VISUAL COMFORT ASSESSMENT OF STEREOSCOPIC IMAGES USING DEEP VISUAL AND DISPARITY FEATURES BASED ON HUMAN ATTENTION

Hyunwook Jeong, Hak Gu Kim and Yong Man Ro[*]

School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST)

ABSTRACT

This paper proposes a novel visual comfort assessment (VCA) for stereoscopic images using deep learning. To predict visual discomfort of human visual system in stereoscopic viewing, we devise VCA deep networks to latently encode perceptual cues, which are visual differences between stereoscopic images and human attention-based disparity magnitude and gradient information. To extract the visual difference features from left and right views, a Siamese network is employed. In addition, human attention region-based disparity magnitude and gradient maps are fed to two individual deep convolutional neural networks (DCNNs) for disparity-related features based on human visual system (HVS). Finally, by aggregating these perceptual features, the proposed method directly predicts the final visual comfort score. Extensive and comparative experiments have been conducted on IEEE-SA dataset. Experimental results show that the proposed method can yield excellent correlation performance compared to existing methods.

Index Terms— Visual comfort assessment, deep learning, stereoscopic images, human visual system

1. INTRODUCTION

Stereoscopic three-dimensional (S3D) contents have received a significant interest from industries and research fields due to enhanced viewing experiences. In the stereoscopic imaging, S3D contents can provide a unique viewing experience to viewers with the binocular disparity between left and right views. However, with the increasing interest of the S3D contents, the concerns are emerging for the safety of stereoscopic imaging. Many studies have reported that the current stereoscopic displays could lead to various visual fatigue symptoms such as eye strain, headache, focusing difficulty, and nausea [1] – [3]. To address the perceptual problems, existing solutions require labor-intensive, costly and time-consuming process for manually finding visually uncomfortable scene and performing post-processing [4]. As a result, it is essential to develop a reliable objective visual comfort assessment (VCA), aiming to automatically predict the visual discomfort of the stereoscopic images. In the current stereoscopic displays, there are several determinative factors affecting the visual discomfort such as disparity magnitude, disparity gradient, binocular mismatches, etc [2], [5]. The excessive disparity magnitude is one of the causative factors causing a severe conflict between accommodation and vergence, which could lead to visual discomfort and fatigue [6]. Moreover, the disparity gradient is also the important factor limiting the binocular fusion. Even for the disparity magnitude within a comfortable viewing zone, the excessive disparity gradient could lead to the binocular fusion failure [5].

Despite of the significance of visual discomfort factors, many existing VCA methods for stereoscopic images considered simple features such as global statistics of the disparity magnitude and depth variance in an entire image [7]. There are a few attempts to explicitly take into account the visual discomfort features based on human visual system (HVS) [5]. The authors of [5] tried to apply the human visual attention model to the VCA by combining the saliency information and visual discomfort features. In [8] – [10], various factors affecting the visual discomfort such as the object size, width, motion, and disparity information were utilized to predict the visual discomfort during S3D viewing. In [11], a scene mode classification-based VCA method was proposed considering the disparity types of foreground object and background region. The author of [6] developed a S3D visual discomfort predictor using the disparity-based coarse features and neural activity-based fine features.

In recent years, machine learning-based VCA methods were proposed that established the regression models from visual discomfort features to the subjective scores [12]. They tried to solve the visual comfort prediction problems by learning the nonlinear mapping from objective feature space to subjective scores space with the machine learning algorithms such as support vector machine and random forest. Nonetheless, the state-of-the-art machine learning-based VCA methods still had difficulties for training the regression model which well mapped the learned feature space to the subjective score space.

In this paper, we deploy a novel VCA method for stereoscopic images based on deep learning with the characteristics of visual content of stereoscopic images and human visual attention-based disparity. The proposed deep network for VCA stereoscopic images contains two networks,

[*] Corresponding author (ymro@ee.kaist.ac.kr)

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

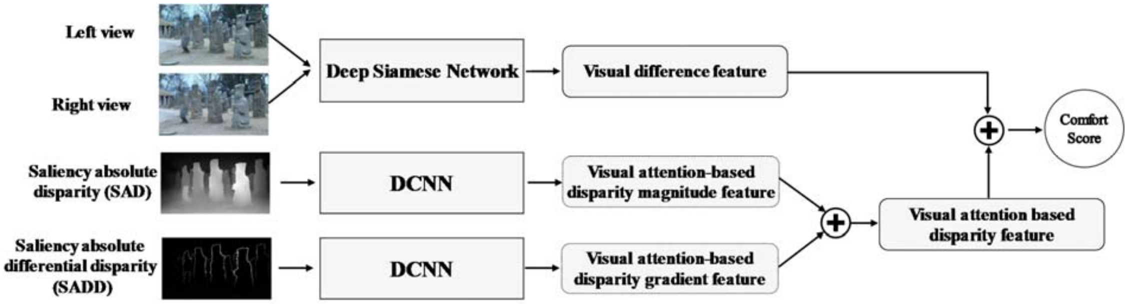
Left view
Right view
Saliency absolute disparity (SAD)
Saliency absolute differential disparity (SADD)
Deep Siamese Network
DCNN
DCNN
Visual difference feature
Visual attention-based disparity magnitude feature
Visual attention-based disparity gradient feature
Visual attention based disparity feature
Comfort Score

Fig. 1. Overall procedure of the proposed stereoscopic image VCA framework. Proposed network has multiple inputs to learn visual difference feature (deep Siamese CNN) and to consider human attention (DCNN)

which are separately trained: 1) Deep Siamese convolutional neural network (CNN) for learning the regression model from visual differences between left and right views to subjective scores, 2) Deep convolutional neural networks (DCNNs) for learning the regression model from the human visual attention-based disparity magnitude and gradient features to the subjective scores. Finally, by learning to aggregate the visual discomfort features from two deep networks, the proposed deep learning-based VCA method effectively predicts visual comfort scores. Also, in the insufficient VCA database for stereoscopic images, we employ the transfer leaning to avoid over-fitting problems. The performance of the proposed method is validated via comparative experiments on the IEEE Standard Association (IEEE-SA) stereoscopic image database [6]. Experimental results have shown that the proposed VCA method has a high correlation with human subjective scores.

The rest of this paper is organized as follows. In Section 2, we describe the proposed deep learning-based VCA method to objectively assess the degree of the visual comfort in stereoscopic images. In Section 3, we describe experiments and results. Finally, conclusions are drawn in Section 4.

2. PROPOSED METHOD

Fig. 1 shows the overall procedure of the proposed VCA based on deep learning. It consists of three main parts. In the first part, the visual difference features from left and right views are extracted by deep Siamese CNN. By the deep Siamese CNN, the relation between the visual difference of stereoscopic images and degree of visual discomfort is learned. In the second part, disparity-related features considering the human attention model are extracted by two different DCNNs. To that end, saliency-weighted absolute disparity (SAD) and saliency-weighted absolute differential disparity (SADD) maps based on the human attention regions are fed to two DCNNs. By the DCNNs, the relation between the human attention region-based disparity features and degree of visual discomfort is learned. In the last part, the perceptual feature aggregating the visual difference, disparity magnitude and gradient features are regressed to the subjective score. The details are described in the following sub-sections.

2.1. Deep Siamese CNN for visual difference features from left and right views

In the stereoscopic images, there are various visual differences such as binocular mismatches [13] and window violation [4] since the left and right views are captured from different viewpoints. In particular, excessive visual difference can cause severe visual discomfort and disturb the binocular fusion in stereoscopic viewing.

In this paper, to extract the visual difference feature inducing visual discomfort, a deep Siamese CNN is employed. Since the deep Siamese CNN consists of two identical sub-networks sharing the same parameters, it is very effective to simultaneously learn the desirable visual difference features from left and right views. Fig. 2 shows the proposed deep Siamese CNN for the visual difference features of stereoscopic images. To learn a visual difference feature highly correlated with visual comfort score, the visual difference feature is generated by calculating difference between outputs of each DCNN. Then, the visual comfort score is predicted from the visual difference feature using fully connected layers. Let f_{θ_1} and g_{ϕ_1} denote the transfer function of each DCNN with parameter θ_1 in the deep Siamese CNN and the regression function with parameter ϕ_1, respectively. As a result, by projecting the difference features of outputs in each DCNN for left and right views onto the subjective score space using g_{ϕ_1} , the visual comfort score is predicted. For extracting the desirable visual difference feature, the transfer function and regression function is learned by minimizing the mean absolute error (MAE) between predicted score from visual difference feature and true comfort score. The loss function can be written as

$$L_1 = \frac{1}{N}\sum_{i=1}^{N}\left|g_{\phi_1}(f_{\theta_1}(\mathrm{I}_i^L) - f_{\theta_1}(\mathrm{I}_i^R)) - s_i\right|, \qquad (1)$$

where $\mathbf{I}_i^L$ and $\mathbf{I}_i^R$ are i-th left and right view. s_i is i-th mean opinion score (MOS) obtained from subjects.

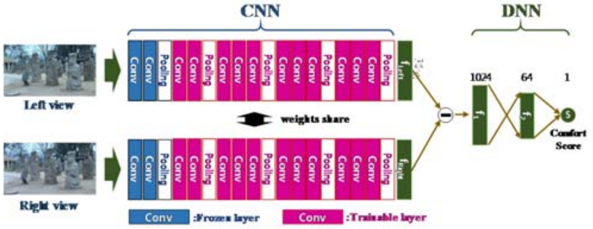

Fig. 2. Proposed deep Siamese CNN for learning visual difference.

2.2. DCNNs for disparity magnitude and gradient feature based on human attention region

It is well known that excessive disparity magnitude and disparity gradient is one of the major factors causing severe visual discomfort [5]. When these factors are in the human attention region, the visual discomfort can be more induced. Therefore, human attention region-based disparity magnitude and gradient maps are considered to extract latent disparity features, which are highly correlated with the visual discomfort. In our experiment, we employ the visual importance map in [5]. Let $\mathbf{S}_I$ denote image saliency map and $\mathbf{S}_D$ denote disparity saliency map. The visual importance map is given by $\mathbf{S}$=0.5$\mathbf{S}_I$ + 0.5$\mathbf{S}_D$. Fig. 4 (c) and (d) show image saliency map and depth saliency map from color and disparity maps. Fig. 4 (e) shows visual importance map. Based on the visual importance map, SAD and SADD [5] are obtained. The SAD is related with disparity magnitude based on visual importance map whereas the SADD refers differential disparity based on visual importance map. Let $\mathbf{D}_R$ denote disparity map and $\Delta\mathbf{D}_R$ denote differential disparity map, which is the gradient map of $\mathbf{D}_R$. Let $\mathbf{X}_{SAD}$ and $\mathbf{X}_{SADD}$ denote a SAD map and a SADD map, which can be written as

$$\mathbf{X}_{SAD} = \mathbf{S} \otimes |\mathbf{D}_R|, \tag{2}$$

$$\mathbf{X}_{SADD} = \mathbf{S} \otimes |\Delta\mathbf{D}_R|, \tag{3}$$

where $\otimes$ denotes element-wise multiplication.

The goal of the proposed DCNNs is to learn collaborative disparity features considering discomfort factors caused by absolute and relative disparities. Fig. 3 shows the proposed DCNNs for disparity features based on human attention model. As shown in Fig. 3, two different disparity features of SAD and SADD by each DCNN are trained together to regress to comfort score by fully connected layers. Suppose that $f_{\theta_2}(\cdot)$ and $f_{\theta_3}(\cdot)$ denote transfer functions of each DCNN. Let $g_{\phi_2}(\cdot)$ denote a regression function for disparity feature based on human attention region. By minimizing the loss between the predicted score from the combined disparity features and true comfort score, the transfer function and regression function for disparity magnitude and gradient features, are well trained. The loss function in this case can be written as

$$L_2 = \frac{1}{N}\sum_{i=1}^{N}\left|g_{\phi_1}(f_{\theta_2}(\mathrm{X}_{SAD}) + f_{\theta_3}(\mathrm{X}_{SADD})) - s_i\right|, \tag{4}$$

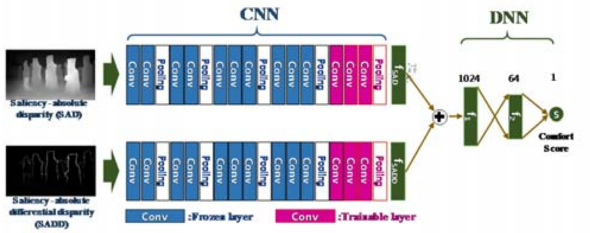

Fig. 3. Proposed DCNNs for learning human attention based disparity feature.

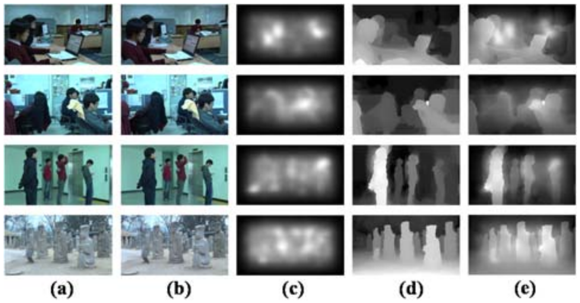

Fig. 4. Examples of visual importance map estimation. (a) Left view. (b) Right view. (c) Image saliency map. (d) Depth saliency map. (e) Visual importance map.

2.3. Regression for comfort score

In the last part, visual difference feature and human attention based disparity magnitude and gradient features are aggregated to predict visual comfort score more correctly using the combination of visual and disparity information. To that end, we aggregate the output of the deep Siamese CNN for visual difference features and output of the DCNNs for disparity features using linear combination. The transfer and regression functions learned in Section 2.1 and 2.2 are transferred as initial parameters in the CNN part. All CNN parameters in this section are frozen. Our network is fine-tuned in DNN part. As a result, the DNN directly regresses the aggregated feature to the MOS value by minimizing the MAE loss function as follows

$$L_3 = \frac{1}{N}\sum_{i=1}^{N}\left|g_{\phi_4}(\mathbf{T}_{diff} + \mathbf{T}_{SAD} + \mathbf{T}_{SADD}) - s_i\right|, \tag{5}$$

where $\mathbf{T}_{diff}$ denotes visual difference feature. $\mathbf{T}_{SAD}$ and $\mathbf{T}_{SADD}$ denote human attention-based disparity magnitude features and human attention-based disparity gradient features, respectively. g_{ϕ_4} denotes DNN transfer function.

3. EXPERIMENTS AND RESULTS

3.1. Database

In our experiment, we used IEEE-SA stereoscopic image database [6]. The database consists of 800 stereoscopic images with a spatial resolution of 1920x1080 pixels and the corresponding MOSs of visual comfort. These images consist

Table 1. Prediction performance on IEEE-SA Database

VCA metric for S3D	PLCC	SROCC	RMSE
Parallax distribution [17]	0.685	0.611	0.609
Stereoscopic impairments [18]	0.702	0.616	0.584
Statistical + fine features [6]	0.861	0.784	0.449
PUP [19]	0.860	0.780	0.420
Proposed Method	0.900	0.825	0.351

of 160 scenes with five separated convergence points.

3.2 Experimental setup

In our experiment, left and right view images (1920x1080) were resized to 384x216 to efficiently handle the memory and the computational load of the proposed deep network. Also, the image saliency map, depth saliency map, disparity map and differential disparity map were also resized to 384x216. Likewise, SAD and SADD were generated of size 384x216.

For performance evaluation, we used 10-fold cross-validation. For each validation set, the database was randomly subdivided into 90% of stereo pairs for training and 10% for prediction [14].

3.3 Learning the proposed network

In this paper, we used the IEEE-SA stereoscopic image data, which has the largest number of images among the existing stereoscopic VCA datasets. In particular, to prevent the over-fitting, we employed transfer learning from VGG-16 model [15]. The structure and parameters of the CNN in the VGG-16 model were transferred to Siamese CNN and DCNNs in the proposed method. In the CNN part, the lower convolution layers could capture ordinary features (e.g., color blobs and edges) [16]. The higher convolutional layers could capture the specific features for training purpose [16]. In the proposed method, the Siamese CNN was designed for learning visual difference of stereoscopic image and the DCNNs were designed for learning the characteristics of the disparity magnitude and gradient. Therefore, the parameters in higher layer were re-trained for learning the characteristics of the discomfort features in stereoscopic viewing, while parameters in lower layer were frozen for learning the basic characteristics of the images. In our experiment, convolution layers were frozen from the first to the 11-th in the Siamese CNN for visual difference features (see Fig. 2). In the DCNNs for disparity features, convolution layers are frozen from the first to the 10-th (see Fig. 3).

3.4 Performance comparisons

For performance evaluation, we employed three performance metric values (Pearson linear correlation coefficients (PLCC), Spearman rank order correlation coefficients (SROCC) and root mean square errors (RMSE)), which were measured between MOSs and predicted comfort scores. Table 1 shows performance comparing proposed method with others. We compared four previous methods. In [17], VCA metric was developed based on parallax distribution. In [18], by detecting stereoscopic impairments caused by inappropriate shooting parameters or camera misalignment, visual discomfort in S3D was predicted. The author of [6] used two kinds of features for VCA in S3D, which were statistical features computed from horizontal disparity map and fine features inducing neural activity in disparity perception and eye movement. In [19], for description of the presence of disparity, the percentage of un-linked pixels (PUP) was developed for VCA. As shown in Table 1, the proposed method achieved the highest correlation (PLCC: 0.900, SROCC: 0.825) and the lowest error (RMSE: 0.351). The results indicated that proposed deep perceptual features which are visual differences and human attention-based disparity features could reflect visual discomfort in stereoscopic viewing.

Table 2. Prediction performance for proposed method

Deep features	PLCC	SROCC	RMSE
Visual difference feature	0.712	0.677	0.748
Human attention based disparity features	0.896	0.823	0.356
Visual difference features + Human attention based disparity features	0.900	0.825	0.351

Further, we analyzed the correlation between aforementioned deep features and subjective score. Table 2 shows prediction performance results according to each feature in our network. When considering visual difference feature, PLCC was 0.712. When considering human attention based disparity features, PLCC was 0.896, which were higher than when considering the visual difference feature. The result indicated that the disparity feature more affected visual comfort in the proposed deep network. Finally, we aggregated all features so that the highest prediction performance could be achieved due to the complementary effect between features.

4. CONCLUSION

In this paper, we proposed a deep learning based novel stereoscopic VCA method. Through the proposed deep network, latent features from visual difference between left-right views and disparity features based on human attention region were encoded. The latent discomfort features and associate aggregation were represented through the deep network. To represent visual difference feature, deep Siamese CNN was revised. In addition, DCNN was utilized in order to represent human attention based disparity feature. Finally, by aggregating these deep features, the proposed method could predict visual comfort score. The experimental results showed that the proposed method achieved outperformed prediction performance.

5. ACKNOWLEDGEMENT

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (No.2017-0-00780, Development of VR sickness reduction technique for enhanced sensitivity broadcasting)

6. REFERENCES

[1] H. Sohn, Y. J. Jung, S. Lee, H. W. Park, and Y. M. Ro, “Attention model based visual comfort assessment for stereoscopic depth perception,” *in Proc. IEEE Int. Conf. Digital Signal Processing*, 2011, pp. 1–6.

[2] W. Tam, F. Speranza, S. Yano, K. Shimono, H. Ono, "Stereoscopic 3D-TV: Visual comfort", *IEEE Trans. Broadcast*., vol. 57, no. 2, pp. 335-346, Jun. 2011.

[3] M. Lambooij, W. Ijsselsteijn, M. Fortuin, I. Heynderickx, "Visual discomfort and visual fatigue of stereoscopic displays: A review", *J. Imag. Sci. Technol*., vol. 53, no. 3, pp. 1-14, Mar. 2009.

[4] Y. J. Jung, H. Sohn, S. Lee, and Y. M. Ro, “Visual comfort improvement in stereoscopic 3D displays using perceptually plausible assessment metric of visual comfort,” *IEEE Trans. Consum. Electron.,* vol. 60, no. 1, pp. 1–9, Apr. 2014

[5] Y. J. Jung, H. Sohn, S. Lee, H. W. Park, and Y.M. Ro, "Predicting Visual Discomfort of stereoscopic images using human attention model*," IEEE Trans. Circuits Syst. Video Technol*., vol. 23, no. 12, pp. 2077 – 2082, Dec. 2013

[6] J. Park, H. Oh, S. Lee and A. C. Bovik, “3D Visual Discomfort Predictor : Analysis of Disparity and Neural Activity Statistics,” *IEEE Trans. on Image Processing.*, vol. 24, no. 3, pp. 1101-1114, 2015.

[7] K. Ha and M. Kim, “A perceptual quality assessment metric using temporal complexity and disparity information for stereoscopic video,” *in Proc. IEEE Int. Conf. Image Process*., Sep. 2011, pp. 2525–2528.

[8] H. Sohn, Y. J. Jung, S. Lee, and Y. M. Ro, “Predicting visual discomfort using object size and disparity information in stereoscopic images,” *IEEE Trans. Broadcast.,* vol. 59, no. 1, pp. 28–37, Mar. 2013.

[9] M. T. M. Lambooij, W. A. IJsselsteijn, M. Fortuin, and I. Heynderickx, “Visual discomfort and visual fatigue of stereoscopic displays: A review,” *J. Imag. Sci. Technol.,* vol. 53, no. 3, pp. 030201–030201–14, Apr. 2009.

[10] K. Ha and M. Kim, “A perceptual quality assessment metric using temporal complexity and disparity information for stereoscopic video,” *in Proc. IEEE Int. Conf. Image Process*., Sep. 2011, pp. 2525–2528.

[11] H. Ying, G. Jiang, M. Yu, F. Shao, Z. Peng, and Y, Yang, “New stereo visual comfort assessment method based on scene mode classification,” *in Proc. IEEE Int. Conf. Quality of Multimedia Experience.*, 2015, pp.1-6.

[12] Q. Jiang, F. Shao, W. Lin, and G. Jiang, “On predicting visual comfort of stereoscopic images: A learning to rank based approach,*” IEEE Signal Process. Lett*., vol. 23, no. 2, pp. 302–306, Feb. 2016.

[13] J. Chen, J. Zhou, J. Sun, and A. C. Bovik, “Binocular mismatch induced by luminance discrepancies on stereoscopic images,” *in Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2014, pp. 1–6

[14] R. Huang, J. Zhou, X. Gu, Y. Zhang, and A. Bovik, “Comparison of regressors on 3D visual discomfort prediction,” *in Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast*., 2016 pp. 1-6.

[15] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv*:1409.1556, 2014.

[16] A. Abdulnabi, G. Wang, J. Lu, and K. Jia, “Multi-Task CNN Model for Attribute Prediction,” *IEEE trans. on Multimedia.*, vol. 17, no. 11, pp. 1949-1959, Nov. 2015.

[17] Y. Nojiri, H. Yamanoue, A. Hanazato, and F. Okano, “Measurement of parallax distribution and its application to the analysis of visual comfort for stereoscopic HDTV,” *Proc. SPIE*, vol. 5006, pp. 195–205, May 2003.

[18] D. Kim and K. Sohn, “Visual fatigue prediction for stereoscopic image,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 2, pp. 231–236, Feb. 2011.

[19] J. Chen, J. Zhou, J. Sun, and A. Bovik, "Visual discomfort prediction on stereoscopic 3D images without explicit disparities," *Signal Processing: Image Communication.*, vol 51, pp. 50-60, Jan 2017.