Automatic Prediction of Cybersickness for Virtual Reality Games

Weina Jin, Jianyu Fan, Diane Gromala, Philippe Pasquier
School of Interactive Arts and Technology, Simon Fraser University
Surrey, Canada
{weinaj, jianyuf, gromala, pasquier}@sfu.ca

Abstract—**Cybersickness, which is also called Virtual Reality (VR) sickness, poses a significant challenge to the VR user experience. Previous work demonstrated the viability of predicting cybersickness for VR 360° videos. Is it possible to automatically predict the level of cybersickness for interactive VR games? In this paper, we present a machine learning approach to automatically predict the level of cybersickness for VR games. First, we proposed a novel ranking–rating (RR) score to measure the ground-truth annotations for cybersickness. We then verified the RR scores by comparing them with the Simulator Sickness Questionnaire (SSQ) scores. Next, we extracted features from heterogeneous data sources including the VR visual input, the head movement, and the individual characteristics. Finally, we built three machine learning models and evaluated their performances: the Convolutional Neural Network (CNN) trained from scratch, the Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) trained from scratch, and the Support Vector Regression (SVR). The results indicated that the best performance of predicting cybersickness was obtained by the LSTM-RNN, providing a viable solution for automatically cybersickness prediction for interactive VR games.**

Keywords— Machine Learning; Virtual Reality; Cybersickness.

I. INTRODUCTION

With the release of consumer Virtual Reality (VR) products, VR has gained in popularity and been widely used in many fields, such as games, film, medical training, psychological therapy, education, museum, and sports. VR is defined as a computer-generated interactive virtual world that the user is effectively immersed in and has dynamic control of the viewpoint [1]. However, the VR experience is often accompanied by VR sickness, which poses a great challenge and safety issue. VR sickness, also called cybersickness, is the motion-sickness-like symptoms that occur during exposure to the virtual environment [2]. The symptoms of cybersickness include nausea, retching, vomiting, increased salivation, cold sweating, drowsiness, pallor, dizziness, etc. Since the reported incidence of cybersickness is 61–80% [3], cybersickness is a major barrier to the wider use of VR. In particular, it limits the effective use of VR for training, rehabilitation, and therapeutic purposes.

The pathological cause of cybersickness is unknown. The most common theory in cybersickness research is the sensory mismatch theory. It states that the sickness is caused by the conflict of different sensory input channels, such as visual, auditory, and vestibular [2]. Because cybersickness cannot be completely eliminated at the current stage, automatically predicting the level of cybersickness can help the systematic control, for example, will enable the VR games to set individualized breakpoints based on the player's cumulative cybersickness level, or allow the players to review the cybersickness score before purchasing a new VR game. Machine learning provides techniques to build predictive models from real-world gameplay data, and previous works evaluated the viability of using machine learning to automatically predict cybersickness for 360° videos [4,5]. To the best of our knowledge, no work has utilized a machine learning approach to predict cybersickness for interactive VR games. The difficulties of predicting cybersickness in the real-world scenarios of VR gameplay include rich interactions, dynamic changes of viewpoints, and subjective experiences of cybersickness. To approach this problem, we first designed a ranking–rating (RR) score to measure the degree of cybersickness, and aimed to reduce the cognitive load and achieve consistency within and between annotators. Then we collected VR gameplay data and trained machine learning models based on each piece of gameplay data, rather than on the averaged data for each game, to allow our model to capture the interactive and individualized nature of VR gameplay. Our contributions to the VR community are the following:

- Designed a ranking–rating score to annotate the level of cybersickness and compared it with the Simulator Sickness Questionnaire (SSQ), which is the most widely used questionnaire in cybersickness studies.

- Utilized heterogeneous data sources including the VR visual input, the head movement, and the individual characteristics.

- Built machine learning models to predict the level of cybersickness based on the heterogeneous data and the ranking-rating annotations.

The paper is structured as follows: we first review the previous work on predicting cybersickness (Section II). Next, we describe our methods of dataset construction (Section III), data preprocessing (Section IV), and feature extraction (Section V). We built three machine learning models and compared their performances (Section VI). Finally, we present our conclusion and future work (Section VII).

TABLE I.   The Factors of Cybersickness

	Hardware	Software Content	Individual
Measured Factors	• Head movement	• Motion in a scene • Scene texture • Color in a scene	• Video game experience • VR game experience • Susceptibility to motion sickness
Controlled Factors	• Field of view • Resolution	• Sitting vs. standing	• Duration of VR exposure
Unused Factors	• None	• Independent visual backgrounds (elements of the visual field that remain stable relative to the user [6]) • The degree of control (to what extent do users have control over their movement within a virtual environment [7]) • Scene content • Change of Color (hue, saturation, and brightness) over a scene	• Postural instability (the ability of an individual to maintain balance and postural control, usually measured by body sway [8]) • History of headaches/migraine • Age • Gender

II. RELATED WORK

A. Factors of Cybersickness

Previous cybersickness research usually attributes the factors of sickness to three categories [2,6]: 1) **hardware**: the VR device and its configuration; 2) **software**: the VR content; and 3) **individual**: the user who interacts with the VR environment. Here, we list the most relevant factors in Table I as adapted from Rebenitsch's thorough literature review and prioritization (Table XI: Confirmed Factors, Table XII: Probable factors, and Table XIII: Possible factors [6]).

B. Traditional Approaches to Predict Cybersickness

As Table I shows, there are many cybersickness-related factors, whose configurations are quite flexible. Moreover, the interactions among factors are complex. Therefore, previous studies focused on controlled experiments in laboratory settings to determine the causality of one or more factors that account for cybersickness.

Kennedy et al. identified 16 symptoms as statistical indicators that showed significant changes from pre-exposure to post-exposure in 1,119 trial data [9]. Thus they devised an instrument called Simulator Sickness Questionnaire (SSQ) to quantify and predict the level of simulator sickness. It is worth pointing out that cybersickness, simulator sickness, and motion sickness share similar symptoms but are caused by exposure to different situations. While motion sickness usually involves actual motions, simulator sickness and/or cybersickness are subsets of motion sickness and are experienced while the users remain stationary [2]. The SSQ is the most widely used questionnaire in cybersickness studies [10]. It asks the degree of 16 symptoms, with a four-level severity scale of "0-none, 1-slight, 2-moderate, and 3-severe". The 16 symptoms are attributed to three subscales of nausea, oculomotor, and disorientation. The SSQ score is calculated as 3.74 times the sum of the three subscales. According to [9], after testing nine simulators from 3,691 samples, the mean ± SD of SSQ is 9.8 ± 15.0, the median is 3.7, and the SSQ score ranges from 0.0 to 108.6.

Since cybersickness is a subjective experience, to predict individual differences, Golding devised the Motion Sickness Susceptibility Questionnaire short version (MSSQ) [11]. It is an 18-item self-report questionnaire that asks participants about their motion sickness experience (scored with numbers 0–3) when riding different transportations in childhood and adulthood. The MSSQ shows good reliability and validity for predicting individual susceptibility to motion sickness. To calculate the MSSQ score, for each subscale of childhood and adulthood, subscore = (total sickness score) × 9 / (9 – the number of types not experienced). The MSSQ score is the summation of the childhood and adulthood subscores. It ranges from 0 to 54. According to [11], the mean ± SD of MSSQ score was 12.9 ± 9.9 with a positively skewed distribution from a normative sample of 257 university students.

Rebenitsch developed several single-factor linear models that incorporate multiple factors to predict cybersickness [6]. Based on a set of controlled experiments involving 24 participants, the model on individual factors explains 37% of the adjusted variance of cybersickness. It included factors of MSSQ, headache, and video gameplay. Based on previous studies with reported configurations, the researcher built linear models on hardware and software factors, and it explained 55% of the adjusted variance. The model included factors of duration, tracking, controller type, seated or not, the realism of a virtual environment, the field of view, and movement in VR.

C. Machine Learning Approach of Predicting Cybersickness for 360° Videos

While the traditional controlled experimental approaches achieved high internal validity [12] in providing the causal correlation between specific factors and cybersickness, they sacrificed external validity of generalization to the real world, and may not capture the complex interactions among factors. Machine learning provides an alternative approach to this problem. It learns directly from real-world observational data

and saves the experimental labor of controlling variables one at a time to determine their correlations. After a thorough literature search, we list the recent work related to cybersickness and machine learning.

Padmanaban et al. built a machine learning model to predict the sickness level when watching a 360° stereoscopic video [4]. They collected a set of 109 one-minute videos annotated with SSQ from 96 participants. They then trained a bagged decision tree as the machine learning model on hand-crafted features (quantifying speed, direction, and depth) from video content. Their model generally outperformed a naive estimate but was ultimately limited by the size of the dataset. During the data collection, the participant's head motion was constrained using a headrest to ensure that all users saw the same scene and the participant was not allowed to look around. This limits their model to be generalized to a fully interactive VR environment in which the viewpoint changes frequently.

Kim et al. adopted an unsupervised learning approach to detect irregular motions in 360° videos [5]. The deep convolutional autoencoder extracts features from five consecutive video frames in the encoder part and reconstructs the video sequences in the decoder part. Since the reconstruction errors were high for irregular motions, this network was used as an objective measure of the exceptional motions, and such measures had a high correlation with a subjective measure of SSQ. However, because the training and test videos used in this study were consecutive driving videos of cities and roads, this may make the model not generalize well to complex scenarios in VR gameplay that involve dynamic changes of viewpoints and scenes. Moreover, their model only took account of the single motion velocity factor to measure cybersickness.

Despite the previous progress in predicting cybersickness, none has used machine learning models to predict the level of cybersickness for interactive VR games. Built upon the previous work, we present a machine learning approach to predict cybersickness for VR games.

III. Dataset Construction

A. Study Design

1) Factors Included in the Study

We designed a study and recruited participants to collect data for building the predictive model. In our experiment, we tried to include as many factors of cybersickness as possible based on previous literature reviews. Among them, some were included as *measured factors*, while others were included as *controlled factors* that were fixed throughout the study. The rows of "*Measured Factors*" and "*Controlled Factors*" in Table I summarize the factors included in this study. We describe the methods of collecting these factors below.

a) Hardware

We included *head movement* as a hardware factor of cybersickness. Head movement was recorded as the changes of position and rotation of the head-mounted display (HMD), with a sample rate of 500 Hz. To represent the head movement, the captured data were later processed to extract the following

factors: linear speed and acceleration in 3-D space, and the first and second difference of quaternion for spatial rotation.

b) Software Content

We recorded the *eye screen video* that was displayed in the HMD. The visual output from the HMD was projected to the desktop and was recorded with a both-eye view at 60 frames per second (fps). The captured videos were then processed to extract factors of motion, texture, and color.

c) Individual

We designed a *demographic questionnaire* to record individual factors. The video game experience and VR game experience were collected as binary data (0 for rarely, 1 for often). The susceptibility to motion sickness was measured by MSSQ.

TABLE II. Factors Controlled in this Study

Hardware	Software Content	Individual
VR device = HTC Vive Field of view = 110° Resolution = 1080 ×1200 Refresh rate = 90 Hz	Position = Sitting	Duration = 3 minutes

Some factors, although related to cybersickness, were controlled throughout the study. Although these controlled factors were not included in our present models, we reported their configurations in Table II, so that the dataset can be extended by collecting more data with different configurations in future work. Due to the nature of our pilot study, other possible factors were not collected in our study, or not included in our present models, as shown in the row of "Unused Factors" in Table I. We plan to take them into consideration in future work.

2) VR Games Used in the Study

We chose five "off-the-shelf" VR games for the study based on the distribution of each factor. The games are: 1) *The Night Cafe: A VR Tribute to Vincent Van Gogh* [1]; 2) *NoLimits 2 Roller Coaster Simulation* [2]; 3) *Endless Labyrinth* [3]; 4) *InCell VR* [4]; and 5) *Audioshield* [5]. The game configurations are listed in Table III.

3) Ranking-Rating Measure of Cybersickness

For quantifying subjective assessment, ratings are the most often applied instrument. However, because contextual situations differ, the meaning of ratings may change over time or across individuals [13]. Compared with ratings, ranking-based approaches show consistency within participants and over time [14, 15]. However, a ranking system itself cannot easily assess the distance between values for one participant, and it is difficult to compare the rankings between participants.

---

[1] https://store.steampowered.com/app/482390
[2] https://store.steampowered.com/app/301320
[3] https://store.steampowered.com/app/495830
[4] https://store.steampowered.com/app/396030
[5] https://store.steampowered.com/app/412740

TABLE III.   THE FIVE VR GAMES PLAYED IN THIS STUDY

Name	*Audioshield*	*The Night Cafe*	*InCell VR*	*Endless Labyrinth*	*NoLimits 2 Roller Coaster Simulation*
Game level played in the study	Difficult: Normal; Music: Chopin Nocturne Op.9, No.2; Skin: Private amphitheater	The first 3 minutes of the game	Level 1	The first 3 minutes of the game	"Dive Park" theme in the default library
Control of movement	No full-body movement in space	Teleport	The movement along the axis is forced by the game. The player could move the head to control the orientation.	Use a controller to glide in the labyrinth.	The movement is forced by the game
Use of controller	Two controllers	One controller to teleport	None	One controller	None
Game Screenshot					
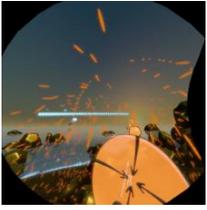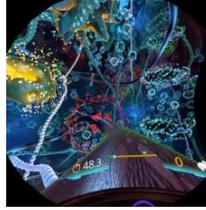
Table IV lists the pros and cons of the two measures [13,15,16].

TABLE IV.   PROS AND CONS OF RATING AND RANKING.

	Rating	**Ranking**
Pros	• Widely used • Easy to design	• A higher inter-annotator reliability • The relative ranking task is simple for participants.
Cons	• Imposes heavy cognitive load on the participants	• It is relatively difficult to design a ranking-based measure • The analysis method of ranking is less commonly known

The two measures may not be mutually exclusive [17]. Based on the advantages of both measures, we designed an instrument, named "ranking-rating" (RR) measure, that combines the ranking and rating at the same time to assess cybersickness across games and individuals. To compare the RR measure with the previous measures, after each gameplay, we also asked the participants to answer the SSQ, which is a rating-based measure that widely used in cybersickness studies.

Here, we describe the procedure of using the RR measure for cybersickness. The following rating and ranking processes are based on an 11-point scale (0–10, 0 for no sickness and 10 for the highest level of sickness). For each participant, the RR measure consists of two steps.

First step: After playing the first game, participants were asked to rate the score of cybersickness based on their understanding of the scale. Starting from the second game, after each gameplay, participants were asked to evaluate their cybersickness based on the relative comparisons (ranking) among the games they had played so far. By ranking, the participants faced less cognitive load compared to solely rating the next game. It also avoids inconsistencies potentially caused by a changing understanding of the scale. These scores were used as references in the second step.

Second step: After playing all the games, participants were asked to adjust all the scores they marked in the first step. They could modify the order of the ranking, and adjust the distance between each other. The use of fractions or decimals was encouraged, and if decimals were used, all participants chose to report their scores with a resolution of 0.5. The scores obtained in the second step were the final RR scores of cybersickness. The idea of designing this second step is that after playing all the games, participants now have a more general understanding of the various levels of cybersickness. This procedure also helps to minimize the bias introduced by the initial rating of the first game.

B. Study Procedure

The data collection was conducted in an experimental setting. The study was approved by Simon Fraser University's Research Ethics Board. The inclusion criterion is people above 18 years old. The exclusion criteria are people with migraines or other severe diseases which may affect the study. Participants were recruited using convenience sampling. Written informed consent was obtained from participants before enrolment. A total of 25 participants were enrolled in this study. One participant's data were discarded because of withdrawal and data incompletion. The data of 24 participants (Female = 10, Male = 14; Age 28.0 ± 7.5 years) were included in the final data processing.

During the study, the participants were asked to play five "off-the-shelf" VR games as stated above. The games were purchased from Steam gaming platform[6]. Each gameplay lasted

---

[6] http://store.steampowered.com/

for three minutes. The gameplay sequence was randomly assigned by drawing lots. Participants remained seated on a swivel chair to play the VR games. They could move their body in the virtual environment by moving the swivel chair in the 2.5 × 2.5 meters physical space.

We used HTC Vive as the VR gaming device. The Vive headset has a refresh rate of 90 Hz and a 110° field of view. The display resolution for each eye screen is 1080 × 1200. The games were run on an Alienware Desktop (Intel Core i7-5820K CPU @ 3.30 GHz, 12 Logical Processors; 16 GB RAM; Dual NVIDIA GTX 980) with Windows 10 Enterprise operating system.

After each gameplay, the participants were asked to evaluate their levels of cybersickness using RR and SSQ measure, as stated in Section III.A.3. Then, the participants had a washout period between each gameplay for at least five minutes until they did not feel any sickness from the previous gameplay. After the whole gameplay session, the participants had a semi-structured interview to talk about their feelings of sickness during each gameplay. Then, they answered the demographic questionnaire before completing the study. The study session last for 1–1.5 hours in total. The participants were thanked with $10 CAD cash for their time and effort participating in this study.

IV. DATA PREPROCESSING AND AUGMENTATION

A. Sickness Score Calculation

The SSQ score was calculated according to the method described in Section II.B. The RR score was collected directly in the study without any effort to calculate, as stated in Section III.A.3. It has the range of [0.0, 10.0].

The Pearson's correlation coefficient r between the SSQ and RR scores is 0.838, which indicates that the two have a strong positive linear correlation. Therefore, we used the RR score as the ground-truth label for the construction of the following predictive models. Figure 1 visualizes the distribution of the RR and SSQ scores and their correlation.
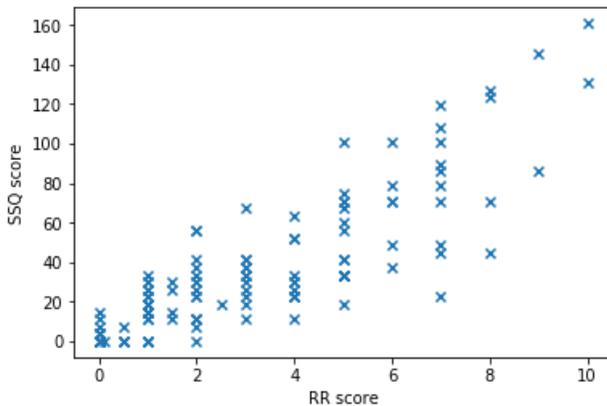

Figure 1. The distribution and correlation between RR and SSQ score.

It is worth noting that instead of using the averaged RR score of each game as the ground-truth label, we used the raw individual RR scores as the ground truth, and regarded each gameplay as one data point. This is because each gameplay information (eye video recordings, head movement, individual characteristics, cybersickness experience) is different from the other. We ended up having 120 raw data points (5 games per participant × 24 participants).

B. Data Preprocessing

For the eye screen video data, first, we truncated the original recordings to obtain the first 125 seconds of the eye screen videos. Second, we rescaled the video recordings. For each video recording in the corpus, the size is 37 × 20 (width × height). Third, considering the low-frequency nature of the visual stimuli and the head movement, and the computational recourses, we downsampled our video recordings to 2 fps.

For the head movement data, we truncated the first 125 seconds of the raw data, which was in the same period aligned with the video data. We then downsampled the original 500 Hz data to 2 Hz to align with the video data. The time steps of video and head movement data were precisely aligned. The raw data consisted of two vectors: the position in 3-D space (x, y, z), and the quaternion spatial rotation (qw, qx, qy, qz). A quaternion uses four numbers to simplify the way to encode an axis–angle rotation to a position vector. The vector (qx, qy, qz) in a quaternion represents the position vector, and the scalar qw encodes the rotation angle.

C. Data Augmentation

Because each of the original video recordings is 125 seconds long, each converted video recording has 250 frames in total. Then, we adopted a windowing method to perform data augmentation to enlarge the training set artificially. We experimented with a couple of settings and chose the window size of 60 frames and a step size of 10 frames. For one video recording, we kept selecting 60 consecutive frames as one augmented recording and moved one step ahead, until we reached the end of a video recording. After the data augmentation, we ended up having 2,400 video recordings. One augmented video recording is 30 seconds long and contains 60 frames. The annotations of each augmented video recording are the same as the annotations of the original video recording.

The head movement data were augmented in the same way as the video recording. We ended up having a set of 2,400 augmented data points, and each has the shape of 14 × 60. The details of head movement features are described in Section V.B.

V. FEATURE EXTRACTION

A. Individual Features

For the video game experience and VR game experience, we collected binary data with the demographic questionnaire. These data were treated as categorical features for the machine learning model. The details of these two features are described in Section III.A.1.c. In addition, we calculated the MSSQ score (numerical feature) according to the method described in Section II.B.

B. Head Movement Features

We computed hand-crafted head movement features based on the original position vectors (*x, y, z*) and quaternion vectors (*qw, qx, qy, qz*). For the position vector, we computed the speed feature (3 dimensions) as the first difference of two consecutive position vectors, and the acceleration feature (3 dimensions) as the first difference of two consecutive speed vectors. For the quaternion vector, we computed the first difference (4 dimensions) and the second difference (4 dimensions) of two consecutive quaternion vectors. The size of the head movement features is 14×60, where 14 is the total number of dimensions, and 60 is the total number of time steps (2 Hz for 30 seconds).

C. Video Features (Eye Screen Video)

We extracted low-level video features including color, motion, and texture. The ViVid [18] software package was used for the video feature extraction.

1) Color: Hue, Saturation, and Brightness

Regarding color features, we selected hue, saturation, and brightness, which are related to human perception of colors. For each color feature, the original range of the value is from 0 to 255. Here, we divided the range into 32 bins. Instead of computing these features frame by frame, we computed these color features over the entire 30-second video recording (all frames together). We ended up having 96 dimensions of color features for one video recording.

Hue is defined as "the degree to which a stimulus can be described as similar to or different from stimuli that are described as: red, orange, yellow, green, blue, violet" [19]. Saturation represents the level of purity of the color. A higher saturation value indicates less mixing of colors. By mixing more colors, the picture will turn gray. Therefore, the purity will decrease, and the saturation will decrease. The brightness feature represents the level of a source appears to be radiating or reflecting light, which is the perception elicited by the luminance of a visual target [20]. Similar to saturation, we separated the brightness value into 32 bins.

2) Motion: Motion Intensity

For motion features, we investigated the motion intensity, which is the level of motion between frames within the video. While the color features were computed over the entire video recording and presented based on its distribution, the motion intensity feature was extracted based on two consecutive frames. Since there are 60 frames in each augmented video recording, we have 60 time-steps of motion features in total.

3) Texture: Contrast, Smoothness, Entropy

We chose to compute contrast, smoothness and entropy features to provide the texture information of the video recordings. These are also frame-level features. Contrast is determined by the difference of color between the objects within an image. Smoothness represents the homogeneity of the gray level distribution of a frame. It is approximately inversely correlated with contrast. Entropy represents the level of randomness. Specifically, it is a measure of the amount of information, which must be coded by a compression

algorithm. If the image has higher entropy, it contains more information. Each feature has 60 time-steps.

D. Two Feature Sets

We used two sets of features for different machine learning models separately. Feature Set 1 (Table V) contains the time sequence features (frame by frame) and was fed into the Convolutional Neural Network (CNN) and Long-Short Term Memory Recurrent Neural Networks (LSTM-RNN) models. We ended up having an 18×60 feature matrix, where 18 is the total number of dimensions, and 60 is the number of time steps. Each feature was normalized by subtracting its average and dividing by its standard deviation.

TABLE V.   FEATURE SET 1, PREPARED FOR CNN & LSTM-RNN

Features	Feature Types	Dimensions	Time Steps
Head Movement	Temporal	14	60
Motion Intensity	Temporal	1	60
Contrast	Temporal	1	60
Smoothness	Temporal	1	60
Entropy	Temporal	1	60

TABLE VI.   FEATURE SET 2, PREPARED FOR SVR

Features	Feature Types	Dimensions
Video Game Experience	Categorical	1
VR Experience	Categorical	1
MSSQ	Numerical	1
Head Movement	6 Statistics	84
Hue	Distribution	32
Saturation	Distribution	32
Brightness	Distribution	32
Motion Intensity	6 Statistics	6
Texture	6 Statistics	18

The Feature Set 2 (Table VI) was prepared for Support Vector Regression (SVR) model. These features are either categorical features (video game experience and VR game experience), color features (distributions of hue, saturation, and brightness), or the statistics of time sequence features. For the statistics of each time sequence feature, we computed 6 features: mean, standard deviation, skewness, kurtosis, maximum and minimum. We ended up having a 207-

dimensional feature vector as listed in Table VI. Each feature was normalized by subtracting its average and dividing by its standard deviation.

VI. MACHINE LEARNING MODELS AND PERFORMANCES

In this section, we describe the construction of three machine learning models: the CNN, LSTM-RNN, and SVR.

A. CNN

We built and trained a CNN model from scratch. We applied grid search to find the best hyperparameters such as the number of kernels in each layer, kernel size, learning rate, and decay. Our CNN is composed of two convolutional layers and one fully connected layer. The first convolutional layer filters the $18 \times 60 \times 1$ input features with 16 kernels of size $3 \times 3 \times 1$ and a stride of 1. The second convolutional layer, connected to the first one, uses 16 kernels of size $3 \times 3 \times 16$. We used maxpooling (2×2) for the outputs of both convolutional layers. There is a dropout layer between the second convolutional and the fully-connected layer, with a dropout rate of 0.15. The fully connected layer, connected to the second convolutional layer, is composed of 256 neurons. The output layer is composed of 1 neuron. We used the ReLU non-linearity activation function for all convolutional layers and the fully connected layer. For the output layer, we use the linear activation to predict the level of cybersickness. All the weights were initialized based on a Xavier uniform. The CNN was trained using the RMSProp optimizer with a batch size of 32 examples, learning rate of 1×10^{-3} and decay of 1×10^{-6}. We trained 60 epochs before testing the model.

B. LSTM-RNN

We built and trained an LSTM-RNN from scratch. To find the best hyperparameters such as the number of neurons in each layer, learning rate, and decay, we used a grid search method. Our LSTM-RNN is composed of two stacked LSTM units. The input size is 18×60, where 18 is the number of the feature extracted from head movement and eye screen video data, and 60 is the number of time steps (See Table V). There are 64 neurons in each LSTM unit. The output layer is composed of 1 neuron. We used the tanh non-linearity activation function for LSTM units, and linear activation function for the output layer to predict the level of cybersickness. Similar to CNN, all the weights were also initialized based on a Xavier uniform. We trained the LSTM-RNN using the RMSProp optimizer with a batch size of 32 examples, learning rate of 1×10^{-2} and decay of 1×10^{-6}. We trained 60 epochs before testing the model. We implemented the CNN and LSTM-RNN with Keras 2.0.

C. SVR

The SVR model maps the input data into a higher dimensional feature space using nonlinear mapping and builds a linear model in this feature space to make predictions. We fed the SVR with Feature Set 2 (See Table VI), selected the Radial Basis Function (RBF) kernel and used a grid search method to find the best hyperparameters of C and $gamma$. We implemented the SVR in scikit-learn 0.19.0.

D. Performances

To train and evaluate the above three models, we applied repeated random sub-sampling validation. The augmented dataset composing a total of 2,400 30-seconds data was shuffled 10 times. Each time, 90% of the data were randomly selected for training the model, and the remaining 10% was used for testing. We use R^2 and MSE to evaluate the performance of the prediction. Table VII presents the performance results of CNN, LSTM-RNN, and SVR. Among the three models, the LSTM-RNN model achieved the highest performance both on R^2 (0.868) and MSE (0.009).

TABLE VII. MODEL PREDICTION PREFORMANCES

	R^2 [a]	MSE [b]
CNN	0.462	0.036
LSTM-RNN	**0.868**	**0.009**
SVR	0.793	0.014

[a.] R^2: Coefficient of determination

[b.] MSE: Mean Square Error

E. Discussion

Our results showed that LSTM-RNN model outperformed CNN and SVR models in predicting cybersickness. One possible explanation is that the LSTM-RNN could remember things and find patterns across time to make predictions, which is suitable for the problem of predicting cybersickness based on the time-series events of VR gameplay. Another possible explanation is that the LSTM-RNN may be capable of modeling the interaction and "mismatch" between the head motion and visual motion stimuli. For example, it may capture a rapid visual change with no head movement, which corresponds to the sensory mismatch theory that explains the cause of cybersickness. In addition, the LSTM-RNN model could make predictions without the individual features; it only used the eye screen videos and head movement data as input. This would save the players from manually inputting their individual features for predicting cybersickness.

The SVR model also achieved good, but not equivalent performance to the LSTM-RNN, probably due to the lack of capacity of representing the interactions among features. Regarding the CNN, it is reasonable that the CNN did not achieve satisfying results because it is more suitable to capture spatial rather than temporal information.

Our system of predicting cybersickness can be utilized in the following usage scenarios: *1) Quantify cybersickness in a retrospective manner*: it can be applied to automatically quantify the cumulative cybersickness over a period of exposure to VR games. It will allow the VR systems to set individual-based breakpoints to control cybersickness below a certain threshold and thus improve VR gaming experience. *2) Predict cybersickness in a prospective manner*: our system enables the VR game developers and platforms to compute the cybersickness score automatically, thereby encourages the initiation of a cybersickness rating system. Such system allows the players to easily check the level of cybersickness of new

games. However, since our model requires eye screen video recordings and head movement data as input, such information cannot be acquired before the gameplay. One way to solve this problem is to invite a number of trial players to play the new game first and record their eye screen videos and head movement data. The data collected from the trial players will be fed to our model to get a number of predicted RR scores. Then, an overall RR score for the new game can be computed by averaging the predicted RR scores from each trial player.

VII. CONCLUSIONS & FUTURE WORKS

In this paper, we present a machine learning approach to automatically predict the level of cybersickness for interactive VR games. We reviewed the literature and selected the possible factors for cybersickness, and designed an experiment to construct the dataset. Different from previous study approaches, which mainly considered the static viewpoint of visual stimuli in VR as input, our study included the dynamic visual stimuli, head movement, and individual features. We designed the RR measure and aimed to achieve an easier and consistent assessment for cybersickness. We built three machine learning models and compared their performance of predicting cybersickness. The results indicated that the LSTM-RNN is a more viable model for the problem of predicting cybersickness.

The present dataset can be expanded in future work by collecting more data varying in the configurations of the measured and controlled factors as shown in Table I. As for the features extracted from eye screen video recordings, at the current stage, we only extracted low-level features such as color, motion, and texture. Since cybersickness is a multiplex reaction mingled with physiological and psychological processes, in the future work, we also intend to include higher-level information, such as semantic or emotional features to improve our model's performance.

ACKNOWLEDGMENT

REFERENCES

[1] F. P. Brooks, "What's real about virtual reality?," *American Industrial Hygiene Association Journal, ieee computer graphics and applications, IEEE computer graphics and applications.*, vol. 19, no. 6, pp. 16–27, Nov. 1999.

[2] S. Davis, K. Nesbitt, and E. Nalivaiko, *A Systematic Review of Cybersickness*. New York, NY, USA: ACM, 2014.

[3] B. Lawson, "Motion Sickness Symptomatology and Origins," in *Handbook of Virtual Environments*, CRC Press, 2014, pp. 531–600.

[4] N. Padmanaban, T. Ruban, V. Sitzmann, A. M. Norcia, and G. Wetzstein, "Towards a Machine-learning Approach for Sickness Prediction in 360° Stereoscopic Videos," *IEEE Transactions on Visualization and Computer Graphics*, vol. PP, no. 99, pp. 1–1, 2018.

[5] H. G. Kim, W. J. Baddar, H. Lim, H. Jeong, and Y. M. Ro, "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*, New York, NY, USA, 2017, pp. 36:1–36:7.

[6] L. R. Rebenitsch, *Cybersickness Prioritization and Modeling*. Michigan State University, 2015.

[7] K. M. Stanney and P. Hash, "Locus of User-Initiated Control in Virtual Environments: Influences on Cybersickness," *Presence*, vol. 7, no. 5, pp. 447–459, Oct. 1998.

[8] S. V. G. Cobb, "Measurement of postural stability before and after immersion in a virtual environment," *Applied Ergonomics*, vol. 30, no. 1, pp. 47–57, Feb. 1999.

[9] R. S. Kennedy, N. E. Lane, K. S. Berbaum, and M. G. Lilienthal, "Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness," *The International Journal of Aviation Psychology*, vol. 3, no. 3, pp. 203–220, Jul. 1993.

[10] B. Lawson, "Motion Sickness Scaling," in *Handbook of Virtual Environments*, CRC Press, 2014, pp. 601–626.

[11] J. F. Golding, "Predicting individual differences in motion sickness susceptibility by questionnaire," *Personality and Individual Differences*, vol. 41, no. 2, pp. 237–248, Jul. 2006.

[12] R. W. Proctor and E. J. Capaldi, "Internal and External Validity," in *Why Science Matters*, Wiley-Blackwell, 2008, pp. 180–194.

[13] G. N. Yannakakis and H. P. Martínez, "Ratings are Overrated!," *Front. ICT*, vol. 2, 2015.

[14] O. L. Bock and C. M. Oman, "Dynamics of subjective discomfort in motion sickness as measured with a magnitude estimation method.," *Aviation, Space, and Environmental Medicine*, vol. 53, no. 8, pp. 773–777, 1982.

[15] J. Fan, M. Thorogood, and P. Pasquier, "Emo-soundscapes: A dataset for soundscape emotion recognition," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2017, pp. 196–201.

[16] A. Metallinou and S. Narayanan, "Annotation and processing of continuous emotional attributes: Challenges and opportunities," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013, pp. 1–8.

[17] S. Ovadia, "Ratings and rankings: reconsidering the structure of values and their measurement," *International Journal of Social Research Methodology*, vol. 7, no. 5, pp. 403–414, Sep. 2004.

[18] J. Fan, P. Pasquier, L. M. Fadel, and J. Bizzocchi, "ViVid: A Video Feature Visualization Engine," in *Design, User Experience, and Usability: Understanding Users and Contexts*, vol. 10290, A. Marcus and W. Wang, Eds. Cham: Springer International Publishing, 2017, pp. 42–53.

[19] J. J. V. Wijk and E. R. V. Selow, "Cluster and calendar based visualization of time series data," in *1999 IEEE Symposium on Information Visualization, 1999. (Info Vis '99) Proceedings*, 1999, pp. 4–9, 140.

[20] A. Vadivel, S. Sural, and A. k. Majumdar, "Robust histogram generation from the HSV space based on visual colour perception," *International Journal of Signal and Imaging Systems Engineering*, vol. 1, no. 3–4, pp. 245–254, Jan. 2008.