PHYSIOLOGICAL FUSION NET: QUANTIFYING INDIVIDUAL VR SICKNESS WITH CONTENT STIMULUS AND PHYSIOLOGICAL RESPONSE

Sangmin Lee[1], *Seongyeop Kim*[1], *Hak Gu Kim*[1], *Min Seob Kim*[2], *Seokho Yun*[2], *Bumseok Jeong*[2], *Yong Man Ro*[1]

[1]Image and Video Systems Lab, School of Electrical Engineering, KAIST, South Korea
[2]Graduate School of Medical Science and Engineering, KAIST, South Korea

ABSTRACT

Quantifying Virtual Reality (VR) sickness is demanded in industry to address viewing safety issue. In this paper, we develop a new method to quantify VR sickness. We propose a novel physiological fusion deep network which estimates individual VR sickness with content stimulus and physiological response. In the proposed framework, content stimulus guider and physiological response guider are devised to effectively represent feature related with VR sickness. Deep stimulus feature from the content stimulus guiders reflects the content sickness tendency while deep physiology feature from the physiological response guider reflects the individual sickness characteristics. By combining those features, VR sickness predictor quantifies individual Simulation Sickness Questionnaires (SSQ) scores. To evaluate the performance of the proposed method, we built a new dataset that consists of 360-degree videos with physiological signals and SSQ scores. Experimental results show that the proposed method achieved meaningful correlation with human subjective scores.

Index Terms— Virtual reality, individual sickness, content stimulus, physiological response

1. INTRODUCTION

Virtual Reality (VR) can provide immersive experience. With the rapid development of VR equipment and 360-degree video acquisition device, VR contents have increasingly attracted attention in industry and research fields. However, as the VR environment expands, concerns over the safety of viewing VR contents are rising. Several studies reported that symptoms containing headache, dizziness, and focusing difficulty are triggered when viewing VR contents [1,2]. Generally, 80% to 95% of people feel VR sickness [3]. Therefore, in order to handle the VR sickness, it is needed to quantify the VR sickness caused by viewing VR contents and to provide a safety guide of VR content creation and viewing.

In recent years, VR sickness quantification methods have been introduced [4–6]. Kim et al. [4] proposed a sickness quantification method with deep learning-based generative model. This generative model was trained by VR contents with normal motions. At testing phase, this generative model could not reconstruct VR videos with exceptional motion that causes sickness. Therefore, the degree of the VR sickness could be quantified based on the difference between the original video and the generated video. In [5], a deep network that consists of generator and VR sickness predictor was reported for sickness quantification. In this model, the difference between the original video and the generated video is regressed to the Simulation Sickness Questionnaires (SSQ) [7] score. The aforementioned VR sickness quantification methods estimated mean value of SSQ score, not individual VR sickness. Another study [6] quantified VR sickness caused by visual-vestibular conflict. In this work, SVM [8] was used on motion feature from visual-vestibular interaction and content feature from VR content. This method did not consider the deviation from subjects even on the same stimulus. Also, used stimulus contents are controlled graphical video.

In this paper, we propose a novel physiological fusion deep network that predicts individual VR sickness considering real-world content stimulus and subject. There were clinical studies that validated the correlation between subjective sickness and physiological responses [9–13]. Based on the physiological relationship with sickness, the proposed deep network consists of content stimulus guider, physiological response guider, and VR sickness predictor. The content stimulus guider extracts content characteristics related to the sickness level of VR videos. The content stimulus guider is composed of a visual expectation generator and a stimulus context extractor. The purpose of the visual expectation generator is to extract features that deviate from the normal VR videos. The stimulus context extractor outputs a deep stimulus feature by receiving VR video and features from the visual expectation generator. The physiological response guider extracts individual sickness features by receiving physiological signals (EEG, EKG, and GSR). Each physiological signal is encoded in a frequency domain and a time domain, and then fused. The domain fused features for EEG, EKG, and GSR are integrated once again to create a deep physiology feature.

This work was supported by IITP grant (No. 2017-0-00780). The corresponding author of this project is Yong Man Ro.

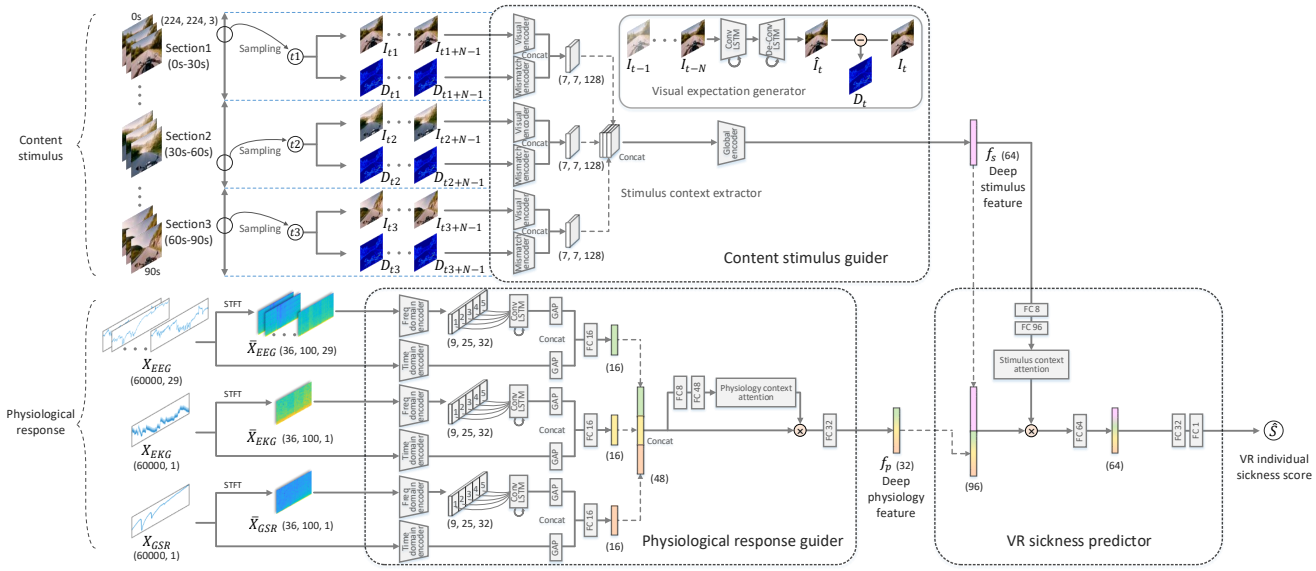

Fig. 1. Proposed physiological fusion network for predicting individual sickness.

This physiology feature reflects individual sickness characteristics. Finally, the VR sickness predictor estimates the SSQ score by combining the deep stimulus feature that includes sickness tendency of VR video, with the deep physiology feature that contains individual sickness characteristics.

To validate the proposed method, we collected real-world 360-degree video data with corresponding SSQ scores and physiological signals (EEG, EKG, and GSR). The collected stimulus videos have various motion patterns with two types of frame rate (10Hz, 60Hz). The subjective experiment was conducted under the supervision of neuropsychiatry specialists. The performance of the proposed model was evaluated with the human SSQ scores.

The major contributions of this work are as follows.

- We propose a novel deep learning framework that predicts individual VR sickness with content stimulus and physiological response. To the best of our knowledge, this is the first work that quantifies individual VR sickness considering individual subject with stimulus.
- For evaluation of the proposed model, we built newly collected real-world 360-degree video dataset with corresponding physiological signals (EEG, EKG, and GSR) and SSQ scores. This VR sickness assessment dataset will be publicly available.

2. PROPOSED METHOD

Fig. 1 shows the proposed physiological fusion network for predicting individual VR sickness. The overall network is divided into three parts which are content stimulus guider, physiological response guider, and VR sickness predictor. Given a VR content, the content stimulus guider extracts the deep stimulus feature that reflects the content characteristics. The physiological response guider utilizes physiological signals being collected during watching the VR content to extract deep physiology feature. With the deep stimulus feature and the deep physiology feature, the VR sickness predictor predicts subjective VR sickness score. When predicting individual VR sickness, physiology feature is considered as well as content feature in the proposed method.

2.1. Content stimulus guider

VR sickness could arise if sensory information that an individual perceives does not correspond with the normal experience [14]. Based on this observation, we design the content stimulus guider, which consists of visual expectation generator and stimulus context extractor. Actual viewport of VR contents is used as the input of the content stimulus guider.

The visual expectation generator takes previous N frames $I_{t-N}, ..., I_{t-1}$ to generate the next frame $\hat{I}_t \in \mathbb{R}^{224\times224\times3}$ ($N = 11$). The generator consists of ConvLSTM [15] and DeConvLSTM which replaces convolution with deconvolution. The generator is pre-trained with videos including only normal motion with high frame rate (60Hz). Therefore, the generated frame has a large difference from the original frame for abnormal (sickness-inducing) VR content that could contain exceptional motion. To generate a desirable next frame, a pixel-wise generation loss is used for training the generator. Let G denote the generator function. The generation loss can be written as

$$\mathcal{L}_{gen} = \frac{1}{K} \sum_{t \in batch} \|G(I_{t-N}, ..., I_{t-1}) - I_t\|_2^2, \qquad (1)$$

where K is a mini batch size at training phase.

Based on the visual expectation generator, the stimulus context extractor outputs deep stimulus feature which is re-

Table 1. Network structures of proposed model.

Network	Module	Layer	Filter/ Stride / Output channel
Content stimulus guider	Visual expectation generator	4×ConvLSTM	3×3/ [4×(2, 2)] / [16, 32, 64, 128]
		4×DeConvLSTM	3×3/ [4×(2, 2)] / [64, 32, 16, 3]
	Visual encoder	5×3D-Conv	3×3×3/ [5×(1, 2, 2)] / [8, 16, 32, 64, 64]
	Mismatch encoder	5×3D-Conv	3×3×3/ [5×(1, 2, 2)] / [8, 16, 32, 64, 64]
	Global encoder	1×2D-Conv	3×3/ (1, 1)/ [64]
		1×FC	[64]
Physiological response guider	Freq-domain	6×2D-Conv	3×3/ [(1, 1), (1, 1), (2, 2), (1, 1), (1, 1), (2, 2)] / [32, 32, 32, 32, 32, 32]
		1×ConvLSTM	3×3/ (1, 1)/ [32]
	Time-domain	1×1D-Conv	3/ (2)/ [32]
		4×1D-Resblock	3/ [(2), (2), (2), (2)] / [32, 32, 32, 32]

lated to the content. Given a video content, three temporal sections with equal lengths are divided up. From each section, randomly sampled content video sequence $(I_t, ..., I_{t+N-1})$ and generation difference sequence $(D_t, ..., D_{t+N-1})$ are used as inputs at training phase. Note that $D_t = |\hat{I}_t - I_t|$, and midst frames of each section were sampled at testing phase. Content and difference sequences are fed into a visual encoder and a mismatch encoder, respectively. In this process, visual context and visual mismatch of VR content for each section are encoded with 3D-Conv layers. The output features of the three sections are then combined through a global encoder for extracting the overall characteristics of the content. Output deep stimulus feature $f_s \in \mathbb{R}^{64}$ represents the tendency of sickness-inducing stimulus about the VR content.

2.2. Physiological response guider

The physiological response guider takes individual subject characteristics into consideration to estimate VR sickness. The physiological responses (EEG, EKG, and GSR) are acquired while the subjects watching VR content. Those signals are used as inputs of the physiological response guider. Each original time-domain signal $X \in \mathbb{R}^{60000 \times C}$ passes through a time-domain encoder that consists of stride 1D-Resblock [16]. Note that C is the channel size of the input signal. It is known that the characteristic of frequency band is related to cybersickness [17–19]. In order to consider the frequency characteristics, spectrogram image $\bar{X} \in \mathbb{R}^{60000 \times C}$ of each signal is obtained through Short-Time Fourier Transform (STFT) [20]. $\bar{X}$ is fed into a freq-domain encoder which is composed of 2D-Conv layers. Then, the hidden feature drawn by the freq-domain encoder is divided into five patches in terms of temporal axis. Patches enter the ConvLSTM in temporal order. In this process, the short-term and long-term characteristics can be encoded through the convolutional kernel and the LSTM structure. Then, time domain and frequency domain features are fused. Each fused feature becomes VR sickness related feature of EEG, EKG, and GSR, respectively. The fused features of EEG, EKG, and GSR are again concatenated. Physiology context attention is applied element-wise to the concatenated feature for emphasizing important physiological parts to infer VR sickness. The output of the physiological response guider, deep physiology feature $f_p \in \mathbb{R}^{32}$ reflects the physiological characteristics related with individual VR sickness.

2.3. VR sickness predictor

The VR sickness predictor combines the deep stimulus feature f_s with the deep physiology feature f_p to predict individual SSQ scores. Once f_s and f_p are concatenated, a stimulus context attention is elementwise multiplied to the concatenated feature. This attentive fusion determines which physiological features to be emphasized based on the context of specific stimulus. Then the VR sickness predictor finally estimates the individual SSQ score through fully connected layers. Let P denote the sickness predictor function. The sickness score loss for training can be represented as

$$\mathcal{L}_{SSQ} = \frac{1}{K} \sum_{t \in batch} \|P(f_s, f_p) - SSQ_{indiv}\|_2^2, \quad (2)$$

where SSQ_{indiv} is a ground truth individual SSQ score. At training phase, $\mathcal{L}_{SSQ}$ is back-propagated to overall networks except for the visual expectation generator. The architecture details are shown in Table 1. ReLU [21] was used as an activation function for each layer.

3. BENCHMARK DATABASE

3.1. 360-degree video datasets

We collected normal motion 360-degree videos from Blend and Vimeo to pre-train the visual expectation generator. Each video consists of normal motion with high frame rate (60Hz). Total 32 videos (60s length) include various normal scenes such as slowly driving car and moving drone. In addition, we collected assessment 360-degree videos from Vimeo for subjective experiment and model evaluation. 10 types of video (90s length) were collected, and two versions of frame rate (10Hz, 60Hz) were made. It is known that video with exceptional motion and low frame rate causes cybersickness [4,12]. As a result, total 20 contents with various degrees of sickness were constructed for VR sickness assessment.

3.2. Subjective experiment

A total of 20 subjects participated in the VR content viewing experiment. Three subjects who had withdrawn during the subjective experiment were excluded. Each subject was guided to watch a 90s video twice, and then fill in SSQ sheet [7]. In this process, SSQ score and physiological

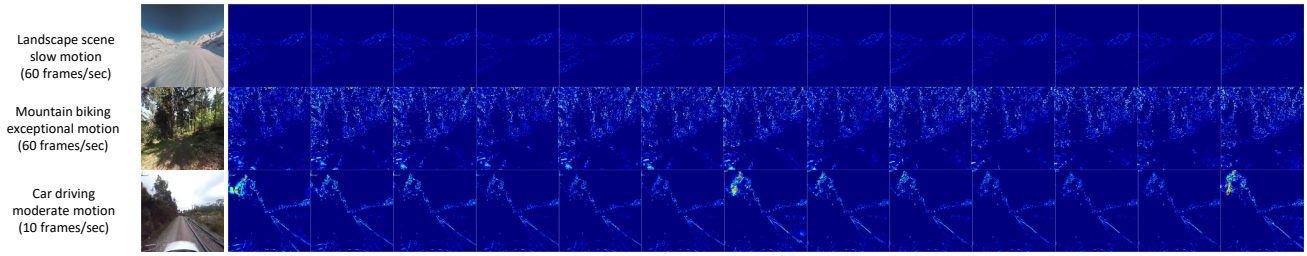

Fig. 2. Difference frame visualization by the visual expectation generator.

Table 2. Prediction performance for individual SSQ score.

Method	PLCC	SROCC	RMSE
Proposed method (physiological response)	0.791	0.551	19.171
Proposed method (physiological response + content stimulus)	0.854	0.700	17.877

Table 3. Prediction performance for mean SSQ score.

Method	PLCC	SROCC	RMSE
Proposed method (physiological response)	0.649	0.635	9.567
Proposed method (physiological response + content stimulus)	0.830	0.819	7.341

signals (EEG, EKG, and GSR) were obtained under the supervision of qualified neuropsychiatry specialists. Experimental settings followed the guideline, ITU-BT.500-13 [22] and BT.2021 [23]. LG 34UC98, Cognionics Quick-30, and Cognionics AIM were used in the experiment.

4. EXPERIMENTAL RESULTS

4.1. Implementation

Considering actual perception, 10Hz video frames are repeated six times to be matched with the length of 60Hz video. The intermediate 120s of each physiological signal was utilized for eliminating the noise of both ends. We used Adam [24] to optimize the proposed network with a learning rate of 0.0002 and a batch size of 16.

4.2. Performance evaluation

We conducted 5-fold cross-validation [25] with the benchmark database. Pearson linear correlation coefficient (PLCC), spearman rank order correlation coefficient (SROCC), and root mean square error (RMSE) were used as performance evaluation metrics.

Table 2 shows prediction performance for the individual SSQ score. Physiological response model indicates that only deep physiology feature was used to regress the SSQ score. The proposed method with physiological response and content stimulus indicates the proposed physiological fusion network. As shown in the table, the proposed method achieved higher performance in terms of all evaluation metrics when stimulus and response were used together. The proposed method achieved meaningful correlation performance of PLCC$\geq$0.8 and SROCC$\geq$0.7 with p-value$\leq$0.05. Table 3 represents prediction performance for the mean SSQ score over each content. We estimated the mean SSQ score of each content by averaging the estimated individual SSQ scores. As shown in the table, the content stimulus feature significantly contributed to the performance for the mean SSQ score. This experimental result indicates that the content stimulus feature could provide VR sickness tendency in terms of the mean SSQ score. Note that the proposed model was not trained to predict the mean SSQ score. Nevertheless, predicting mean SSQ score was achieved with valid performance of PLCC$\geq$0.8 and SROCC$\geq$0.8 with p-value$\leq$0.05.

Fig. 2 shows difference maps between original frames and generated frames. The function of the visual expectation generator was visualized. It can be seen that the large difference occurred for the contents including exceptional motion or low frame rate. This result shows that the content stimulus guider could actually capture the sickness-inducing regions of the VR content.

5. CONCLUSION

In this paper, we proposed the novel deep learning framework that quantifies individual VR sickness with content stimulus and physiological response. To effectively represent the sickness related features, the content stimulus guider and the physiological response guider were devised. These guiders encoded stimulus sickness tendency and individual sickness characteristics to predict individual SSQ scores. The experimental results showed that the proposed method achieved meaningful correlation with both individual and mean SSQ scores. In addition, we contributed to the VR sickness assessment field by constructing the dataset that consists of 360-degree videos with corresponding physiological signals and SSQ scores.

6. REFERENCES

[1] 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, 1993.

[2] K. Carnegie and T. Rhee, "Reducing visual discomfort with hmds using dynamic depth of field," *IEEE computer graphics and applications*, vol. 35, no. 5, pp. 34–41, 2015.

[3] S. Sharples, S. Cobb, A. Moody, and J.R. Wilson, "Virtual reality induced symptoms and effects (vrise): Comparison of head mounted display (hmd), desktop and projection display systems," *Displays*, vol. 29, no. 2, pp. 58–69, 2008.

[4] 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 *VRST*. ACM, 2017, p. 36.

[5] H.G. Kim, H. Lim, S. Lee, and Y.M. Ro, "Vrsa net: Vr sickness assessment considering exceptional motion for 360 vr video," *IEEE transactions on image processing*, vol. 28, no. 4, pp. 1646–1660, 2019.

[6] J. Kim, W. Kim, S. Ahn, J. Kim, and S. Lee, "Virtual reality sickness predictor: Analysis of visual-vestibular conflict and vr contents," in *QoMEX*. IEEE, 2018, pp. 1–6.

[7] S. Bruck and P.A. Watters, "Estimating cybersickness of simulated motion using the simulator sickness questionnaire (ssq): A controlled study," in *CGIV*. IEEE, 2009, pp. 486–488.

[8] C. Chang and C. Lin, "Libsvm: a library for support vector machines," *ACM transactions on intelligent systems and technology*, vol. 2, no. 3, pp. 27, 2011.

[9] Y.Y. Kim, H.J. Kim, E.N. Kim, H.D. Ko, and H.T. Kim, "Characteristic changes in the physiological components of cybersickness," *Psychophysiology*, vol. 42, no. 5, pp. 616–625, 2005.

[10] M.S. Dennison, A.Z. Wisti, and M. DZmura, "Use of physiological signals to predict cybersickness," *Displays*, vol. 44, pp. 42–52, 2016.

[11] D. Egan, S. Brennan, J. Barrett, Y. Qiao, C. Timmerer, and N. Murray, "An evaluation of heart rate and electrodermal activity as an objective qoe evaluation method for immersive virtual reality environments," in *QoMEX*. IEEE, 2016, pp. 1–6.

[12] M. Meehan, B. Insko, M. Whitton, and F.P. Brooks Jr, "Physiological measures of presence in stressful virtual environments," *Acm transactions on graphics*, vol. 21, no. 3, pp. 645–652, 2002.

[13] A. Singla, S. Fremerey, W. Robitza, and A. Raake, "Measuring and comparing qoe and simulator sickness of omnidirectional videos in different head mounted displays," in *QoMEX*. IEEE, 2017, pp. 1–6.

[14] J.T. Reason, "Motion sickness adaptation: a neural mismatch model," *Journal of the royal society of medicine*, vol. 71, no. 11, pp. 819–829, 1978.

[15] S. Xingjian, Z. Chen, H. Wang, D. Yeung, W. Wong, and W. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *NIPS*, 2015, pp. 802–810.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[17] S. Wibirama, H.A. Nugroho, and K. Hamamoto, "Depth gaze and ecg based frequency dynamics during motion sickness in stereoscopic 3d movie," *Entertainment computing*, vol. 26, pp. 117–127, 2018.

[18] C. Lin, S. Chuang, Y. Chen, L. Ko, S. Liang, and T. Jung, "Eeg effects of motion sickness induced in a dynamic virtual reality environment," in *EMBS*. IEEE, 2007, pp. 3872–3875.

[19] B. Patrao, S. Pedro, and P. Menezes, "How to deal with motion sickness in virtual reality," *Sciences and Technologies of Interaction*, pp. 40–46, 2015.

[20] J. Allen, "Short term spectral analysis, synthesis, and modification by discrete fourier transform," *IEEE transactions on acoustics, speech, and signal processing*, vol. 25, no. 3, pp. 235–238, 1977.

[21] V. Nair and G.E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010, pp. 807–814.

[22] "Methodology for the subjective assessment of the quality of television pictures," *ITU-R BT.500-13*, 2012.

[23] "Subjective methods for the assessment of stereoscopic 3dtv systems," *ITU-R BT.2021*, 2012.

[24] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

[25] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal of the royal statistical society. Series B (Methodological)*, pp. 111–147, 1974.