wilson JR (2008) Virtual reality induced symptoms and effects (vrise): comparison of head mounted display (hmd), desktop and projection display systems. Displays 29(2):58–69

Stanney KM, Hale KS, Nahmens I, Kennedy RS (2003) What to expect from immersive virtual environment exposure: influences of gender, body mass index, and past experience. Hum Factors 45(3):504–520

Treisman M (1977) Motion sickness: an evolutionary hypothesis. Science 197(4302):493–495

Uddin MZ, Hassan MM (2018) Activity recognition for cognitive assistance using body sensors data and deep convolutional neural network. IEEE Sens J 19(19):8413–8419

Um T.T, Babakeshizadeh V, Kulić D (2017) Exercise motion classification from large-scale wearable sensor data using convolutional neural networks. In: 2017 IEEE/RSJ international conference on intelligent robots and systems (IROS), pp. 2385–2390. IEEE

Wang Y, Chardonnet JR, Merienne F (2019) Vr sickness prediction for navigation in immersive virtual environments using a deep long short term memory model. In: 2019 IEEE conference on virtual reality and 3d user interfaces (VR), pp. 1874–1881. IEEE

Weech S, Kenny S, Barnett-Cowan M (2019) Presence and cybersickness in virtual reality are negatively related: a review. Front Psychol 10:158

You Q, Jin H, Wang Z, Fang C, Luo J (2016) Image captioning with semantic attention. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4651–4659