This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

A Deep Cybersickness Predictor Based on Brain Signal Analysis for Virtual Reality Contents

Jinwoo Kim[1], Woojae Kim[1], Heeseok Oh[2], Seongmin Lee[1], and Sanghoon Lee[1]

[1]Yonsei University
[2]Electronics & Telecommunications Research Institute

Abstract

What if we could interpret the cognitive state of a user while experiencing a virtual reality (VR) and estimate the cognitive state from a visual stimulus? In this paper, we address the above question by developing an electroencephalography (EEG) driven VR cybersickness prediction model. The EEG data has been widely utilized to learn the cognitive representation of brain activity. In the first stage, to fully exploit the advantages of the EEG data, it is transformed into the multi-channel spectrogram which enables to account for the correlation of spectral and temporal coefficient. Then, a convolutional neural network (CNN) is applied to encode the cognitive representation of the EEG spectrogram. In the second stage, we train a cybersickness prediction model on the VR video sequence by designing a Recurrent Neural Network (RNN). Here, the encoded cognitive representation is transferred to the model to train the visual and cognitive features for cybersickness prediction. Through the proposed framework, it is possible to predict the cybersickness level that reflects brain activity automatically. We use 8-channels EEG data to record brain activity while more than 200 subjects experience 44 different VR contents. After rigorous training, we demonstrate that the proposed framework reliably estimates cognitive states without the EEG data. Furthermore, it achieves state-of-the-art performance comparing to existing VR cybersickness prediction models.

1. Introduction

Although Virtual Reality (VR) devices are effectively integrated into a variety of applications, such as movies, games and medical cares, the cybersickness that occurs while experiencing VR is considered as an obstacle to the VR industries. Unlike the stereoscopic 3D display, the VR environment is accompanied by complicated cognitive-physiological factors in the brain. For this reason, it is difficult to determine the exact cause of the cybersickness. In particular, the level of cybersickness distributes differently according to individual differences (e.g., prior experience, susceptibility, gender, age, etc.). In this respects, the biosignals are treated as one of the most objective ways to reflect individual differences (e.g., galvanic skin response (GSR), photoplethysmogram (PPG), electroencephalography (EEG), skin temperature (SKT), etc.). However, there is no solid publication to predict the cybersickness using a model while reflecting the individual difference even if researchers recognize that this is the most important factor for the prediction. Thereby, the measurement using the sensors is regarded as a reliable way for the cybersickness prediction.

Currently, there are some major works to predict this visual discomfort over 3D applications [22, 23]. The performance has been thresholded due to the failure of including the individual difference into their metrics which are formulated or modeled to find common factors in general. Recently, with the breakthrough evolution of convolutional neural network (CNN) [9, 10, 18], there have been significant applications in the image/video content analysis field [15, 16]. In this paper, we make a pioneer work to generalize individual differences of VR cybersickness by utilizing the VR video sequences only. However, the general VR video sequences are recorded according to the user's head motion in the limited virtual space, in this reason, the VR videos have similar characteristics over the spatial and temporal axes. Therefore, the CNN-based model, especially visual feature-driven manner strongly depends on the extracted features from the viewed VR video sequence without any cue of individual differences.

To overcome this limitation, we devise a novel deep learning framework to identifies the human cognitive feature space for cybersickness prediction by analyzing brain activity. Furthermore, the framework interprets individual differences by relying on VR contents rather not on brain

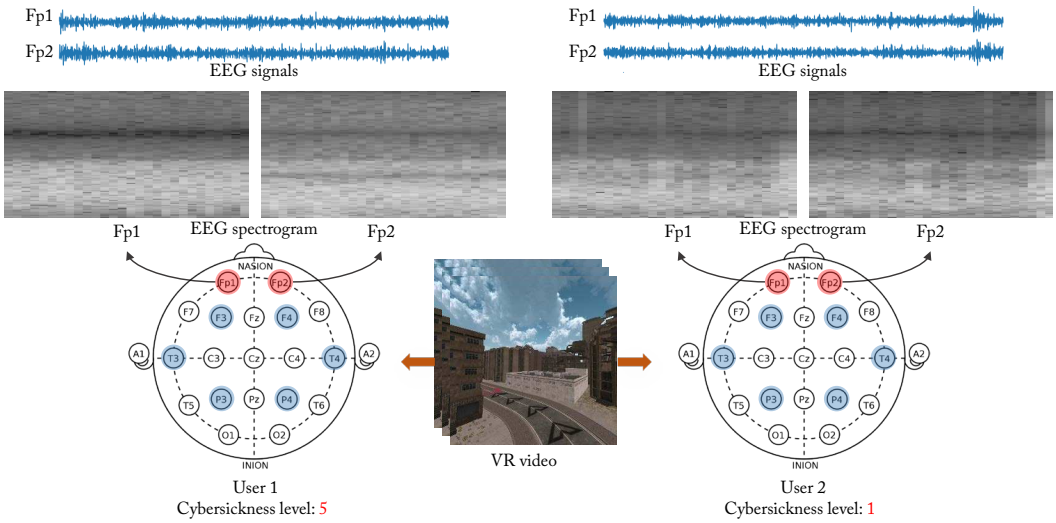

Figure 1. Example of EEG data evoked by different users: The raw EEG signals and transformed spectrograms are represented for FP1 and FP2, which record brain activity of the frontal lobe among the eight markers. Each user shows different cybersickness levels in the same content.

signals. To realize this, we start from an observation by prior arts on VR cybersickness [5–7, 19, 20, 31, 32]. Fig. 1 shows the recorded EEG data when the two users experience the same content, and Fp1 and Fp2 are the examples which designate the brain activity of the frontal lobe among the eight markers. As shown in the figure, the cybersickness score of each user gives a different level, although they experienced the same content. Here, we can find that the EEG data is much more distinct than the recorded VR video, depending on the level of cybersickness.

Based on this observation, we aim to encode EEG signals to the cognitive representation relative to VR cybersickness. Moreover, by transferring the cognitive representation onto the VR video-based deep model, we perform the cybersickness prediction without the EEG signal. Through our framework, it enables the machine to analyze and understand the pattern of the EEG signal which is one of the important goal in the brain-computer interface (BCI) research. Since the main purpose of BCI is to directly classify the specific patterns from EEG data, we believe the proposed framework has a significant impact beyond the BCI approach. To this end, the fundamental ideas of the framework are as follows: the *cognitive representation learning* by classifying the EEG signals, and the *cybersickness learning* that expresses the visual and cognitive features at an intermediate state using VR video.

Among these steps, the *cognitive representation learning* plays an important role since it captures both inter- and intra-individual differences of the cybersickness. For more detail, we first transform the EEG data into a spectrogram and it is then encoded by CNN. Note that the spectrogram includes the temporal and spectral domain. However, since the generic CNN filters deal with omnidirectional correlation over the 2D axis, it is difficult to apply the spectrogram directly into the general CNN network. Therefore, we propose a new CNN approach by geometrical processing dedicated to the spectrogram domain, i.e., temporal and spectral.

Our contributions are summarized as follows:

- We propose a novel deep learning architecture for estimating cognitive state using EEG spectrogram by discriminating the feature spaces related to VR cybersickness levels.
- We present a method for computing and combining visual and cognitive features with VR videos alone for cybersickness prediction.
- We will release a massive VR content database including the recorded EEG data, and it also contains a simulator sickness questionnaire (SSQ) measurements for various subjects.

2. Related Work

Currently, a number of theoretical papers have been published describing the mechanism of cybersickness. The sensory conflict theory is stated that cybersickness arises from conflicts between information coming from the visual-vestibular systems [28]. The subjective vertical theory is stated that cybersickness is caused by the collision between perceived and expected information from the body sensor and brain, respectively [3]. In [4], it is suggested that the two theories could be integrated to develop a mechanism of cybersickness occurrence. Based on these observations,

many cybersickness prediction models have been developed. However, since each sensor module of describing the mechanism is a black-box model that is not defined as a deterministic function, the general strategy for cybersickness prediction follows a top-down framework. In other words, the models are designed on the assumption that the cognitive conflict by motion is the main factor of cybersickness. For example, the authors in [12, 14, 24] used a feature vector extracted from an optical flow containing motion information.

Computational models using brain signals have been developed in the literature. For example, the self-organizing neural fuzzy inference network (SONFIN) is a model based on the assumption that the power spectrum of the EEG data reflects the correlation with the cybersickness [20]. In particular, the EEG data has been analyzed through sequential models utilized raw 1-D signals for seizure detection [1]. However, since the EEG data processing based on sequential models is strongly optimized in the temporal feature space, the models tend to fail to generalize spectral correlation as well as inter-channel interaction. Developed from the previous study, the authors of [2] proposed a deep learning approach that preserves spatial, spectral and temporal structures by transforming the EEG data into a sequence of topology-preserving multi-spectral images. While extracting significant features that are less sensitive to distortion and variation in each dimension, this method fails to show satisfactory performance when a small number of markers are used to record brain signals.

There are some other approaches, that learn EEG manifold for image classification by estimating cognitive state at the intermediate stage [25, 30]. The primary objective of these studies is to interpret the human mind from the image and to transfer it to the learned EEG manifold, while our approach aims to look for visual and cognitive representation simultaneously from the image sequences.

3. Proposed Algorithm

The approach described in this paper is based on the following intuitions.

- The visual information, which is a feature vector derived from the VR videos, is superior in the performance prediction of inter-content cybersickness prediction, but weak in predicting the sickness level made by the subject in the same content.

- The EEG data evoked from VR videos transmits cognitive information that conveys inter-subjective differences, i.e., individual differences, about VR cybersickness. Fig. 1 shows quantitative differences in the same VR content of EEG data according to different subjects.

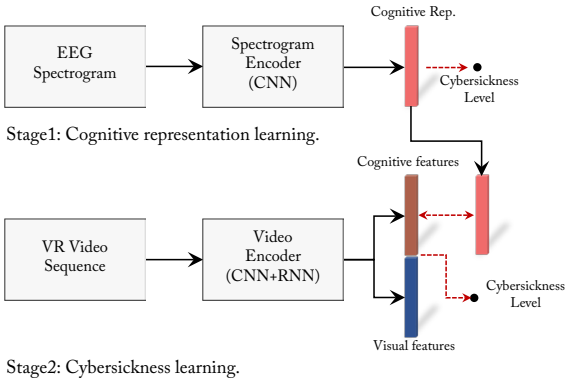

Figure 2. Overview of the proposed approach. Our work is composed of two stages. Stage1: The network trains the cognitive states related to VR cybersickness using the EEG spectrograms. Stage2: The network learns complementary visual and cognitive features for cybersickness prediction using the VR video sequence.

- We assume that if the visual-cognitive information is learned complementary through VR videos, then the model will generalize inter- and intra-subjective differences.

Through above-described intuitions, our proposed approach is designed with two-stage learning for the cybersickness prediction as shown in Fig. 2. The first stage of our work – *cognitive representation learning* – seek to generate a decision boundary that determines the cybersickness level with a low-dimensional representation within the EEG space. To learn this representation, the recorded EEG data is transformed into the spectrogram. Then, the *spectrogram encoder* is trained to extract a meaningful feature vector that describes cybersickness from EEG spectrogram. This was implemented in CNN with a dynamic filter shape for multi-channel aware spectral and temporal correlation analysis. The training process is supervised by the cybersickness level, while the fully connected layer (FCL) for cognitive representation is learned in the process.

The use of EEG data is unreasonable in the application aspect, as it requires an additional device to obtain signals and reduces the practicality of the device. Therefore, the second stage – *cybersickness learning* – aims at learning visual and cognitive features jointly through VR video only for cybersickness prediction. By training the *video encoder* combined with CNN and RNN, the visual features are extracted, and the features after FCL are mapped to cognitive representation learned in stage 1. In the end, the last feature vector, concatenated with visual and cognitive features, is classified to the cybersickness level through the FCL.

The cybersickness level	5
The number of VR content	44
The number of subjective	202
The number of data	8,888
Visualization order	Sequential
Time for subjective test	30 sec.
Time for pause	3 min.

Table 1. The parameters for the subjective experiment.

3.1. Data Acquisition

To our knowledge, there are no public databases for cybersickness prediction. Hence we introduce a new cybersickness database named ETRI-VR including a variety of visual motion and different types of reference scenes: *'urban'* scenes with high complex component; *'astrospace'* scenes consisting of relatively simple object arrangement. Note that the path of each scene is scripted in advance and the control operation is not reflected except the user's head motion. By using the constructed ETRI-VR contents, we collected human's opinions for each scene in terms of cybersickness. During the experiment, 44 VR contents were divided into 3 sessions. In each session, VR contents were continuously shown to 202 subjects. The rest time between sessions was given 3 minutes, and subjects were asked to perform the subjective evaluation at the end of each content according to the Likert-like scale: 5=Extreme sickness, 4=Strong sickness, 3=Sickness, 2=Mild sickness, and 1=Comfortable. In the end, we collected different subjective scores for $44 \times 202 = 8,888$ contents. HTC VIVE was used for the subjective experiments and the frame rate was kept above 96 fps to minimize cybersickness caused by motion to photon latency that was irrelevant to the psychophysical aspect. A summary of the experimental procedure is shown in Table. 1.

The EEG data was also collected using 8 scalp electrodes during VR content usage as shown in Fig. 1 following the international system [29]. The sampling rate and resolution of the EEG data were set to 250 Hz and 16 bits, respectively. A bandpass filter $(0.3 \sim 100Hz)$ and notch filter $(at\ 60Hz)$ were applied to minimize the effect of power line noise [20]. The length of collected EEG data varied depending on the VR contents. From each EEG data, each of the first and last 250 samples $(1.0s)$ were discarded in order to exclude any possible interference from the previously shown experience according to [30]. Then, the 3,450 samples $(14.0s)$ EEG data in the middle area were employed for the experiments. After acquiring the EEG data, the spectrogram transformation proceeded using a Fourier transform (FFT) through a sliding window, i.e., the data block was determined to be $0.5s$ with a Hann window. Thus, EEG spectrogram can be denoted by I_s, which is a multidimensional array of the form $I_s \in \mathbb{R}^{8\times64\times53}$, where each spectrogram has dimension 64×53 and 8 is the number of channels.

3.2. Stage 1: Cognitive Representation Learning

This stage is primarily intended to encode cybersickness as a low-dimensional representation of the EEG data. The details are depicted in Fig. 3. The existing approaches have focused on the temporal feature space to discriminate EEG data [30]. On the other hand, the proposed method take into account the inter-correlation of the EEG channles and the intra-correlation over spectral and temporal domains. For this reason, the EEG data is transformed into a spectrogram and stacked in the input pipeline, i.e., 8 channels stacked spectrogram. Note that each axis of the spectrogram indicates temporal and spectral domains, respectively.

There is a problem in applying the existing CNN method directly to the EEG spectrogram. The CNN operation takes the omnidirectional correlation of local pixels by square shape filter (e.g., 3×3 and 5×5). However, the coefficients of the spectrogram are only correlated in the horizontal or vertical directions. To overcome this, we encode the spectrogram of EEG data by following networks inspired by previous audio signal processing work [27]. First, *temporal* network trains temporal dependency by taking various sizes of the horizontal kernel. Second, *spectral* network learns spectral dependency through various sizes of the vertical kernel. Third, the *temporal* and *spectral* networks are concatenated to encode the temporal and spectral features jointly. The details are as follows:

- *Temporal* dependency kernels *(1-by-m)*: are capable to learn temporal cues by capturing the horizontal coefficients of the spectrogram. For example, such filters are specialized to make temporal representations related to cybersickness. As shown in the model of Fig. 3, deep convolutional operations with $1 \times m_i$ kernels are used for the *temporal* network, where $i \in \{1,\ 2,\ and\ 3\}$ represents the i^{th} convolution layer. The kernel length of each convolution layer gradually decreases by half of the previous layer. Note that, due to the convolution operation procedure in the temporal axis, the spectral resolution is preserved.

- *Spectral* dependency kernels *(n-by-1)*: are designed to learn the spectral cues by using vertical coefficients of the spectrogram. To capture spectral correlation, convolutional network with $n_j \times 1$ kernels are applied to the EEG spectrogram, where $j \in \{1,\ 2,\ and\ 3\}$ represents the j^{th} convolution layer. As same as the temporal network, the kernel length of each convolution layer gradually decreases by half of the previous layer. Note that the *spectral* network only learns spectral correlation by reducing the spectral dimension while preserving the temporal resolution.

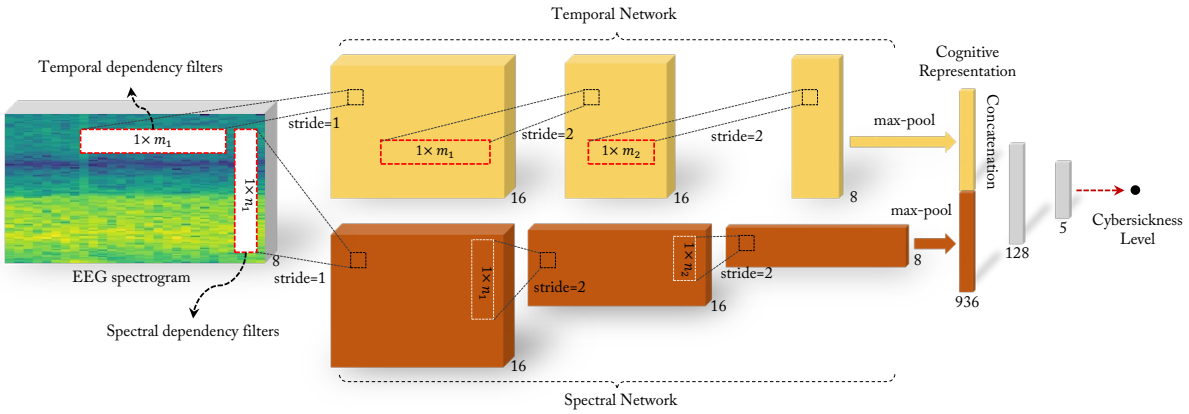

Figure 3. Overall architecture for cognitive representation learning. The network consists of two parts: the *temporal network* and the *spectral network*. Both networks extract the temporal and spectral feature vectors from the EEG spectrogram, respectively, and the feature vectors are concatenated to express the *cognitive representation*. The last feature vector is classified as a cybersickness levels after passing through the FCL.

In both the *temporal* and *spectral* networks, two strided convolutions are used for subsampling except for the firtst layer. In each convolutional layer, batch normalization [11] and a leaky rectified linear unit (LReLU) [21] are employed continuously. After the end of each network, the max-pooled temporal $\mathbf{x_t}$ and spectral $\mathbf{x_s}$ feature vectors are concatenated, and we denote this vector as a cognitive representation $\mathbf{x_{cr}}$. After then, two FCLs are used to discriminate features onto the cybersickness level. After training, the *temporal* and *spectral* networks are used as a ground truth of the cognitive features in stage 2.

3.3. Stage 2: Cybersickness Learning

Our goal is to estimate the cognitive state from a visual stimulus without the EEG data. Toward this, we propose a deep model that predicts the cybersickness while mimicking the cognitive representation encoded in stage 1. The overall architecture is illustrated in Fig. 4. The input VR video is first sampled as same as EEG samples to synchronize both data. Then the frames in each sampled VR video are fed to *ResNet18* [9] to extract spatial features.

After the spatial features are extracted from the VR video sequences, the temporal features are then considered in a sequential network. Here we use a stacked long short term memory (LSTM) network as a sequential network. Then, we take the visual feature vector $\mathbf{x_v}$ from the last step of the *stacked LSTM* network. For precise expressions, let combined *ResNet18* and *stacked LSTM* be the *video encoder*. After passing through the *video encoder*, the FCL is used to extract cognitive features $\mathbf{x_c}$. Then it is concatenated with the visual features $\mathbf{x_v}$ to produce the final feature vector. At the end of the model, the network two FCLs are utilized to classify the final feature vector onto the cybersickness levels.

The final objective function is formed by two terms; the prediction loss and the regression loss as

$$\mathcal{L} = \mathcal{L}_{pre} + \beta \cdot \mathcal{L}_{reg}, \tag{1}$$

where $\mathcal{L}_{pre}$ indicates the standard cross-entropy loss between cybersickness levels and output unit of last FCL, $\mathcal{L}_{reg}$ denotes the regression loss on the cognitive features, and β is a constant to tune the trade-off between the two terms. The regression loss is defined by the mean squared error (MSE) between the cognitive features and the cognitive representation

$$\mathcal{L}_{reg} = \|\mathbf{x_c} - \mathbf{x_{cr}}\|_2^2 . \tag{2}$$

4. Experimental Results

In this section, we first describe the implementation details. Then, we analyze the contribution of individual components in our proposed network. Performance analysis is split into two parts – Test 1: cognitive representation learning and Test 2: cybersickness learning. In test 1, we verify that the EEG spectrogram driven network truly encodes the cognitive state as low-dimension vector relative to cybersickness proposed in Section 3.2. In test 2, we demonstrate that the proposed network learns the cognitive state using only VR video for cybersickness prediction by transferring the encoded cognitive representation to the intermediate feature space.

4.1. Implementation Details

For the experiments, the ETRI-VR database mentioned in Section 3.1 is utilized. We randomly divided the database into training (80%), validation (10%) and testing (10%)

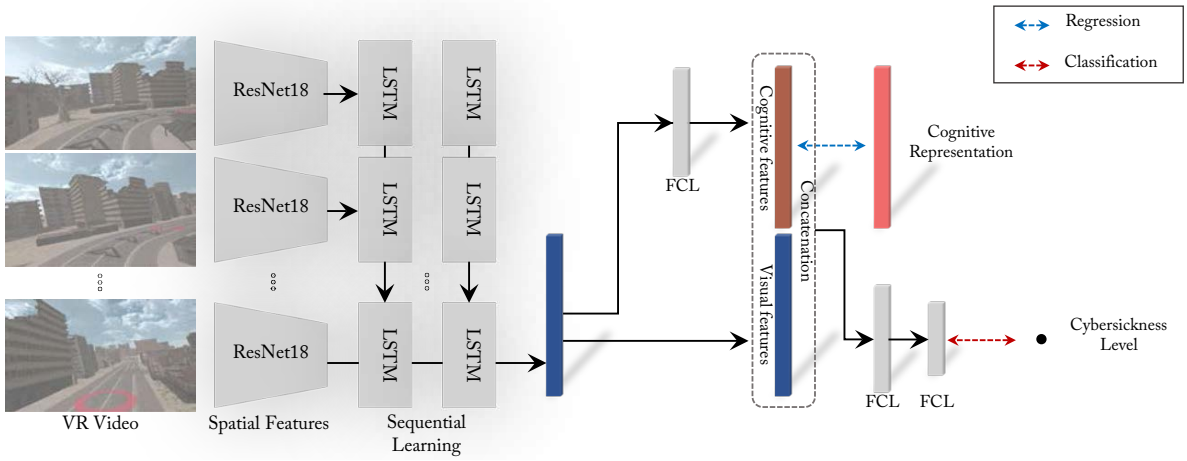

Figure 4. Overall architecture for cybersickness learning. The input VR video pass through combined CNN-RNN network to extract visual features. The visual features are given to the fully connected layer to represent cognitive features and both featuers are concatenated.

set. To ensure performance validity, Monte-Carlo cross-validation with 20 repetition was then conducted. We implemented the proposed networks on the pytorch framework [26]. The initial weight value of all the networks was applied to the method proposed by [8], which normalizes to the variance according to the input dimensionality. We used the Adam optimizer to train the networks, with the momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and regularization parameter $\epsilon = 10^{-6}$ [17]. The learning rate was initially set to 5×10^{-4}.

The overall experimental evaluations were performed with the split validation and test set. However, in the final model, the overall samples were used as a training set to encode. In stage 2, we resized the input dimension to $3 \times 224 \times 224$. The VR videos were uniformly sampled at a sampling interval r. The performance comparisons for each sampling interval will be analyzed in Section 4.3.

4.2. Test 1: Cognitive Representation Learning

In this section, we tested three individual ablation sets: the *temporal* network, the *spectral* network, and the *proposed* model which made by combining the *temporal* and *spectral* networks as shown in Fig. 3. Here, one *baseline* network was compared. The *baseline* network adopts a general CNN architecture that utilizes the square shape kernel. Accordingly, the structure of the baseline is the same as the proposed model except for the shape of the filter kernel.

In Table 2, the achieved performance by different configurations is shown according to the detail kernel shape to analyze the temporal and spectral dependencies of EEG spectrogram. The top model for each evaluation criterion is shown in boldface. The *baseline* network with $(n_1, m_1) = (3, 3)$ has the worst validation (test) accuracy with $58.32\ (56.89)\%$. In case of the *temporal* network, the validation (test) accuracy is $81.38\ (80.57)\%$ at $(n_1, m_1) = (1, 28)$ where the filter size is the most horizontal, i.e., the temporal dependency is the longest. On the other hand, in the *spectral* network, the prediction performance is superior to the narrow kernel shape that captures small frequency resolution. The best performance of validation (test) is $85.51\ (83.74)\%$ at $(n_1, m_1) = (8, 1)$. This agrees with the fact that features in a specific frequency rather than broad band are highly correlated with cybersickness [20]. Besides, the *spectral* network is more robust to the kernel shape than the *temporal* network. Overall, the prediction performance of the *proposed* network with $(n_1, m_1) = (1, 28)$-$(8, 1)$ is the most outperformed than other configurations.

The BCI study has examined how the brain region of the EEG signals affects VR cybersickness. To further contribute to this, we tested the model with an independent channel of the spectrogram as an input, and compared prediction accuracy. More specifically, the proposed networks were learned using independent spectrogram dimensions. Table 3 shows the performance comparison according to the eight brain regions (Fp1, Fp2, F3, F4, T3, T4, P3, and P4). The prediction accuracy for each region is almost similar, but the P3 and P4 regions are slightly better. Therefore, it can be concluded that our work agrees with previous BCI research since the VR cybersickness levels are highly correlated with responses in the occipital midline brain area than in other brain areas [20].

4.3. Test 2: Cybersickness Learning

To verify whether the proposed model truly learns the cognitive state using the VR video alone, we tested two scenarios: *visual predictor*; *visual – cognitive predictor*. The *visual predictor* takes only visual features as an input of

Models	**Kernel Shape** (n_1, m_1)	**Max-pool**	**VA**: (mean±std.)	**TA**: (mean±std.)
baseline network	$(3,3)$	$(2,2)$	$58.32 \pm 1.71\%$	$56.89 \pm 1.86\%$
	$(5,5)$	$(2,2)$	$60.77 \pm 2.41\%$	$61.09 \pm 2.33\%$
	$(7,7)$	$(2,2)$	$61.31 \pm 1.89\%$	$61.83 \pm 1.41\%$
temporal network	$(1,7)$	$(1,12)$	$63.49 \pm 2.13\%$	$65.17 \pm 2.88\%$
	$(1,14)$	$(1,8)$	$67.16 \pm 2.72\%$	$69.16 \pm 2.38\%$
	$(1,28)$	-	$81.38 \pm 3.01\%$	$80.57 \pm 2.49\%$
spectral network	$(8,1)$	$(13,1)$	$85.51 \pm 1.33\%$	$83.74 \pm 2.81\%$
	$(16,1)$	$(9,1)$	$81.57 \pm 1.76\%$	$80.27 \pm 1.49\%$
	$(32,1)$	-	$73.51 \pm 1.33\%$	$77.74 \pm 2.81\%$
proposed network	$(1,7)$-$(8,1)$	$(1,12)$-$(13,1)$	$84.33 \pm 1.19\%$	$83.72 \pm 1.87\%$
	$(1,28)$-$(32,1)$	-	$86.01 \pm 1.12\%$	$85.91 \pm 1.22\%$
	$(1,28)$-$(8,1)$	$(0,0)$-$(13,1)$	$\mathbf{87.46 \pm 2.37\%}$	$\mathbf{87.13 \pm 1.51\%}$

Table 2. Maximum validation accuracy ("**VA**") and test accuracy ("**TA**") at **VA** for different networks according to kernel shape. The top model for prediction accracy is in bold.

Regions	Fp1	Fp2	F3	F4	T3	T4	P3	P4
VA (%)	82.21	82.15	82.52	81.19	83.55	83.13	**85.93**	**86.43**
TA (%)	81.78	83.23	81.37	83.42	82.10	82.71	**84.66**	**86.16**

Table 3. Maximum validation accuracy ("**VA**") and test accuracy ("**TA**") at **VA** according to brain marker region. The best two marker regions are in bold.

the last FCL in stage 2. The *visual – cognitive predictor* is the full version as depicted in Fig. 4. Here, the performance are benchmarked using with two existing handcrafted feature based methods [14, 24] and one deep learning based method [13]. Since the benchmarked methods are not designed to reflect individual differences, so the predicted cybersickness scores were regressed onto the mean opinion score (MOS) of each content. However, the proposed model is based on classification over the individual cybersickness levels. Therefore, we modified the last regressor of the benchmark methods as a classifier to match with discrete cybersickness levels and then trained benchmarked methods using the ETRI-VR database.

In Table 4, the validation accuracy and test accuracy of the benchmarked methods are compared to the proposed model. As shown n the table, the *visual – cognitive predictor* outperformed the other methods. In particular, compared to the *visual – cognitive predictor* and the *visual predictor*, it is noteworthy that the integrated learning of the visual and cognitive features is much superior to using only visual features. Interestingly, the validation (test) performance of the proposed model shows 26.55 $(26.20)\%$ higher than the *visual predictor*. From this observation, it is evident that taking the cognitive features helps the model generalize individual differences to achieve higher accuracy. In particular, the cognitive features are expressed only from the visual features, which is an impressive result, considering that any objective indicator, such as EEG data, was not used.

In addition, the prediction accuracy of the *visual – cognitive predictor* is superior to EEG-driven cybersickness prediction comparing to Tables 2 and 4. This result implies that both the cognitive and visual features are meaningful information relative to cybersickness and lead to more powerful performance when they are integrated. We used confusion matrics to illustrate the discordance between the *visual – cognitive predictor's* predictions and user's cybersickness levels. When the VR cybersickness level is severe, such as 4 and 5, it shows accurate prediction performance, while levels 1 and 2 are more confused. We expect that this is due to the low-level VR cybersickness boundaries where the user feels ambifuous to make a decision.

To further discuss whether our model truly learns individual differences, we tested the network by transferring cognitive features to stage 1. For more detail, we used the cognitive features learned in stage 2 as input in the last FCL learned in stage 1. In Table 5, the results of the prediction accuracy using the cognitive representation and cognitive features are reported. The prediction model transferring cognitive features was comparable to that of the model directly learned from the EEG data. In fact, after finishing optimization, the MSE between the cognitive representation and features is 0.52, which implies that the distance between the two spaces is very close. In our experiment,

Type	Models	**VA**: (mean±std.)	**TA**: (mean±std.)
benchmark	Padmanaban *et al.* [24]	$51.94 \pm 2.33\%$	$50.48 \pm 3.81\%$
	Kim *et al.* [14]	$55.72 \pm 4.31\%$	$58.37 \pm 4.17\%$
	Kim *et al.* [13]	$63.12 \pm 1.72\%$	$65.83 \pm 1.88\%$
proposed	*visual predictor*	$68.93 \pm 1.65\%$	$69.03 \pm 1.24\%$
	visual – cognitive Predictor	$\mathbf{90.48 \pm 1.99\%}$	$\mathbf{89.16 \pm 1.87\%}$

Table 4. Maximum validation accuracy ("**VA**") and test accuracy ("**TA**") at **VA** for different networks. The top model among the accuracies is in bold.

	cognitive representation	cognitive features
VA (%)	87.46%	85.91%
TA (%)	87.13%	86.35%

Table 5. Maximum validation accuracy ("**VA**") and test accuracy ("**TA**") at **VA** for cognitive representation and cognitive features.

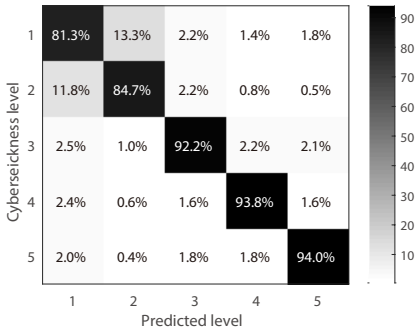

Figure 5. Confusion matrix for the predicted VR cybersickness levels versus the subjective evaluation levels.

the regression loss is converged to around zero, hence it can be concluded that the model estimates the cognitive state from the visual features without any prior physiological cue. Therefore, the inter- and intra-individual differences are usefully addressed to predict VR cybersickness. These observations are an extension of cognitive neuroscience studies that estimate the EEG response from visual stimuli [30].

As mentioned in Section 4.1, the input VR video frames are uniformly sampled at the sampling interval r and pass through the network. Therefore, at high sampling interval, the model is expected to detect rapidly changing motion patterns than a low sampling interval. The results according to the sampling interval are shown in Fig. 6. Actually, the prediction accuracy according to the sampling interval is almost similar, and slightly better performance at a low sampling interval, i.e., $r = 4$. It is expected that there will be a difference in the prediction accuracy in low-level cybersickness contents depending on the sampling interval, rather than in high-level cybersickness contents.

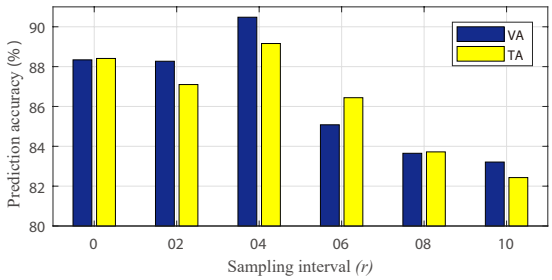

Figure 6. Cybersickness prediction accuracy according to input VR video sampling interval.

5. Conclusion

In this paper, we have proposed a novel framework for VR cybersickness prediction where the deep learning model learns individual differences using only VR video. Our framework consists of two stages. In the first stage, the EEG spectrogram driven classification was performed to cybersickness levels by representing cognitive states with the low-dimensional vector. To extract a meaningful cognitive representation of cybersickness, our novel architecture holistically considered the inter-correlation of the EEG channels and intra-correlation over the spectral and temporal informations in each spectrogram. Through the rigorous tests, we demonstrated that designing an architecture to reflect temporal and spectral dependency is essential cues to describe cybersickness. In the second stage, by transferring cognitive representation to the intermediate feature space, an RNN-based network aimed at extracting visual cognitive features for cybersickness prediction. Above all, the proposed model achieves a state-of-the-art performance on the ETRI-VR database and reliably estimated individual differences through the VR video without the EEG data.

Acknowledgment. This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No.2017-0-00289, Development of a method for regulating human-factor parameters for reducing VR-induced sickness)

References

[1] U Rajendra Acharya, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, and Hojjat Adeli. Deep convolutional neural network for the automated detection and diagnosis of seizure using EEG signals. *Computers in biology and medicine*, 100:270–278, 2018. 3

[2] Pouya Bashivan, Irina Rish, Mohammed Yeasin, and Noel Codella. Learning representations from EEG with deep recurrent-convolutional neural networks. *arXiv preprint arXiv:1511.06448*, 2015. 3

[3] Willem Bles, Jelte E Bos, Bernd De Graaf, Eric Groen, and Alexander H Wertheim. Motion sickness: only one provocative conflict? *Brain research bulletin*, 47(5):481–487, 1998. 2

[4] Jelte E Bos, Willem Bles, and Eric L Groen. A theory on visually induced motion sickness. *Displays*, 29(2):47–57, 2008. 2

[5] WE Chelen, MATTHEW Kabrisky, and SK Rogers. Spectral analysis of the electroencephalographic response to motion sickness. *Aviation, space, and environmental medicine*, 64(1):24–29, 1993. 2

[6] Yu-Chieh Chen, Jeng-Ren Duann, Shang-Wen Chuang, Chun-Ling Lin, Li-Wei Ko, Tzyy-Ping Jung, and Chin-Teng Lin. Spatial and temporal EEG dynamics of motion sickness. *NeuroImage*, 49(3):2862–2870, 2010. 2

[7] Yu-Chieh Chen, Jeng-Ren Duann, Chun-Ling Lin, Shang-Wen Chuang, Tzyy-Ping Jung, and Chin-Teng Lin. Motion-sickness related brain areas and EEG power activates. In *International Conference on Foundations of Augmented Cognition*, pages 348–354. Springer, 2009. 2

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 6

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 5

[10] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 1

[11] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015. 5

[12] Hak Gu Kim, Wissam J Baddar, Heoun-taek Lim, Hyunwook Jeong, and Yong Man 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*, page 36. ACM, 2017. 3

[13] Hak Gu Kim, Heoun-Taek Lim, Sangmin Lee, and Yong Man Ro. VRSA Net: VR Sickness Assessment Considering Exceptional Motion for 360 VR Video. *IEEE Transactions on Image Processing*, 28(4):1646–1660, 2019. 7, 8

[14] Jaekvung Kim, Woojae Kim, Sewoong Ahn, Jinwoo Kim, and Sanghoon Lee. Virtual Reality Sickness Predictor: Analysis of visual-vestibular conflict and VR contents. In *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, 2018. 3, 7, 8

[15] Jongyoo Kim and Sanghoon Lee. Deep learning of human visual sensitivity in image quality assessment framework. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1676–1684, 2017. 1

[16] Woojae Kim, Jongyoo Kim, Sewoong Ahn, Jinwoo Kim, and Sanghoon Lee. Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 219–234, 2018. 1

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 1

[19] Chin-Teng Lin, Shang-Wen Chuang, Yu-Chieh Chen, Li-Wei Ko, Sheng-Fu Liang, and Tzyy-Ping Jung. EEG effects of motion sickness induced in a dynamic virtual reality environment. In *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3872–3875. IEEE, 2007. 2

[20] Chin-Teng Lin, Shu-Fang Tsai, and Li-Wei Ko. EEG-based learning system for online motion sickness level estimation in a dynamic vehicle environment. *IEEE transactions on neural networks and learning systems*, 24(10):1689–1700, 2013. 2, 3, 4, 6

[21] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3, 2013. 5

[22] Heeseok Oh, Sewoong Ahn, Sanghoon Lee, and Alan Conrad Bovik. Deep visual discomfort predictor for stereoscopic 3d images. *IEEE Transactions on Image Processing*, 27(11):5420–5432, 2018. 1

[23] Heeseok Oh, Sanghoon Lee, and Alan Conrad Bovik. Stereoscopic 3d visual discomfort prediction: A dynamic accommodation and vergence interaction model. *IEEE Transactions on Image Processing*, 25(2):615–629, 2016. 1

[24] Nitish Padmanaban, Timon Ruban, Vincent Sitzmann, Anthony M Norcia, and Gordon Wetzstein. Towards a machine-learning approach for sickness prediction in 360 stereoscopic videos. *IEEE transactions on visualization and computer graphics*, 24(4):1594–1603, 2018. 3, 7, 8

[25] Simone Palazzo, Concetto Spampinato, Isaak Kavasidis, Daniela Giordano, and Mubarak Shah. Generative adversarial networks conditioned by brain signals. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3410–3418, 2017. 3

[26] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

[27] Jordi Pons, Thomas Lidy, and Xavier Serra. Experimenting with musically motivated convolutional neural networks. In *2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1–6. IEEE, 2016. 4

[28] James T Reason and Joseph John Brand. *Motion sickness.* Academic press, 1975. 2

[29] F Scharbrough, GE Chatrian, R Lesser, H Luders, M Nuwer, and T Picton. Guidelines for standard electrode position nomenclature. *Am. EEG Soc*, 1990. 4

[30] Concetto Spampinato, Simone Palazzo, Isaak Kavasidis, Daniela Giordano, Nasim Souly, and Mubarak Shah. Deep learning human mind for automated visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6809–6817, 2017. 3, 4, 8

[31] CS Wei, SW Chuang, WR Wang, LW Ko, TP Jung, and CT Lin. Development of a motion sickness evaluation system based on EEG spectrum analysis. In *Proceedings of the 2011 IEEE International Symposium on Circuits and Systems (ISCAS 2011)*, 2011. 2

[32] JIAN-P WU. EEG changes in man during motion sickness induced by parallel swing. *Space Medicine and Medical Engineering*, 5(3):200–205, 1992. 2