ESTIMATING VR SICKNESS CAUSED BY CAMERA SHAKE IN VR VIDEOGRAPHY

*Seongyeop Kim, Sangmin Lee, and Yong Man Ro**

Image and Video Systems Lab, School of Electrical Engineering, KAIST, South Korea

ABSTRACT

Recent development of Virtual Reality (VR) technology provides more realistic experience for viewers with a variety of contents. While the viewing safety of the viewers is one of the important issues in VR industry, the necessity of VR sickness estimation has been drawing attentions. Inspired by the observations that camera shake in VR videography is one of the major causes of VR sickness, we propose a novel deep network that predicts VR sickness level of individuals caused by camera shake. The proposed method is designed to comprehensively identify changes in direction and speed of the VR video scenes with camera shake. Sparse selection of optical flow maps with different intervals allows the proposed network to efficiently extract stimulus features with a variety of camera shake patterns. We built a new benchmark database for the evaluation of the proposed method that consists of 360-degree videos including various camera shake movements, physiological signals, and Simulation Sickness Questionnaires (SSQ) scores of the experimental participants. Experimental results of the sickness prediction show the effectiveness of the proposed method on the built benchmark database.

***Index Terms*—** Virtual reality, VR sickness assessment, camera shake, deep learning

1. INTRODUCTION

Virtual Reality (VR) contents are now wide-spread for the public providing realistic viewing experiences. 360-degree videos are easily accessible for viewers with VR displays such as head-mounted display (HMD). VR industry is rapidly growing in the glare of public attention, but in the meantime, concerns over possible safety problems also come along. Over 80% of VR contents viewers are reported to experience VR sickness which includes physical symptoms such as nausea, headache, and sweating [1–3]. Although objective assessment of VR sickness is a challenging task due to the existence of various determinants of VR sickness, such research is a necessary step to provide safety guidance for VR contents creators and viewers.

This work was supported by IITP grant (No. 2017-0-00780).
*Corresponding author(ymro@kaist.ac.kr).

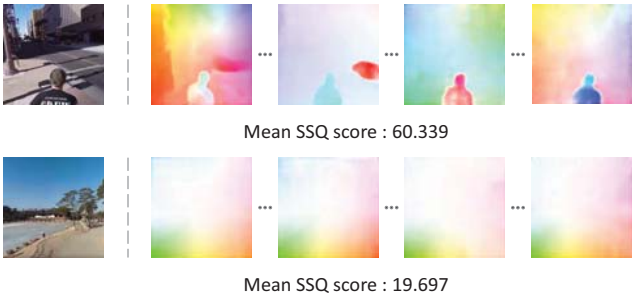

Fig. 1. Visualization results of the optical flow maps acquired from a video with extreme camera shake movements (above) and another video with stable camera movements (below) with the corresponding mean SSQ scores. The visualized optical flow maps are selected with 1 second time interval.

Several attempts for the objective assessment of VR sickness level have been proposed toward various VR sickness inducing factors [4–8]. Kim et al. [4] proposed a deep generative model to assess VR sickness level of 360-degree video scenes including high acceleration. An autoencoder is trained with slow and stationary videos to reconstruct corresponding video frames, so that the autoencoder faces difficulties reconstructing videos with high acceleration at testing phase. In consideration of this characteristic, reconstruction loss is used for quantifying VR sickness. In [5], a deep generative model is used to obtain difference features between the ground truth frames and the reconstructed frames. The difference feature is further regressed to predict Simulation Sickness Questionnaires (SSQ) [9] score. In [6], visual-vestibular sensory conflict of the viewers is pointed out as a sickness inducing factor for VR sickness level assessment. Perceptual motion features of the viewers and statistical content features of the VR contents are regressed to predict VR sickness level with a support vector machine (SVM) [10]. Kim et al. [7] tried to predict VR sickness caused by the quality degradation of VR contents considering spatio-temporal characteristics. Lee et al. [8] proposed a deep network for VR sickness quantification of individual viewers, demonstrating that VR contents viewers experience much VR sickness as the VR contents are displayed with lower frame rates.

In addition to the aforementioned inducing factors, one of the most influential factors that induces sickness while watch-

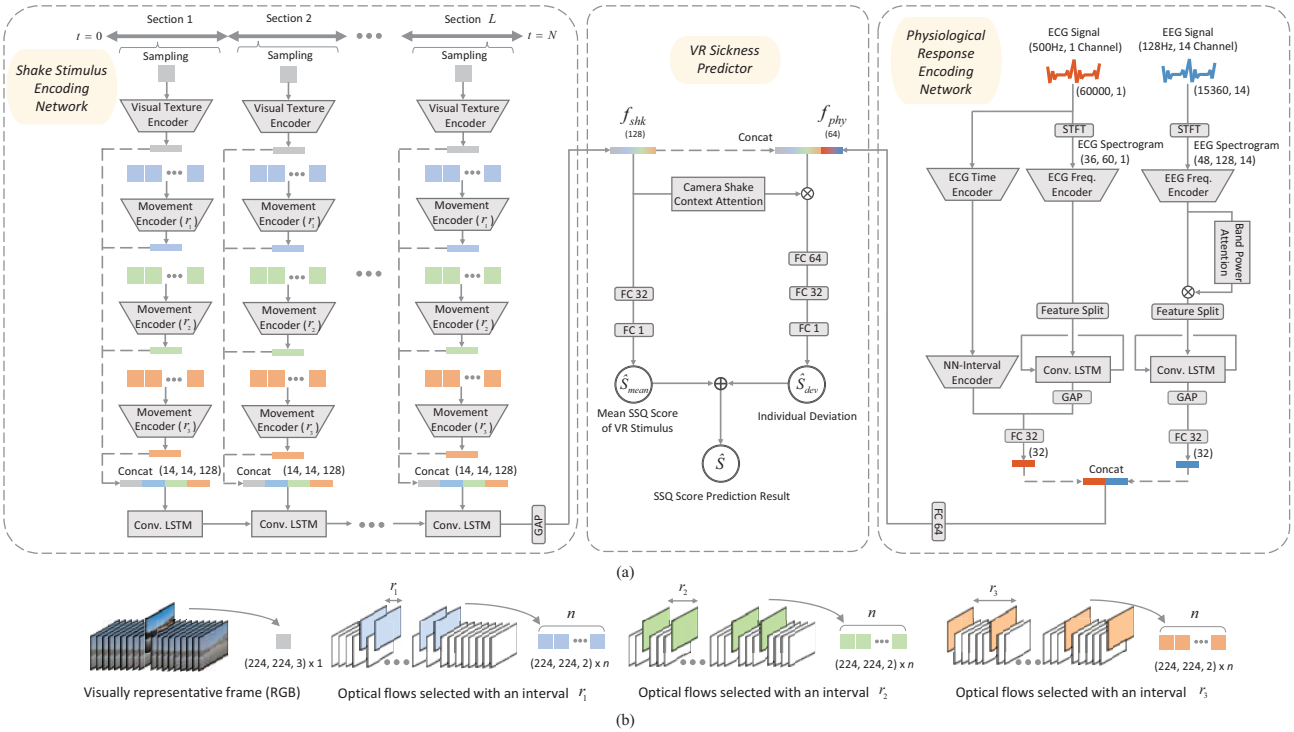

Fig. 2. (a) The overall framework of the proposed VR sickness assessment network caused by camera shake. (b) The input selection procedure for the visually representative frame and the optical flow maps with three different intervals (r_1, r_2, r_3).

ing VR contents is camera shake [11]. Camera shake movement appears at irregular moments throughout the video sequence with random changes in direction and speed. Such visual changes do not accompany with nonvisual inputs (e.g., vestibular) in the VR environment, resulting in a high degree of sensory conflict [11]. Fig. 1 shows the difference between the video with extreme camera movements and the one with stable camera movements, with the optical flow visualization.

In this paper, we propose a deep learning-based VR sickness assessment framework that efficiently captures camera shake movement in 360-degree videography. The proposed deep network is composed of shake stimulus encoding network, physiological response encoding network, and VR sickness predictor. The shake stimulus encoding network takes advantage of movement patterns sampled from different range of intervals to consider the redirection of camera that occurs at various time intervals. The physiological response encoding network estimates VR sickness-related deviations between different individuals on the basis of electrocardiogram (ECG) and electroencephalogram (EEG) of VR contents viewers. The VR sickness predictor analyses features that are acquired from the aforementioned modules to estimate SSQ score of VR contents viewers.

For the validation of the proposed framework, we collected 360-degree video dataset with various degrees of camera movement patterns. The dataset includes various scenes such as driving, bike riding, and dronography with diverse degrees of camera shake movement. 15 subjects participated in the subjective experiment, and the dataset includes corresponding SSQ scores and physiological signals (ECG and EEG) of the participants upon watching the VR contents.

The major contributions of this work are as follows.

- We propose a novel deep learning-based framework that comprehensively identifies camera shake movement throughout the video sequence to assess VR sickness caused by camera shake.
- To evaluate the proposed VR sickness assessment framework, we built a new 360-degree benchmark dataset. The dataset includes 360-degree videos with corresponding physiological signals and SSQ scores of the experimental participants. The 360-degree benchmark dataset is publicly accessible for research purpose. [1]

2. PROPOSED METHOD

Fig. 2 (a) shows the overall architecture of the proposed network. The network consists of shake stimulus encoding network, physiological response encoding network, and VR sickness predictor. The shake stimulus encoding network is designed to encode changes in direction and speed of the 360-degree video scenes from sparsely selected optical flow maps

[1] http://ivylabdb.kaist.ac.kr/base/dataset /VR/Various_Camera_Shaking_Patterns.php

with different intervals. The physiological response encoding module adjusts individual deviations that can be observed from different VR contents viewers for the sickness prediction. The VR sickness predictor predicts the SSQ score on the basis of the VR sickness related features extracted from the shake stimulus encoding network and the physiological response encoding network.

2.1. Shake stimulus encoding network

Changes in direction and speed of the 360-degree video scenes cause sensory conflict for the viewers in VR environment [11]. We design the shake stimulus encoding network to comprehensively grasp randomly changing direction and speed of the video scene throughout the video frame sequence. The shake stimulus encoding network consists of visual texture encoders, multi-interval movement encoders, and shake stimulus extractor. Note that we use the viewport of the 360-degree videos as input of the shake stimulus encoding network.

Given a sequence of N video frames $I_1, \ldots, I_N$, we divide the sequence of N video frames into L sections with the same number of frames. The middle frame of each section is selected as a visually representative frame, and fed into the visual texture encoder. The visual texture encoder consists of four 2D-Conv layers. The visual texture feature $f_{vis} \in \mathbb{R}^{14\times14\times32}$ represents the texture characteristics of the 360-degree video of each section.

The movement encoder takes optical flow maps of the video as input. The optical flow maps are obtained from the viewport of the 360-degree video using PWC-Net [12]. We sparsely select optical flow maps with each frame interval (r_1, r_2, r_3) from every section. We set the frame interval (r_1, r_2, r_3) corresponding to (0.3s, 1s, 2s) of videos in the time scale, respectively. Fig. 2 (b) describes the optical flow map selection procedure. Each set of optical flow maps is fed into the movement encoder, which consists of five 3D-Conv layers. For every section, each movement encoder (r_i) outputs movement feature $f_{mov\_r_i} \in \mathbb{R}^{14\times14\times32}$.

The visual texture feature f_{vis} and the movement features $f_{mov\_r_1}$, $f_{mov\_r_2}$, $f_{mov\_r_3}$ in a section are concatenated, forming L concatenated features. The L concatenated features are recurrently encoded by the shake stimulus extractor which has a ConvLSTM structure [13] and an average pooling layer. The deep feature $f_{shk} \in \mathbb{R}^{128}$ extracted by the shake stimulus extractor represents the sickness-inducing characteristics of the 360-degree video mainly caused by camera shake.

2.2. Physiological response encoding network

The physiological response encoding network is devised to consider individual deviations that may appear from different viewers watching the same VR content. We collected physiological signals (EEG and ECG) from the experimental subjects while watching the VR contents. The network design is based on research findings on the physiological changes that VR sickness causes [14–17].

We consider both the time domain and the frequency domain characteristics of ECG signals to derive VR sickness-related features. To encode the NN-interval deviation from the ECG [15], we devise ECG time encoder and NN-interval encoder. The ECG time encoder and the NN-interval encoder consists of ten and five 1D-Conv layers, respectively. The five output features of the NN-interval encoder layers are concatenated in a channel-wise manner at last to consider different length of NN-interval receptive fields. Spectrogram image of the ECG is obtained by Short-Time Fourier Transform (STFT), and the spectrogram image is encoded by ECG frequency encoder. The ECG frequency encoder consists of six 2D-Conv layers. The resulting feature of the ECG frequency encoder divides into six patches with the same frequency band interval, and the patches are recurrently fed into a sequence of ConvLSTM to consider low frequency to high frequency (LF/HF) ratio of the ECG [14]. The deep feature obtained from the time domain and the frequency domain of the ECG is represented by a vector $f_{ECG} \in \mathbb{R}^{32}$, being combined through a fully connected layer.

Several studies reported observations of the frequency band power changes of EEG signals influenced by VR sickness [16, 17]. We convert the 14 channels of EEG signal of the viewer into spectrogram maps by STFT, and EEG frequency encoder takes the spectrogram maps as input. The EEG frequency encoder consists of three 2D-Conv layers. Then band power attention module gives a guidance of which frequency band the network should concentrate on, among delta, theta, alpha, beta, and gamma waves [17]. The output feature of the EEG frequency encoder is divided into four patches with the same time interval, and recurrently fed into the ConvLSTM. The deep feature of the EEG is represented by a vector $f_{EEG} \in \mathbb{R}^{32}$ through a fully connected layer.

The features f_{ECG} and f_{EEG} are fused with another fully connected layer, thus the output deep feature $f_{phy} \in \mathbb{R}^{64}$ of the physiological response network represents the individual deviation of the VR viewer.

2.3. VR sickness predictor

VR sickness predictor is designed to predict the SSQ score of a viewer while given the two VR sickness representative features f_{shk} and f_{phy}. First, the VR sickness predictor estimates the mean SSQ score of a VR content given f_{shk}. A loss function for the mean SSQ score prediction can be written as

$$\mathcal{L}_{mean} = \|P_m(f_{shk}) - SSQ_{mean}\|_2^2, \tag{1}$$

where P_m denotes the mean SSQ score prediction function. Having f_{shk} as prior knowledge, the rest procedure is to predict the level of individual deviation watching the VR content. Both the f_{shk} and f_{phy} are considered for the individual deviation prediction with an emphasis on the camera shake con-

text of the same VR content. A loss function for the individual deviation prediction can be described as

$$\mathcal{L}_{ind} = ||P_m(f_{shk}) + P_d(f_{shk}, f_{phy}) - SSQ_{ind}||_2^2, \quad (2)$$

where P_d denotes the individual deviation level prediction function. The sum of the predicted mean SSQ score of the VR content and the level of individual deviation is the final SSQ score prediction result of the viewer watching the VR content. The total loss function for the network training is as follows.

$$\mathcal{L}_{\mathcal{SSQ}} = \mathcal{L}_{mean} + \mathcal{L}_{ind} \quad (3)$$

The overall network is trained in an end-to-end manner by minimizing the loss term $\mathcal{L}_{\mathcal{SSQ}}$ with the gradient descent algorithm.

3. BENCHMARK DATABASE

3.1. 360-degree video datasets

We collected twenty 360-degree videos that include various degrees of camera shake patterns from YouTube and Vimeo. The dataset includes scenes such as driving, bike riding, dronography, etc. All the videos have 4K resolution (3840×2048 or 4096×2048) with 30 Hz frame rates, and we clipped the videos to have 90 seconds of duration.

3.2. Subjective experiment

A total of 15 subjects participated in the subjective experiment. The 360-degree videos were displayed with Pimax 5K Plus and Whirligig VR player. We collected the physiological signals (ECG and EEG) of the participants while watching the VR contents with Cognionics AIM Gen2 and EMOTIV EPOC+ 14 Channel Mobile EEG. The participants were guided to watch each video twice in a row (180 seconds) and to answer the SSQ for the corresponding video. The experiment had been conducted under the supervision of qualified neuropsychiatry specialists, and we followed ITU-BT.500-13 [18] and BT.2021 [19] as experimental setting guidelines.

4. EXPERIMENTAL RESULTS

4.1. Implementation

We implemented the proposed work on Python 2.7 environment with Tensorflow version 1.3.0 [21]. We used Adam [22] to optimize the proposed network with a learning rate of 0.0002 and a batch size of 16 on NVIDIA TITAN Xp GPUs. In order to stabilize the training process of the network, we conducted data augmentation by sampling 120 seconds of the acquired physiological signal dataset with five different time ranges.

4.2. Performance evaluation

We evaluated the proposed network with three metrics regarding how well the proposed network predicted SSQ score of individuals watching VR contents. The three metrics are: Pearson linear correlation coefficient (PLCC), Spearman's rank order correlation coefficient (SROCC), and root mean square error (RMSE). We conducted 5-fold cross-validation [23] with the benchmark database.

Table 1. Individual subject VR sickness prediction performance on the benchmark database.

Method	PLCC	SROCC	RMSE
Peak interval feature -based method (ECG) [20]	0.340	0.237	46.469
Band power feature -based method (EEG) [17]	0.492	0.352	35.157
Velocity-based method	0.673	0.536	30.229
Proposed method with Interval (r_1)	0.732	0.662	29.407
Proposed method with Intervals (r_1, r_2)	0.769	0.650	27.635
Proposed method with Intervals (r_1, r_2, r_3)	**0.804**	**0.698**	**27.406**

The performance comparison for VR sickness prediction is shown in Table 1. Peak interval feature-based method (ECG) predicts VR sickness based on four major ECG features (MeanRR, SDRR, pNN50, and NN50) [20]. Band power feature-based method (EEG) [17] utilizes the power of theta, alpha, low-beta, and gamma. The mean velocity of each optical flow map is used as velocity feature in the velocity-based method as in [24]. Note that the velocity features are regressed to predict the SSQ score along with the ECG and EEG signal of individuals. As shown in the Table 1, the proposed method outperforms other sickness assessment methods in terms of all evaluation metrics.

We also conducted an ablation study on the advantage of the multi-interval movement encoding. Selection of optical flow maps with different intervals positively affects the performance of VR sickness prediction by capturing redirection occurring at various time intervals.

5. CONCLUSION

In this paper, we proposed a deep network estimating VR sickness caused by camera shake in VR videography. The proposed method comprehensively considers camera shake movement patterns throughout the video sequence to predict the VR sickness level of the viewer. The experimental results showed that the camera movement of video scenes in VR environment evidently affects the level of VR sickness. We newly built 360-degree video dataset for validation, including physiological signals of the participants with the 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, pp. 1–7.

[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, 2018.

[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] K. Kim, S. Lee, H.G. Kim, M. Park, and Y.M. Ro, "Deep objective assessment model based on spatiotemporal perception of 360-degree video for vr sickness prediction," in *ICIP*. IEEE, 2019, pp. 3192–3196.

[8] S. Lee, S. Kim, H.G. Kim, M.S. Kim, S. Yun, B. Jeong, and Y.M. Ro, "Physiological fusion net: Quantifying individual vr sickness with content stimulus and physiological response," in *ICIP*. IEEE, 2019, pp. 440–444.

[9] 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.

[10] C. Chang and C. Lin, "Libsvm: A library for support vector machines," *TIST. ACM*, vol. 2, no. 3, pp. 1–27, 2011.

[11] F. Bonato, A. Bubka, S. Palmisano, D. Phillip, and G. Moreno, "Vection change exacerbates simulator sickness in virtual environments," *Presence: Teleoperators and Virtual Environments*, vol. 17, no. 3, pp. 283–292, 2008.

[12] D. Sun, X. Yang, M. Liu, and J. Kautz, "Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume," in *CVPR*, 2018, pp. 8934–8943.

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

[14] 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.

[15] 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.

[16] 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.

[17] S. Chuang, C. Chuang, Y. Yu, J. King, and C. Lin, "Eeg alpha and gamma modulators mediate motion sickness-related spectral responses," *International journal of neural systems*, vol. 26, no. 02, pp. 1650007, 2016.

[18] BT Series, "Methodology for the subjective assessment of the quality of television pictures," *Recommendation ITU-R BT*, pp. 500–13, 2012.

[19] IT Union, "Subjective methods for the assessment of stereoscopic 3dtv systems," *Recommendation ITU-R BT*, vol. 2021, 2015.

[20] F. Shaffer and J. Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers in public health*, vol. 5, pp. 258, 2017.

[21] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, 2016, pp. 265–283.

[22] D.P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[23] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 36, no. 2, pp. 111–133, 1974.

[24] N.M. Palomino and G.C. Chávez, "Abnormal event detection in video using motion and appearance information," in *Iberoamerican Congress on Pattern Recognition*. Springer, 2017, pp. 382–390.