No-Reference Video Quality Assessment based on Convolutional Neural Network and Human Temporal Behavior

Sewoong Ahn* and Sanghoon Lee*
* Yonsei University, Seoul, Korea
E-mail: anse3832, slee@yonsei.ac.kr Tel: +82-2-2123-7734

***Abstract*—The high performance video quality assessment (VQA) algorithm is a necessary skill to provide high quality video to viewers. However, since the nonlinear perception function between the distortion level of the video and the subjective quality score is not precisely defined, there are many limitations in accurately predicting the quality of the video. In this paper, we propose a deep learning scheme named Deep Blind Video Quality Assessment to achieve a more accurate and reliable video quality predictor by considering various spatial and temporal cues which have not been considered before. We used CNN to extract the spatial cues of each video in VQA and proposed new hand-crafted features for temporal cues. Performance experiments show that performance is better than other state-of-the-art no-reference (NR) VQA models and the introduction of hand-crafted temporal features is very efficient in VQA.**

I. Introduction

As the popularity of mobile devices and the demand for video streaming services increase, video services are performed in various dynamic network environments. Therefore, the final quality enjoyed by the viewer is different according to the channel environment even for the same contents [1][2]. To accurately evaluate the difference, there is a need to measure the quality of the video that the viewer perceives.

However, to date, most quality assessment (QA) studies have focused on image quality assessment (IQA). Recent IQA studies [3]-[7] have resulted in higher performance improvements compared to earlier studies [8]. In particular, J. Kim *et. al.* [9] achieved state-of-the-art performance in IQA through deep neural networks. In [9], it can be seen that various spatial cues such as saliency region, high / low spatial frequency, and natural scene statistics (NSS) are extracted through feature map visualization

On the other hand, due to the difficulty of analyzing visual characteristics, video has difficulty modeling nonlinear perception behaviors with a specific function different from image, and VQA method [10][11] is far below the IQA model. In conclusion, VQA has the lowest performance in QA field and VQA using only distorted video without reference one is a very challenging task.

In order to overcome the above-mentioned problems of making VQA difficult and the limits of existing VQA research, we proposed a new framework of no-reference (NR) VQA algorithm named deep blind video quality assessment (DeepVBQA). We use convolutional neural network (CNN) to improve the performance of NR VQA considering various temporal cues that were not considered in previous studies. We use pre-trained CNN model to extract various features. This introduces the concept of transfer learning, which is advantageous in that the number of training data does not cause overfitting. When the application of the model used in the transfer learning is similar to the original task, the performance of the original task increases. Therefore, NR IQA algorithm [9] similar to original task VQA was used as a model for transfer learning. However, image and video have obvious differences such as frame. Therefore, we extracted the features more suitable for VQA through fine-tuning in the model training process.

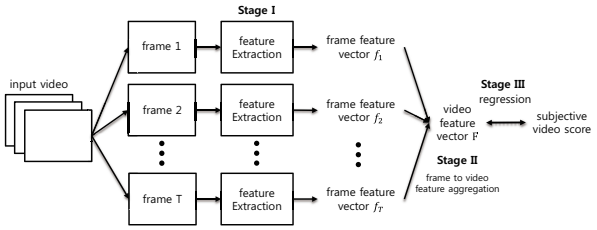

Fig. 1. Overall framework of DeepVQA.

Additional hand-crafted features were used to reduce CNN model complexity. The reason for the increasing depth of CNN model in computer vision field is to extract high level features in the feature extraction process. However, since the features that affect video quality in QA field have been clarified through previous studies, it is possible to add the features in a hand-crafted way so that the deep learning model reduces the efforts to extract existing features. In addition, various temporal cues were considered by introducing new features revealed through experiments.

The remainder of this paper is organized as follows. Section II discusses how to extract spatial features and temporal features at the frame level. And Section III describes a feature vector learning process that regresses to a final subjective quality score through a feature aggregation process. Section IV demonstrates the superiority of the proposed model through various experiments. Conclusions and future work are discussed in Section V.

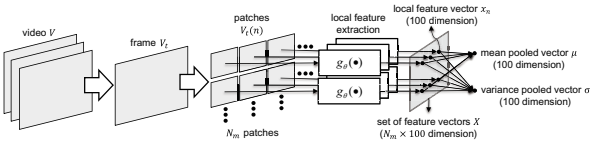

Fig. 2. Framework of extracting spatial features. Spatial features of video frame patches are extracted local feature extraction $g_\theta(\cdot)$ in [9].

II. Frame Feature Extraction

In the feature aggregation stage, a set of frame-level features extracted from the distorted video is combined into a video feature. Finally, at the regression stage, the video feature is trained to predict the subjective score. The proposed Deep-BVQA is shown in Fig. 1.

A. Spatial Features by Pre-trained CNN Model

We used our previous CNN model [9] trained for image quality assessment to extract the spatial feature from the video frame. In particular, patch-based learning is used to solve the problem of lack of database, which is a problem in existing IQA deep learning. The method used to extract spatial features for VQA is shown in Fig. 2.

If the video given in Fig.2 is V and t^{th} frame is V_t, then the video frame V_t is divided into patch units, and n^{th} patch can be called $V_t(n)$. For $V_t(n)$, each patch is passed through the local feature extraction part $g_\theta(\cdot)$ of Fig. 2 to extract the local feature vector x^n for n^{th} patch $(x^n = g_\theta(V_t(n)) = (x_1{}^n, x_2{}^n, ..., x_{100}{}^n))$. If there are N_m patches in t^{th} frame, we get set of feature vector $X = (x^1, x^2, ..., x^{N_m})$ through the corresponding patches.

In order to extract meaningful features from X, mean and variance were used in frame levels. Taking averages is often used to analyze representative characteristics of global quality in QA problems [8], and standard deviation was used to analyze the global variation of local quality [12]. Thus, the mean pooled vector $\mu = (\mu_1, \mu_2, ..., \mu_{100})$ and the variance pooled vector $\sigma = (\sigma_1, \sigma_2, ..., \sigma_{100})$ is derived as follows:

$$\mu_l = \frac{1}{N_m} \sum_{n=1}^{N_m} x_l{}^n, \tag{1}$$

$$\sigma_l = \frac{1}{N_m} \sum_{n=1}^{N_m} (x_l{}^n - \mu_l)^2 \tag{2}$$

where l is the pooled feature index (l=1,2,...100). Therefore, if 200 feature vectors are extracted from each frame V_t of the video (100 features from μ_l and 100 features from σ_l) and a video frame exists up to T, a $200 \times T$-dimensional spatial feature vector can be obtained from one video .

B. Temporal Features by Hand-crafted Methods

Our previous CNN model [9] is specialized to extract spatial features, so it is necessary to add temporal features for the algorithm to measure video quality. In this paper, we extract temporal sharpness variation to extract temporal features.

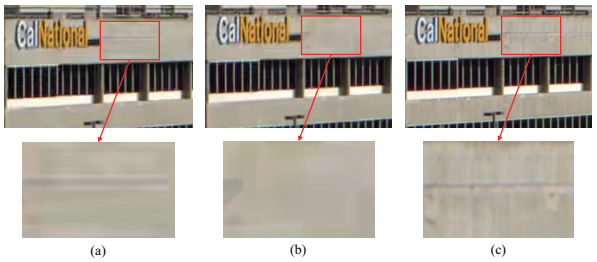

Fig. 3. Temporal variation of frame sharpness due to video distortion in "Pa" sequence of LIVE video database. (a) 10^{th} frame which has low distortion. (b) 11^{th} frame which has high distortion. (c) 12^{th} frame which has very low distortion.

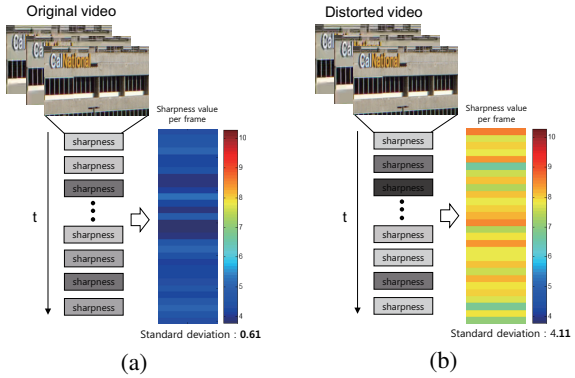

Fig. 4. Sharpness variation of video "BQMall" in CSIQ video database. Sharpness score is measured for each frame and standard deviation of sharpness score are calculated for (a) Original video, (b) Distorted video.

1) Temporal sharpness variation features: Experiments have shown that temporal variation of spatial cues has a significant effect on predicting picture quality. Fig. 3 shows the three consecutive frames of distorted video in the "BQMall" sequence of the CSIQ video database with mosquito noise at the top and an enlarged view of the local patch at the bottom shows the most noticeable changes in video. We can see that the (c) of the three frames is nearest to the original, and the distortion of (b) is the most severely distorted. By using H. Kim's algorithm, which measures the sharpness of video, the sharpness of (c) was measured to be the largest and (b) to be the most blurred. As the temporal variation of frame quality increases, the contrast of the time domain in the temporal CSF [13] increases, so the human recognizes the noise in the video. In conclusion, we use the temporal sharpness variation features because it reflects the human visual system characteristics of temporal variation CSF of frame sharpness.

If the t^{th} frame of video V is V_t, then the sharpness value obtained by applying the sharpness metric $S(\cdot)$ derived from our previous work [14] is $S(V_t)$.

Fig. 4 shows the change of the sharpness value per frame $S(V_t)$ for the original image and the distorted image. We can see that the sharpness variation is larger for the distorted frame. Therefore, we used the variance of the sharpness value per

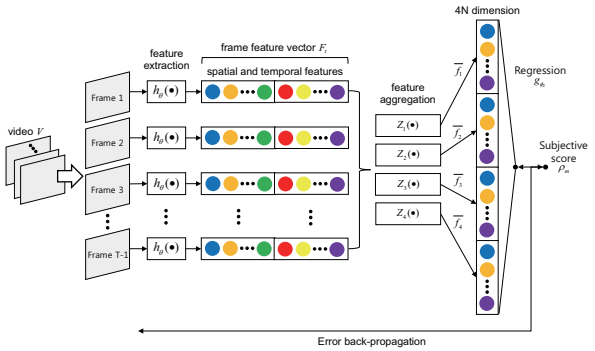

Fig. 5. Framework of frame to video feature aggregation. For each frame, N-dimensional feature vector is extracted. By feature aggregation, the final video feature vector has $4N$ dimension.

frame as a feature of temporal sharpness variation. If the value to be sought is $\sigma(S(V_t))$, it is obtained as:

$$\sigma(S(V_t)) = \frac{1}{T}\sum_{t=1}^{T}(S(V_t) - \mu)^2. \tag{3}$$

here, μ is obtained as the average of sharpness value $S(V_t)$ per frame.

III. Feature Vector Learning

A. Frame to Video Feature Aggregation

The proposed model is designed to extract spatial and temporal feature vectors within a frame through pre-trained CNN models and hand-crafted features. The extracted feature vector only considers adjacent frame information and does not reflect the overall tendency of the entire video frame. Therefore, as shown in Fig. 5, a feature vector of frame level requires an aggregation process to extract a video feature vector which represents a quality of a video.

The process is shown in Fig. 5. It begins by extracting spatial and temporal frame feature vectors for all frames of a video. And then, frame feature aggregation process $\mathbf{Z}(\cdot)$ follows the feature extraction. Therefore, the N-dimensional frame level feature vector extracted through the frame feature extraction function can be expressed as $F_t = \mathbf{h}_\theta(V_t, V_{t+1}) = (f_{t,1}, f_{t,2}, ..., f_{t,N})$. Assuming that an $\bar{\mathbf{f}}_1, \bar{\mathbf{f}}_2, \bar{\mathbf{f}}_3, \bar{\mathbf{f}}_4$ is an N-dimensional feature vector obtained through each of four pooling functions, each feature vector is obtained as follows:

$$\bar{\mathbf{f}}_1 = z_1(\mathbf{F}) = \frac{1}{T}\sum_{t=1}^{T} f_{t,n}, \tag{4}$$

$$\bar{\mathbf{f}}_2 = z_2(\mathbf{F}) = \frac{1}{T}\sum_{t=1}^{T}\left(f_{t,n} - \bar{f}_{1,n}\right)^2, \tag{5}$$

where $f_{t,l}$ denotes n^{th} component of the N-dimensional vector extracted from t^{th} frame. Equation (4) means average pooling, which is often used in pooling to obtain global quality in various VQA algorithms [8]. Equation (5) denotes the variance of frame quality, which reflects the change in frame quality degradation.

In addition, it is known that the quality of the frame with severe distortion in the video greatly affects the overall image quality of the video [15]. To reflect these characteristics, we averaged the upper and lower p^{th} percentiles from the frame feature histogram:

$$\bar{\mathbf{f}}_3 = z_3(\mathbf{F}) = \frac{1}{T^p}\sum_{t>t^{p+}} f^h_{t,n}, \tag{6}$$

$$\bar{\mathbf{f}}_4 = z_4(\mathbf{F}) = \frac{1}{T^p}\sum_{t<t^{p-}} f^h_{t,n} \tag{7}$$

where t^{p+} and t^{p-} indicate the upper and lower p^{th} percentiles in the histogram of frame features $f^h_{t,n}$, respectively. t^p represents the number of p-percentile local qualities, i.e., $t^p = t \cdot p/100$.

B. Regression onto Subjective Video Quality Score

When $4\times N$ dimensional video feature vectors are extracted from m^{th} video through frame to video feature aggregation, the calculated feature vectors are regressed to the corresponding subjective score ρ_m. The process results in finding a parameter that minimizes the loss function l_2:

$$\Theta_2^* = \arg\min_{\Theta_2} \ell_2\left(\{V\}, \hat{\rho}_m; \Theta_2\right), \tag{8}$$

where loss function $l_2(\cdot)$ means mean squared error between predicted video quality score and subjective score.

$$\ell_2\left(\left\{\hat{I}'_{lm}, \hat{I}'_{rm}\right\}, \hat{\rho}_m; \Theta_2 = (\theta, \phi_2)\right) \tag{9}$$
$$= \frac{1}{M_T}\sum_{m=1}^{M_T}\left(g_{\phi_2}\left(\mathbf{z}\left(\mathbf{h}_\theta\left(\{V\}\right)\right)\right) - \hat{\rho}_m\right)^2,$$

where M_t represents the number of videos used in training and $g_{\phi_2}(\cdot)$ represents the regression function with parameter ϕ_2. The computed loss $l_2(\cdot)$ updates the parameter (θ, ϕ_2) in the model through the back-propagation process.

TABLE I
PLCC and SROCC comparison on the LIVE video quality database.

	VQA model	LCC	SROCC
FR	PSNR	0.8158	0.7490
	SSIM [8]	0.8862	0.8884
	MOVIE [16]	0.9124	0.8813
RR	VQM [18]	0.7802	0.7858
	STRRED [17]	0.8917	0.9051
NR	V-CORNIA [19]	0.8534	0.8423
	V-BLIINDS [11]	0.8426	0.8267
	VIIDEO [10]	0.6918	0.6739
	Proposed	**0.8572**	**0.8513**

IV. Experimental Results

State-of-the-art metrics were utilized to compare DeepVQA against the performance of previous VQA models. Two well-known coeffieicnts were used for benchmark: the Pearson linear correlation coefficient (PLCC) and Spearman rank-order

TABLE II
PLCC AND SROCC COMPARISON ON THE CSIQ VIDEO QUALITY DATABASE.

	VQA model	LCC	SROCC
FR	PSNR	0.7932	0.7253
	SSIM [8]	0.8517	0.8661
	MOVIE [16]	0.8912	0.8750
RR	VQM [18]	0.7694	0.7698
	STRRED [17]	0.8734	0.8822
NR	V-CORNIA [19]	0.8315	0.8216
	V-BLIINDS [11]	0.8228	0.8069
	VIIDEO [10]	0.6704	0.6498
	Proposed	**0.8532**	**0.8472**

TABLE III
PLCC AND SROCC COMPARISON ON THE CSIQ VIDEO QUALITY DATABASE, WHERE THE PREDICTION MODEL WAS TRAINED USING THE LIVE VIDEO QUALITY DATABASE.

	VQA model	LCC	SROCC
FR	PSNR	0.7624	0.7028
	SSIM [8]	0.8307	0.8450
	MOVIE [16]	0.8691	0.8548
RR	VQM [18]	0.7382	0.7353
	STRRED [17]	0.8426	0.8524
NR	V-CORNIA [19]	0.8175	0.8083
	V-BLIINDS [11]	0.8024	0.7886
	VIIDEO [10]	0.6523	0.6312
	Proposed	**0.8471**	**0.8398**

correlation coefficient (SROCC). 80% of the databases were randomly selected and used for training, and the remaining 20% of the databases were used for testing. In addition, we doubled the size of the databases by horizontally reversing the videos. This is based on the assumption that the quality score that a person feels will be the same, even if the video is reversed, since the human eye is a symmetric structure. We compared the performances of the proposed model against those of the previous VQA models. Table I and II shows the PLCC ans SROCC of these VQA models on the LIVE and CSIQ video quality database. In terms of performance, DeepVQA shows higher performance than other NR VQA models. Also, our model shows similar performance compared to FR and RR metrics.

To demonstrate the generality of proposed model, we also conducted a cross-database evaluation. After learning a DeepVQA model using 80% of the training data from the LIVE VQA database, predicted quality scores were inferred on the CSIQ video database using the trained model parameters. Table III shows the LCC and SROCC of VQA models using this train-test sequence mentioned before. As shown in Table III, the overall performance of our model was significantly better than that of other NR VQA algorithms and shows similar performance compared to FR and RR metrics.

V. CONCLUSIONS

In this paper, we proposed a deep learning based approach to predict the quality of distorted videos without reference ones. We used a pre-trained CNN model to extract spatial features and hand-crafted features to extract temporal features. As a result, DeepVQA get a PLCC score which is higher than the other state-of-the-art VQA models. However, we manually calculated the hand-crafted temporal features and proceed to deterministic pooling in the process of aggregation. Therefore, it is not true deep learning. Therefore, we will study how to automatically extract temporal features from deep learning model and the method of adaptive temporal pooling according to the characteristics of video contents.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2016R1A2B2014525).

REFERENCES

[1] S. Kim and S. Lee, "Coordinated multicast based on MIMO relay station in a single frequency network," *IEEE Transactions on Vehicular Technology,* vol. 65, no. 2, pp. 685-698, 2016.

[2] H. Lee, B. Kwon, S. Kim, I. Lee and S. Lee, "Theoretical analysis based distributed load balancing over dynamic overlay clustering," *IEEE Transactions on Vehicular Technology,* vol. 65, no. 8, pp. 6532-6546, 2016.

[3] M. A. Saad, A. C. Bovik and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Transactions on Image Processing,* vol. 21, no. 8, pp. 3339-3352, 2012.

[4] J. Kim, T. Kim, S. Lee and A. C. Bovik, "Quality assessment of perceptual crosstalk on two-view auto-stereoscopic displays," *IEEE Transactions on Image Processing,* vol. 26, no. 10, pp. 4885-4899, 2017.

[5] H. Oh, S. Ahn, J. Kim and S. Lee, "Blind deep S3D image quality evaluation via local to global feature aggregation," *IEEE Transactions on Image Processing,* vol. 26, no. 10, pp. 4923-4935, 2017.

[6] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment of natural stereopairs," *IEEE Transactions on Image Processing,* vol. 21, no. 12, pp. 4695-4708, 2012.

[7] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zang and A. C. Bovik, "Deep convolutional neural models for picture quality prediction," *IEEE Signal Processing Magazine,* vol. 34, no. 6, pp. 103-141, 2017.

[8] Z. Wang, A. C. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing,* vol. 13, no. 4, pp. 600-612, 2004.

[9] J. Kim, and S. Lee, "Fully deep blind image quality predictor," *IEEE Journal of Selected Topics in Signal Processing,* vol. 11, no. 1, pp. 206-220, 2017.

[10] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity oracle," *IEEE Transactions on Image Processing,* vol. 25, no. 1, pp. 289-300, 2016.

[11] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind prediction of natural video quality," *IEEE Transactions on Image Processing,* vol. 23, no. 3, pp. 1352-1365, 2014.

[12] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Processing Letters,* vol. 20, no. 3, pp. 209-212, 2013.

[13] J. G. Robson, "Spatial and temporal contrast-sensitivity functions of the visual system," Journal of the Optical Society of America, vol. 58, no. 8, pp. 1141-1142, 1966.

[14] H. Kim, J. Kim, T. Oh and S. Lee, "Blind sharpness prediction for ultra-high-definition video based on human visual resolution," *IEEE Transcations on Circuits and Systems for Video Technology,* vol. 27, no. 5, 951-964, 2017.

[15] J. Park and S. Lee, "Video Quality Pooling Adaptive to Perceptual Distortion Severity," *IEEE Transactions on Image Processing,* vol. 22, no. 2, pp. 610-620, 2013.

[16] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Transactions on Image Processing,* vol. 19, no. 2, pp. 335-350, 2010.

[17] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Transactions on Circuit System and Video Technology,* vol. 23, no. 4, pp. 684-694, 2012.

[18] 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.
[19] J. Xu, P. Ye, Y. Liu and D. Doermann, "No-reference video quality assessment via featuer learning," *IEEE International Conference on Image Processing (ICIP),* 2014.