|  |  |
| --- | --- |
| Project | **HMD based 3D Content Motion Sickness Reducing Technology**  <<http://sites.ieee.org/sagroups-3079/>> |
| Title | **Deep learning-based VR sickness assessment with content stimulus and physiological response** |
| DCN | **3079-19-0021-00-0002** |
| Date Submitted | **July 5, 2019** |
| Source(s) | **Sangmin Lee** [sangmin.lee@kaist.ac.kr](mailto:sangmin.lee@kaist.ac.kr) **(KAIST), Seongyeop Kim** [seongyeop@kaist.ac.kr](mailto:seongyeop@kaist.ac.kr) **(KAIST), Hak Gu Kim** [hgkim0331@kaist.ac.kr](mailto:hgkim0331@kaist.ac.kr) **(KAIST), Yong Man Ro** [ymro@kaist.ac.kr](mailto:ymro@kaist.ac.kr) **(KAIST)** |
| Re: |  |
| Abstract | With the rapid development of VR equipment and 360-degree video acquisition device, VR contents have increasingly attracted attention in industry and research fields. In viewing VR contents, VR sickness could be induced due to visual-vestibular conflict. The degree of the visual-vestibular conflict felt by each person may differs even for the same content stimulus. In this document, we introduce a novel deep learning framework to assess individual VR sickness with content stimulus and physiological response. |
| Purpose | The goal of this document is to deal with a deep learning-based individual VR sickness assessment framework by considering content stimulus and physiological response for evaluating the overall degree of perceived VR sickness in viewing VR content. |
| Notice | This document has been prepared to assist the IEEE 802.21 Working Group. It is offered as a basis for discussion and is not binding on the contributing individual(s) or organization(s). The material in this document is subject to change in form and content after further study. The contributor(s) reserve(s) the right to add, amend or withdraw material contained herein. |
| Release | The contributor grants a free, irrevocable license to the IEEE to incorporate material contained in this contribution, and any modifications thereof, in the creation of an IEEE Standards publication; to copyright in the IEEE’s name any IEEE Standards publication even though it may include portions of this contribution; and at the IEEE’s sole discretion to permit others to reproduce in whole or in part the resulting IEEE Standards publication. The contributor also acknowledges and accepts that IEEE 802.21 may make this contribution public. |
| Patent Policy | The contributor is familiar with IEEE patent policy, as stated in Section 6 of the IEEE-SA Standards Board bylaws <[http://standards.ieee.org/guides/bylaws/sect6-7.html#6](http://standards.ieee.org/guides/bylaws/sect6-7.html%236)> and in *Understanding Patent Issues During IEEE Standards Development* <<http://standards.ieee.org/board/pat/faq.pdf>> |

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. Generally, 80\% to 95\% of people feel VR sickness. 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. Kim et al. 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. 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)[[1]](#footnote-1) score. The aforementioned VR sickness quantification methods estimated mean value of SSQ score, not individual VR sickness. Another study quantified VR sickness caused by visual-vestibular conflict. In this work, SVM 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 document, 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. 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 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.

1. Proposed Method
   1. Overview

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.

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

* 1. Content Stimulus Guider

VR sickness could arise if sensory information that an individual perceives does not correspond with the normal experience. 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 frames to generate the next frame (). The generator consists of ConvLSTM 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 denote the generator function. The generation loss can be written as

|  |  |
| --- | --- |
|  | (1) |

where is a mini batch size at training phase.

Based on the visual expectation generator, the stimulus context extractor outputs deep stimulus feature which is related to the content. Given a video content, three temporal sections with equal lengths are divided up. From each section, randomly sampled content video sequence () and generation difference sequence () are used as inputs at training phase. Note that , 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 represents the tendency of sickness-inducing stimulus about the VR content.

* 1. 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 passes through a time-domain encoder that consists of stride 1D-Resblock. Note that is the channel size of the input signal. It is known that the characteristic of frequency band is related to cybersickness. In order to consider the frequency characteristics, spectrogram image of each signal is obtained through Short-Time Fourier Transform (STFT). 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 reflects the physiological characteristics related with individual VR sickness.

* 1. VR Sickness Predictor

The VR sickness predictor combines the deep stimulus feature with the deep physiology feature to predict individual SSQ scores. Once and 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 denote the sickness predictor function. The sickness score loss for training can be represented as

|  |  |
| --- | --- |
| , | (1) |

where is a ground truth individual SSQ score. At training phase, is back-propagated to overall networks except for the visual expectation generator. ReLU was used as an activation function for each layer.

1. Benchmark database
   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. As a result, total 20 contents with various degrees of sickness were constructed for VR sickness assessment.

* 1. 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. In this process, SSQ score and physiological signals (EEG, EKG, and GSR) were obtained under the supervision of qualified neuropsychiatry specialists. Experimental settings followed the guideline, ITU-BT.500-13 and BT.2021. LG 34UC98, Cognionics Quick-30, and Cognionics AIM were used in the experiment.

1. Experimental Results
   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 to optimize the proposed network with a learning rate of 0.0002 and a batch size of 16.

* 1. Performance Evaluation

We conducted 5-fold cross-validation 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 1 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 PLCC0.8 and SROCC0.7 with p-value0.05. Table 2 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 PLCC0.8 and SROCC0.8 with p-value0.05.

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.

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

Table 1.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 2.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 |

1. Conclusion

In this document, 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.

1. 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 *International Conference on Image Processing (ICIP). IEEE*, 2019. [↑](#footnote-ref-1)