An Objective Assessment Method Based on Multi-level Factors for Panoramic Videos

Shu Yang [1], Junzhe Zhao [1], Tingting Jiang [2], Jing Wang [1*], Tariq Rahim [1], Bo Zhang [1], Zhaoji Xu [3], Zesong Fei [1]

[1] *School of Information and Electronics, Beijing Institute of Technology*
No.5 Zhongguancun South Street, Haidian District, Beijing 100081, China
[2] *School of Electronics Engineering and Computer Science, Peking University*
No.5 Yiheyuan Road, Haidian District, Beijing 100871, China
[3] *China Mobile Communications Corporation Research Institute*
No.32 Xuanwumen West Street, Xicheng District, Beijing 100053, China
[*] Corresponding author: wangjing@bit.edu.cn

***Abstract*—With the development of Virtual Reality (VR) technology, single-viewpoint videos have been replaced by the multi-viewpoint panoramic video owing to the fact that the latter brings people more immersive experiences. To improve the quality of experience (QoE) of panoramic videos, the video quality assessment (VQA) method need to be investigated. However the design of the quality metric for panoramic videos is a more complicated and harder problem, since users' feelings are affected by more psychological and physiological factors. Traditional VQA methods cannot evaluate the quality of panoramic videos accurately. In this paper, we propose a general objective full-reference quality assessment method for panoramic videos. The proposed method is based on multi-level quality factors, which are calculated with region of interest (ROI) maps. The framework is flexible and expandable, and its objective output has a higher correlation with subjective scores than that of traditional VQA methods and existing panoramic video evaluation methods.**

***Index Terms*—Virtual reality, panoramic video, video quality assessment, saliency of panoramic image, region of interest**

I. Introduction

Plenty of work has been done in video quality assessment (VQA)[1][2][3][4] for ordinary videos in the last two decades. However, the panoramic video in VR system is different from ordinary videos. Firstly, the displayer of panoramic videos is the VR Helmet-Mounted Display (HMD). The users are free to change the direction of their sight. At a certain moment, they only see a part of the panoramic frame sphere, which makes other invisible parts less important to the quality of experience (QoE). Particularly, these selected parts always contain viewers' region of interest (ROI), which makes ROIs in panoramic videos more significant to the quality assessment than those in ordinary videos. Secondly, in VR system, there are more complex factors influencing the choice of ROIs, while these ROIs make sense on different levels of visual features, such as the human visual system (HVS), the attention to some objects, and individual interests. As a result, the traditional saliency-based VQA methods may be not suitable for panoramic videos.

Several works have been published about panoramic content [5][6][7][8][9]. Zakharchenko et al. [7] proposed a metric for panoramic images using special zero area distortion projection method. The method considered differences of the viewing areas, but there were few experimental results to prove its validity. Evgeniy et al. [8] introduced a testbed for subjective panoramic images assessment and validated the data with existing objective metrics, such as s-PSNR[9]. S-PSNR is a weighted Peak Signal to Noise Ratio (PSNR) method considering users' viewing habits and reducing the effect of polar area. It is an objective VQA method which is suitable for panoramic videos. However, s-PSNR only considered the viewing difference in the vertical direction, ignoring the difference in the horizontal direction. Meanwhile it didn't provide any experimental results to evaluate its performance. This paper introduces a general VQA method for panoramic videos based on multi-level factors. We extend the idea of s-PSNR and take account of the viewing difference at different levels. Particularly, we evaluate the proposed method with a subjective quality assessment database.

In this paper, an objective full-reference quality assessment method is proposed to evaluate the quality of multi-viewpoint panoramic videos. First we calculate multi-level quality factors based on ROI maps. Then these factors are combined with a fusion model, which is learned from subjective quality assessment scores.

II. VQA with Multi-level Quality Factors

In this paper, a general full-reference objective quality assessment method is introduced for the evaluation of panoramic videos in VR system. The framework includes two stages, the calculation of multi-level quality factors and the multi-factor fusion model, as shown in Fig. 1. In the first stage, the input is the distorted video sequence $s = \{I_1, I_2, ..., I_n\}$ and the reference sequence $s' = \{I'_1, I'_2, ..., I'_n\}$. Firstly, we calculate

978-1-5386-0462-5/17/$31.00 ©2017 IEEE.

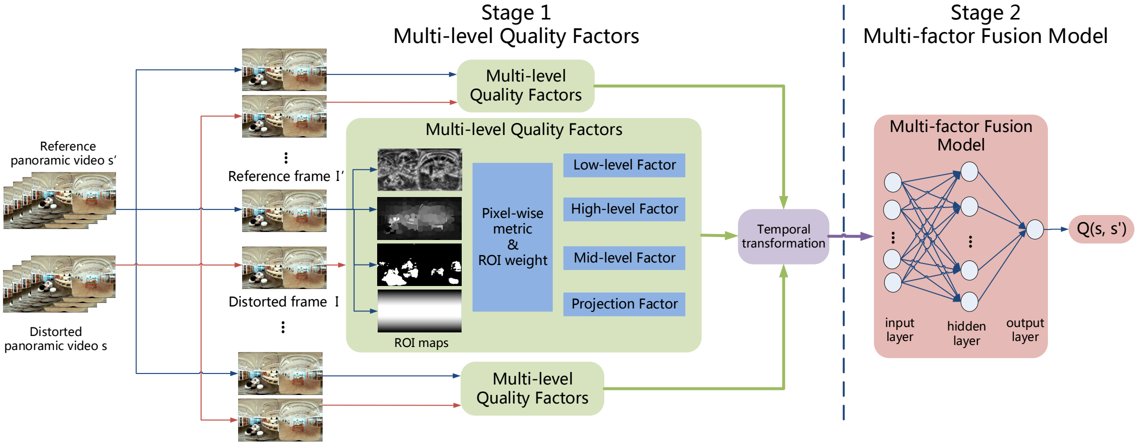

Stage 1
Multi-level Quality Factors
Stage 2
Multi-factor Fusion Model
Reference panoramic video s'
Distorted panoramic video s
Reference frame I'
Distorted frame I
Multi-level Quality Factors
Multi-level Quality Factors
Multi-level Quality Factors
Pixel-wise metric & ROI weight
Low-level Factor
High-level Factor
Mid-level Factor
Projection Factor
ROI maps
Temporal transformation
Multi-factor Fusion Model
input layer
hidden layer
output layer
Q(s, s')

Fig. 1. The proposed framework of full-reference objective quality assessment for panoramic videos.

quality factors of each frame pair (I_t, I'_t). We consider multi-level clues, including low-level, mid-level, high-level clues and projection, which are detailed in Section II.A. Then the quality factor set of the frame pair is defined as $f(I_t, I'_t)$ in Eq.(1), while l, m, h and p indicate low-level, mid-level, high-level and projection, respectively. Next we calculate quality factor set $F(s, s')$ for the sequence pair, as in Eq.(2). It is defined as a transformation of factor sets across all frames based on temporal clues. In the second stage, we build a multi-factor fusion model to combine these factors into one quality score $Q(s, s')$, the quality assessment of the panoramic video in VR system, as in Eq.(3).

$$f(I_t, I'_t) = \{f_l(I_t, I'_t), f_m(I_t, I'_t), f_h(I_t, I'_t), f_p(I_t, I'_t)\} \quad (1)$$

$$\begin{aligned} F(s, s') &= T(f(I_1, I'_1), f(I_2, I'_2), \cdots, f(I_n, I'_n)) \\ &= \{F_l(s, s'), F_m(s, s'), F_h(s, s'), F_p(s, s')\} \end{aligned} \quad (2)$$

$$Q(s, s') = C(F_l(s, s'), F_m(s, s'), F_h(s, s'), F_p(s, s')) \quad (3)$$

A. Multi-level quality factors

We consider quality factors on four different levels, including low-level (pixel), mid-level (region), high-level (object) and projection, as in Eq.(4). They are calculated with ROI weighted pixel-wise metrics. In this paper, the pixel-wise metric is defined by Eq.(5), which is similar to PSNR. The distance d_x between the distorted image I and the reference image I' is defined as the sum of pixel-wise Euclidean distance weighted by the ROI map, as in Eq.(6). W and H represent width and height of the panoramic image, while $I(i, j)$, $I'(i, j)$, and $M_x(i, j)$ are pixel values of the distorted image, the reference image and the ROI map respectively. We use different ROI methods to obtain ROI maps at different levels. There can be multiple maps and multiple quality factors in each level, and these ROI extraction methods are detailed in this section.

$$f_x(I, I') = D(d_x(I, I')) \quad x \in \{l, m, h, p\} \quad (4)$$

$$D(d) = 10 \times \log_{10}(\frac{255^2}{d}) \quad (5)$$

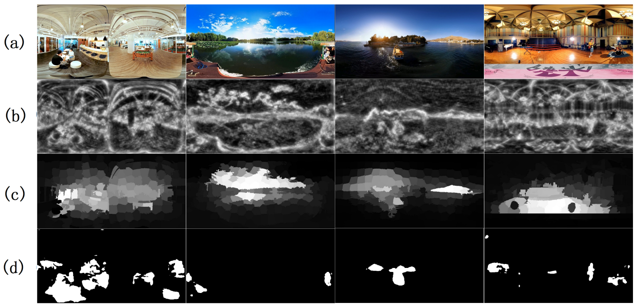


Fig. 2. Panoramic frames and their multi-level ROI maps. (a) The reference frames, (b) The low-level ROI maps, (c) The mid-level ROI maps, (d) The high-level ROI maps.

$$d_x(I, I') = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} [(I(i, j) - I'(i, j))^2 \times M_x(i, j)] \quad (6)$$

Pixel-level We consider HVS clues for low-level quality factors, obtaining saliency pixels with high color contrast or on edges. It is one of the factors to be considered when people evaluate the quality of both traditional videos and panoramic videos. We apply a non-parametric vision model[10] to obtain the low-level ROI map M_l shown in the second row of Fig. 2. M_l is a $W \times H$ dimensional vector taking continuous values in [0, 1], indicating the probability that the pixel in the panoramic image can be noticed. Then we obtain the low-level quality factor $f_l^1(I, I')$ with M_l based on Eq.(5). Particularly, the traditional PSNR method can be regarded as $f_l^2(I, I')$ by using a ‘full-one’ ROI map.

Region-level We consider mid-level information by using region-level features, such as superpixel saliency. By two bottom-up salient detection approaches proposed by Yang et al. [11], we compute two mid-level ROI maps, M_m^1 and M_m^2, and display M_m^1 in the third row of Fig. 2. The pixel value $M_m(i, j)$ is the saliency value of the superpixel which the pixel belongs to. We calculate two mid-level quality factors, $f_m^1(I, I')$ and $f_m^2(I, I')$, based on these two ROI maps.

Object-level Through subjective experiments, we find that

978-1-5386-0462-5/17/$31.00 ©2017 IEEE.

users are usually attracted to some objects, such as people, animals, cars and so on. We apply a semantic segmentation method FCN[12] to extract objects and their semantic information. We use the FCN model trained on PascalVOC dataset which is used in [12], to segment panoramic images. The segmentation method can produce 20 binary masks from one panoramic image according to 20 categories. In this paper, we only consider foreground and background, and we fuse these masks into one high-level ROI map M_h, as shown in the fourth row of Fig. 2. It is obvious that M_h highlights some meaningful areas, like people, boat and table, which usually attract users' attention. The pixel value of M_h is equal to 1 or 0, indicating the pixel of the panoramic image belongs to foreground or background. Then we calculate the high-level quality factor $f_h(I, I')$ with M_h.

Projection There are many projection formats of panoramic videos, such as sphere and cube. When VR users watch panoramic videos with VR HMD, they usually ignore the field of the head and feet, particularly polar area of the sphere or top and bottom area of the cube. Panoramic images with the same projection format obtain the same M_p. Then we calculate the projection quality factor $f_p(I, I')$ with M_p. S-PSNR is one method to calculate the projection quality factor with spherical format.

Considering the four level information above, we obtain six quality factors for each panoramic frame pair, including $f_l^1(I, I')$, $f_l^2(I, I')$, $f_m^1(I, I')$, $f_m^2(I, I')$, $f_h(I, I')$ and $f_p(I, I')$. Then we use these factors to calculate the sequence quality factor set with temporal transformation. Different from ordinary videos, panoramic videos are usually filmed in a gentle way. Given that quick movement of camera or the scene switch can lead to uncomfortable feelings for viewers, such as dizziness and nausea. We simply take the average of each factor across all frames and obtain 6 factors of each video sequence pair, including $F_l^1(s, s')$, $F_l^2(s, s')$, $F_m^1(s, s')$, $F_m^2(s, s')$, $F_h(s, s')$, $F_p(s, s')$, while they are the input for the next stage.

B. *Multi-factor fusion model*

We propose a fusion model to combine these multi-level quality factors into one quality score. Particularly, the input factor set can be regarded as a feature vector of the distorted video, and the output quality score can be regarded as the prediction. As a result, the fusion model can be learned with machine learning methods. In this paper, a back propagation (BP) neural network is designed to establish the fusion model. The neural network consists of three layers, including an input layer, a hidden layer and an output layer. The number of the input neurons is equal to the number of quality factors. The hidden layer contains 10 neurons according to experiments, while the output layer has only one neuron which obtains the quality assessment result for panoramic videos. We name our proposed method as BP-QAVR (BP-based quality assessment of panoramic videos in VR system).

III. EXPERIMENTAL RESULTS

A. *Database and subjective assessment*

We use a panoramic video dataset[13] for training and testing the proposed VQA method. The dataset contains 16 reference panoramic videos with 4k resolution, about concerts, news, sports, and so on. There are 384 distorted videos obtained by encoding reference videos with different bitrates, adding noise to and blurring reference videos. Each distorted video was evaluated with a subjective method in laboratory environment. As the spherical format is used in the dataset, we choose the method of s-PSNR to obtain the projection quality factors.

We divide the dataset into training set and testing set according to reference videos. The training set includes 12 reference videos and their corresponding distorted videos. The testing set include the other 4 reference videos and their corresponding distorted videos. To obtain reliable experimental results, we randomly divide the dataset 50 times. The average results are shown in the following section.

B. *Results and Comparison*

To measure the performance of the proposed method, we apply four widely used metrics, Spearman Rank Order Correlation Coefficient (SROCC), Kendall Rank Order Correlation Coefficient (KROCC), Pearson Linear Correlation Coefficient (PLCC) and Root Mean Square Error (RMSE). They are defined in [14], which can be used to evaluate the correlation between objective quality scores and subjective quality scores (DMOS). Please note that the objective results are mapped to DMOS space by 4-parameters logistic regression[14] for computing PLCC and RMSE.

The SROCC is used to study parameters for the neural network. We consider two parameters, including the number of hidden neurons, and quality factors of different levels. Table I shows SROCC results based on different number of hidden neurons. In this experiment, we use all 6 factors as the input. It is found that the network with 10 hidden neurons is the best. Table II shows SROCC results based on different factor sets. The factor sets stand for:'all factors'$= \{1, 2, 3, 4, 5, 6\}$, 'no low-level'$= \{3, 4, 5, 6\}$, 'no mid-level'$= \{1, 2, 5, 6\}$, 'no high-level'$= \{1, 2, 3, 4, 6\}$, 'no proj-level'$= \{1, 2, 3, 4, 5\}$. While 1, 2, 3, 4, 5 and 6 denote $F_l^1(s, s')$, $F_l^2(s, s')$, $F_m^1(s, s')$, $F_m^2(s, s')$, $F_h(s, s')$ and $F_p(s, s')$, respectively. It can be seen that the 'all factors' factor set performs best, and mid-level factors contribute more than other factors.

We compare the proposed method with the best BP model (based on 10 hidden neurons and all factors) with 4 full-reference VQA methods, including PSNR, s-PSNR[9], SSIM[1] and VQM[4]. Particularly, we also build a linear regression (LR) model to combine the quality factors by using the same train/test set as BP model, and obtain the objective quality score named LR-QAVR. Table III shows that the performance of LR-QAVR is better than PSNR and s-PSNR, indicating the multi-level factors method is superior to the single factor method. The BP-QAVR performance is better

TABLE I
SROCC of the testing set based on different number of hidden neurons.

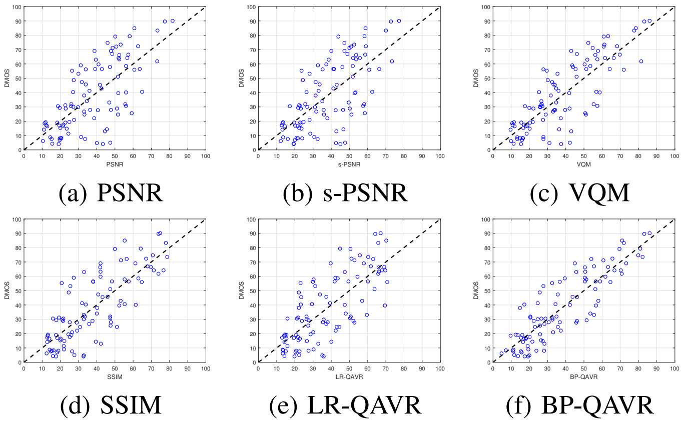
Number of hidden neurons	6	7	8	9	10	11	12
SROCC	0.8225	0.8323	0.8450	0.8349	**0.8485**	0.8394	0.8406

TABLE II
SROCC of the testing set based on different factor sets, in order to evaluate the contribution of these factors.

Factor set	all factors	no low-level	no mid-level	no high-level	no proj-level
SROCC	**0.8485**	0.8253	0.7958	0.8201	0.8314

TABLE III
Performance of the proposed method and other VQA methods.

method	SROCC↑	KROCC↑	PLCC↑	RMSE↓
PSNR	0.6813	0.5003	0.6937	16.0907
s-PSNR[9]	0.6576	0.4795	0.6712	16.5798
VQM[4]	0.7751	0.5907	0.8071	13.1753
SSIM[1]	0.7608	0.5697	0.7917	13.6422
LR-QAVR	0.7462	0.5494	0.7665	14.3543
BP-QAVR	**0.8485**	**0.6622**	**0.8622**	**11.1320**



(a) PSNR (b) s-PSNR (c) VQM

(d) SSIM (e) LR-QAVR (f) BP-QAVR

Fig. 3. Scatter plots of objective quality scores versus subjective DMOS after 4-parameters logistic regression. (a) PSNR, (b) s-PSNR, (c) VQM, (d) SSIM, (e) LR-QAVR, (f) BP-QAVR.

than LR-QAVR and VQM, showing that our BP fusion model is valid compared with LR and traditional multi-factor method. Fig. 3 is scatter plots between the objective and subjective scores on the test set. It can be seen that the data points of BP-QAVR are less spread out than those of other methods, indicating that the proposed method is suitable for panoramic video quality assessment.

IV. Conclusions and Future Work

In this paper, a general objective quality assessment framework for panoramic videos is proposed. The framework consists of multiple quality factors and a fusion model. Quality factors are obtained based on multi-level ROI analysis. Experimental results show that the proposed method is more suitable for panoramic videos. Certainly, there are some limitations in this paper, as different ROI extraction methods may affect the final evaluation result. We choose some ROI methods in this paper, and it is worth to study their influence on quality assessment in future work. However the proposed general framework is flexible. It is expected that its performance can be improved by selecting more effective ROI extraction methods, pixel-wise metrics and fusion models.

Acknowledgment

This work was partially supported by National Basic Research Program of China (973 Program) under contract 2015CB351803 and the Natural Science Foundation of China under contracts 61572042, 61390514.

References

[1] Z. Wang, L. Lu, and A. C. Bovik, “Video quality assessment based on structural distortion measurement,” *Signal Processing Image Communication*, vol. 19, no. 2, pp. 121–132, 2004.

[2] Z. Wang, E. P. Simoncelli, and A. C. Bovik, *Multi-Scale Structural Similarity for Image Quality Assessment*, 2004.

[3] G. H. Chen, C. L. Yang, and S. L. Xie, “Gradient-based structural similarity for image quality assessment,” in *IEEE International Conference on Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings*, 2006, pp. II–II.

[4] M. H. Pinson and S. Wolf, “A new standardized method for objectively measuring video quality,” *IEEE Transactions on Broadcasting*, vol. 50, no. 3, pp. 312–322, 2004.

[5] A. Dalvandi, B. E. Riecke, and T. Calvert, “Panoramic video techniques for improving presence in virtual environments,” in *Jvrc11: Joint Virtual Reality Conference of Egve - Eurovr, Nottingham, Uk, 2011. Proceedings*, 2011, pp. 103–110.

[6] R. X. Bijia Li, Li Song and N. Ling, “Evaluation of h. 265 and h. 264 for panoramas video under different map projections,” in *UMEDIA-2016:PROCEEDINGS OF THE 9TH IEEE INTERNATIONAL CONFERENCE ON UBI-MEDIA COMPUTING*, 2016, pp. 258–262.

[7] V. Zakharchenko, K. P. Choi, and J. H. Park, “Quality metric for spherical panoramic video,” in *SPIE Optical Engineering + Applications*, 2016, p. 99700C.

[8] M. R. Evgeniy Upenik and T. Ebrahimi, “A testbed for subjective evaluation of omnidirectional visual content,” in *32nd Picture Coding Symposium*, 2016, p. 221560.

[9] M. Yu, H. Lakshman, and B. Girod, “A framework to evaluate omnidirectional video coding schemes,” in *IEEE International Symposium on Mixed and Augmented Reality*, 2015, pp. 31–36.

[10] N. Murray, M. Vanrell, X. Otazu, and C. A. Parraga, “Saliency estimation using a non-parametric low-level vision model,” in *Computer Vision and Pattern Recognition*, 2011, pp. 433–440.

[11] C. Yang, L. Zhang, H. Lu, X. Ruan, and M. H. Yang, “Saliency detection via graph-based manifold ranking,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3166–3173.

[12] E. Shelhamer, J. Long, and T. Darrell, “Fully convolutional networks for semantic segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2017.

[13] B. Zhang, J. Zhao, S. Yang, Y. Zhang, J. Wang, and Z. Fei, “Subjective and objective quality assessment of panoramic videos in virtual reality environments,” in *Proc. IEEE International Conference on Multimedia and Expo (ICME’2017)*, Hong Kong, China, 2017.

[14] J. Antkowiak and T. J. Baina, “Final report from the video quality experts group on the validation of objective models of video quality assessment march,” *ITU-T Standards Contribution COM*, 2000.