|  |  |
| --- | --- |
| Project | **HMD based 3D Content Motion Sickness Reducing Technology**  <<http://sites.ieee.org/sagroups-3079/>> |
| Title | **Deep learning-based VR sickness assessment considering content quality factor** |
| DCN | **3079-19-0020-00-0002** |
| Date Submitted | **July 5, 2019** |
| Source(s) | **Sangmin Lee** [sangmin.lee@kaist.ac.kr](mailto:hgkim0331@kaist.ac.kr) **(KAIST), Kihyun Kim** [noddown@kaist.ac.kr](mailto:noddown@kaist.ac.kr) **(KAIST), Hak Gu Kim** [hgkim0331@kaist.ac.kr](mailto:hgkim0331@kaist.ac.kr) **(KAIST), Minho Park** [roger618@kaist.ac.kr](mailto:roger618@kaist.ac.kr) **(KAIST), Yong Man Ro** [ymro@kaist.ac.kr](mailto:ymro@kaist.ac.kr) **(KAIST)** |
| Re: |  |
| Abstract | With the development of 360-degree camera capture system and head mounted display (HMD), HMD-based VR contents have attracted a lot of attention of customers and industries. In viewing VR contents with HMD, VR sickness could be induced by quality degradation of the VR contents. In particular, low resolution contents induce higher simulator sickness than high resolution contents with respect to spatial and temporal inconsistency. In this document, we introduce a novel deep learning framework approach to assess VR sickness caused by quality degradation considering spatio-temporal perceptual characteristics of VR contents. |
| Purpose | The goal of this document is to deal with a deep learning-based objective VR sickness assessment framework by considering spatio-temporal perceptual characteristics of VR contents for evaluating the overall degree of perceived VR sickness in viewing VR content with HMD. |
| 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) becomes popular in many fields because VR can give immersive and dynamic experience for VR viewers. Recently, the VR technology is adopted in various applications such as entertainment, simulation training, health care, education and so on. With the recent development of 360 cameras and VR displays, the popularity of VR is continuously increasing.

Although VR is useful in various applications and the number of VR users is increasing, there are concerns that VR sickness could occur during VR consuming. VR sickness is actually one of main problems hampering the spread of VR. VR sickness is accompanied by physical symptoms: 1) nausea symptoms including sweating, salivation and burping, 2) oculomotor symptoms including visual fatigue and eye strain, and 3) disorientation symptoms including fullness of the head, dizziness and vertigo.

To resolve the problem of VR viewing safety, it is needed to quantify and predict the degree of VR sickness for VR content. Most of previous works made an effort to measure VR sickness with physiological signals or subjective questionnaires in subjective experiments. These approaches were cumbersome and time consuming due to measuring physiological signals with bio sensors and subjective questionnaires such as simulator sickness questionnaires (SSQ)[[1]](#footnote-1). for VR sickness evaluation.

Recently, VRSA (VR sickness assessment) methods have been proposed. Deep networks were proposed to predict VR sickness caused by the exceptional motion in VR video. Using the auto-encoder based deep architecture, they predicted VR sickness caused by exceptional motion. In practical, the quality-degradation in VR video is common and it causes VR sickness as well.

In this document, we propose a novel deep learning-based VR sickness assessment (VRSA) framework that predicts VR sickness caused by quality-degradation of VR video in space-time domain. In testing, the proposed network consists of spatial encoder, temporal encoder and sickness score predictor. In training stage, the spatial encoder and temporal encoder are trained in cooperating with spatial perception-guider and temporal perception-guider, respectively. The spatial and temporal perception-guiders estimate the degraded quantity of input compared to reference so that spatio-temporal perceptional characteristics of input video are encoded in training stage. Finally, the predictor estimates SSQ score from the latent features encoded by spatial and temporal encoders. In testing, VR sickness score is predicted without the guider networks and reference video.

For the validation of the proposed method, we collected a new dataset including 360-degree videos. These videos contain various scenes such as driving, bicycling, sailing and drone. From reference videos (UHD), the degraded videos of FHD, HD and SD were acquired for our subjective experiments. With the 360-degree video datasets consisting of different spatial resolutions, we conducted extensive subjective assessment experiments to verify the effectiveness of the proposed method. In addition, we collected physiological data such as heart rate (HR) signal and galvanic skin response (GSR) signal to compare performance as a benchmark.

1. Proposed Method
   1. Overview

Fig. 1 shows the overall architecture of the proposed deep objective assessment model for VR sickness assessment (VRSA). The proposed network consists of spatial encoder, temporal encoder, spatial perception-guider, temporal per-ception-guider, and sickness score predictor. Let and denote the distorted and reference frame, respectively. To consider the spatio-temporal perception of VR content caused by video encoding and resolution degradation, the spatial perception-guider estimates the spatial inconsistency based on the encoded spatial features of and . The temporal perception-guider estimates the temporal inconsistency based on the encoded spatio-temporal features of and . With the spatial and temporal perception-guiders, the spatial encoder and temporal encoder can extract spatio-temporal perception characteristics affecting the level of VR sickness. Finally, the sickness score is estimated by sickness score predictor.

|  |
| --- |
|  |
|  |
| Fig. 1. Overview of the proposed objective VRSA framework. |

* 1. Spatial Encoder and Spatial Perception-guider

In the proposed spatial encoder, the spatial features of each distorted and reference frame, and . By considering field of view (FoV) of VR display, a viewport with 12001200 pixels is extracted from equirectangular projection and is used as input. It takes five consecutive degraded and reference frames as an input in training. Let and denote the spatial features of -th distorted and reference frames, respectively. The spatial encoder consists total 11 33 convolutional layers and 6 max pooling layers.

In the proposed method, to measure the spatial inconsistency of each frame, we design a spatial perception-guider network. In the spatial perception-guider, the structural similarity (SSIM) is employed to quantify the inconsistency of spatial perception. Estimating the spatial inconsistency of the distorted frame helps the spatial encoder reliably extract the spatial perception at each frame. As shown in Fig. 1, the proposed spatial perception-guider takes the concatenated spatial feature of distorted and reference frames, , as an input. After performing global average pooling, the SSIM index is predicted by three fully connected layers. For training the spatial perception-guider, the spatial inconsistency loss, , can be written as

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

where means a ground-truth SSIM index for -th frame. and indicate the function of spatial perception-guider and the predicted SSIM index, respectively.

is back-propagated to the spatial encoder as well as spatial perception-guider network during training. The spatial encoder can learn how to encode spatial perception of each distorted frame by comparing the corresponding reference frame. The spatial perception-guider is not used in testing.

* 1. Temporal Encoder and Temporal Perception-guider

To encode temporal perception feature of given VR video, we devise a temporal encoder and a temporal perception-guider. The temporal encoder consists of three convolutional LSTM (Conv-LSTM) layers with 33 filter for encoding spatio-temporal feature of , . By iteratively taking consecutive spatial features from spatial encoder as an input, the temporal encoder learns spatio-temporal information such as temporal dynamics.

In addition, we propose a temporal perception-guider in order to give temporal inconsistency information to the temporal encoder. The temporal perception-guider is to make the temporal encoder encode temporal inconsistency by measuring temporal flicker using flicker score, . Let denote the concatenated spatio-temporal feature of distorted and reference frames. As shown in Fig. 1, is used as input of the temporal perception-guider. After the global average pooling is performed, the predicted flicker score is obtained by three fully connected layers. The temporal encoder and temporal perception-guider are trained by minimizing the temporal inconsistency loss, , which can be written as

|  |  |
| --- | --- |
|  | (2) |

where is a ground-truth of flicker score computed by differentiating input frames from reference. and indicate the function of temporal perception-guider and the predicted flicker score, respectively.

is back-propagated to the temporal encoder as well as the temporal perception-guider. The temporal encoder can learn how to encode temporal perception of consecutive distorted frames by comparing the corresponding reference frames. The temporal perception-guider is not used in testing.

* 1. Sickness Score Predictor

After training the spatial encoder, the spatial perception-guider, the temporal encoder and the temporal perception-guider, the sickness score predictor is trained. The sickness score predictor consists of three fully connected layers. The sickness score prediction loss, , can be written as

|  |  |
| --- | --- |
|  | (3) |

where indicates the ground-truth SSQ score of -th VR video content. and indicate the function of the score predictor and the predicted SSQ score, respectively.

In this document, the ground-truth SSQ score of each VR video is obtained by averaging SSQ scores of all subjects. Since each subject's perceived VR sickness level can be different for the same video, we additionally take into account the standard deviation of SSQ with gaussian noise . Therefore, final SSQ prediction loss, , can be defined as

|  |  |
| --- | --- |
|  | (4) |

where indicates the standard deviation of SSQ scores obtained from all subjects for -th video content. Weight parameter is set as 0.2.

1. Benchmark Database
   1. Dataset Generation

To train the proposed network and evaluate the performance, we collected twenty 360-degree video datasets, which are represented in equirectangular projection with 38402160 pixels (UHD), from Vimeo, public video sharing platform. The collected videos contained various scenarios such as driving, bicycling, sailing, drone view, etc. To generate the degradation of spatial resolution in 360-degree videos, we down-sampled the twenty 360-degree videos to three different spatial resolutions, which are SD (640480), HD (1080720), FHD (19201080) using Adobe Premiere 2017. Their frame rate was 30 Hz. As a result, a total number of 80 videos (80 videos = 20 contents 4 spatial resolutions) were obtained for evaluation.

* 1. Subjective Assessment Experiment

A total of 17 subjects participated in our subjective assessment experiment for VRSA. Oculus Rift CV1 with Whirligig player was utilized for displaying 360-degree video datasets. All experimental settings followed the guideline, ITU-BT. 500-13 and BT.2021.

In our subjective assessment experiment, we measured simulator sickness questionnaire (SSQ) scores using the single stimulus (SS) method. Each stimulus was displayed for 60 s. Then, subjects scored their perceived the level of VR sickness in 16-item SSQ sheet and took a rest for 120 s. While subjects watched each stimulus, their HR and GSR were measured by NeuLog sensors, simultaneously. Our experiments consisted of 4 sessions. Each session was conducted on different days. During each session, subjects could immediately stop and take a rest if they felt difficult to continue due to excessive sickness.

* 1. Subjective Experiment Result

Through subjective experiment results, we got SSQ scores for 80 videos that have 20 contents and 4 resolution types (UHD, FHD, HD and SD). As shown in Fig. 2, videos with low resolution generally had higher sickness than those with high resolution. Each shape in resolution types means individual contents. Average SSQ score of all contents according to resolution types is as follows: UHD = 25.855, FHD = 26.822, HD = 28.191, SD = 38.434. Especially, VR sickness score perceived from SD resolution (i.e., 38.434) was higher than 30. It means that watching SD videos in VR environments could be harmful with respect to viewing safety for VR. However, not all contents have the same tendency according to resolution type in VR sickness. This means that not only sickness is caused by the resolution type but also the sickness is determined by the characteristics of contents such as scene complexity. Therefore, it can be inferred that in order to predict the sickness well, the spatio-temporal perception of the content due to the degradation of the resolution type is essentially considered.

|  |
| --- |
|  |
| Fig. 2. SSQ results according to resolution types. |

1. Performance Evaluation Results

To validate the performance of the proposed network, we used three metrics: Pearson linear correlation coefficient (PLCC), Spearman rank order correlation coefficient (SROCC), and Root mean square error (RMSE). PLCC is a measurement of the linear correlation between two variables. SROCC is a nonparametric measure of rank correlation and it evaluates how well the relationship for the monotonic function. RMSE is a metric for assessing simple distances between two variables. We computed PLCC, SROCC and RMSE using non-linear function in order to transform each value of other methods to SSQ score domain in case of using not directly regressing SSQ scores.

HR-based method and GSR-based method preprocessed physiological signals obtained from each subject. In the case of HR-based method, we computed the standard deviation from each subjective HR signal, and then averaged standard deviation values to all subjects. In the GSR method, we directly averaged GSR signals to all subjects. Resolution based method used only numerical value of resolution to regress SSQ score. For example, all contents in UHD, FHD, HD, and SD had the same values as 38402160, 19201080, 1080720 and 640480, respectively. We used these processed values as the value representing sickness score of the corresponding contents.

Table 1 shows the performance comparison for VR sickness assessment on our database. As seen in Table 1, physiological response-based methods such as HR based method and GSR based method had a correlation about less than 0.5 in PLCC. The regression method considering only resolution type without considering characteristics of VR contents had a lower correlation as well. CNN-ConvLSTM based method using deep learning had a higher correlation than those of HR based method, GSR based method, and resolution-based method. However, its performance was still low. On the other hand, the proposed method achieved the highest correlation and the lowest RMSE compared with other methods. As seen in Table 1, the proposed spatial perception guider and temporal perception guider played an important role in VRSA. Since the proposed VRSA predicted SSQ score considering spatio-temporal perception using the proposed guiders, it could acquire higher correlation than conventional CNN-ConvLSTM based method.

Table 1. Prediction performance on our benchmark database

|  |  |  |  |
| --- | --- | --- | --- |
| **Objective metrics** | **PLCC** | **SROCC** | **RMSE** |
| Resolution based method | 0.380 | 0.369 | 12.249 |
| HR based method | 0.468 | 0.295 | 11.700 |
| GSR based method | 0.481 | 0.388 | 11.611 |
| CNN-ConvLSTM based method` | 0.674 | 0.655 | 10.804 |
| Proposed method  with spatial perception-guider | 0.784 | 0.751 | 10.233 |
| Proposed method  with temporal perception-guider | 0.805 | 0.786 | 8.446 |
| **Proposed method** | **0.827** | **0.838** | **8.208** |

1. Conclusions

In this document, we proposed a novel deep learning-based VR sickness assessment framework considering spatio-temporal characteristics of 360-degree videos. To the best of our knowledge, this is the first deep learning framework that quantifies VR sickness caused by quality degradation. Considering VR sickness caused by spatio-temporal perception according to different resolution types, we devised spatial and temporal perception-guider networks to help the spatial encoder and temporal encoder to extract spatio-temporal perception information. The experimental results showed that our proposed network had a meaningful correlation between predicted SSQ score and human subjective SSQ score. Finally, we contributed to the development of research for VR sickness assessment by building a new dataset that consists of 360-degree videos (stimuli), physiological signals and the corresponding SSQ scores.

1. K. Kim\*, S. Lee\*, H. G. Kim, M. Park and Y. M. Ro, “Deep objective assessment model based on spatio-temporal perception of 360-degree video for vr sickness prediction,” in *International Conference on Image Processing (ICIP). IEEE*, 2019. [↑](#footnote-ref-1)