Received: 14 April 2020 | Revised: 5 August 2020 | Accepted: 11 November 2020

DOI: 10.4218/etrij.2020-0160

ORIGINAL ARTICLE

ETRI Journal WILEY

No-reference quality assessment of dynamic sports videos based on a spatiotemporal motion model

Hyoung-Gook Kim[1] | **Seung-Su Shin**[1] | **Sang-Wook Kim**[2] | **Gi Yong Lee**[1]

[1]Department of Electronics Convergence Engineering, Kwangwoon University, Seoul, Rep. of Korea

[2]Department of Software, Chung-Ang University, Seoul, Rep. of Korea

Correspondence
Hyoung-Gook Kim, Department of Electronics Convergence Engineering, Kwangwoon University, Seoul, Rep. of Korea.
Email: hkim@kw.ac.kr

Funding information
The work reported in this paper was conducted during the sabbatical year of Kwangwoon University in 2018. This study was supported by the BK21 Four project (Wellness Care Fusion Technology Based on Hyper-Connected Human Experiences) funded by the Ministry of Education, Department of Electronics, Convergence Engineering, Kwangwoon University, Rep. of Korea (F20YY8101058).

This paper proposes an approach to improve the performance of no-reference video quality assessment for sports videos with dynamic motion scenes using an efficient spatiotemporal model. In the proposed method, we divide the video sequences into video blocks and apply a 3D shearlet transform that can efficiently extract primary spatiotemporal features to capture dynamic natural motion scene statistics from the incoming video blocks. The concatenation of a deep residual bidirectional gated recurrent neural network and logistic regression is used to learn the spatiotemporal correlation more robustly and predict the perceptual quality score. In addition, conditional video block-wise constraints are incorporated into the objective function to improve quality estimation performance for the entire video. The experimental results show that the proposed method extracts spatiotemporal motion information more effectively and predicts the video quality with higher accuracy than the conventional no-reference video quality assessment methods.

KEYWORDS
3D shearlet transform, conditional constraints, deep residual bidirectional gated recurrent neural network, natural scene statistics, no-reference video quality assessment

1 | INTRODUCTION

The rapid development of digital devices, multimedia services, and digital networking technologies has made it easier to create and transmit videos from many different devices and share them on social media platforms, resulting in an explosion of video content for online viewing. However, after a digital video is acquired through a capturing device, the video is accompanied by distortion while being processed, compressed, stored, transmitted, or reproduced. This distorted video must be reproduced accurately within the range of human perception. Additionally, it is imperative for a video service system to realize and quantify the video quality degradations that occur in the system, so that it can maintain, control, and possibly enhance the quality of the video data. To this end, video quality assessment (VQA) methods have been recently proposed, attracting considerable research attention.

VQA methods can be divided into two classes: subjective and objective VQA [1]. In the subjective VQA method, a user's satisfaction can be accurately measured, because the experiment is performed on an actual viewer to measure the video quality perceived and accepted by the viewer. However, performing experiments involving multiple viewers is time-consuming and costly. Moreover, it is difficult to apply the results to real-time in-service quality evaluations. To overcome the disadvantages of subjective VQA, the objective VQA method has been proposed. The

goal of objective VQA is to design mathematical models that can accurately and automatically predict the quality of videos. An ideal objective VQA method should be able to mimic the quality predictions of an average human observer. Depending on the availability of a reference video, in which case there is no distortion and the quality is considered perfect, objective VQA methods can be classified into three categories: full-reference (FR)-VQA [2], reduced-reference (RR)-VQA [3], and no-reference (NR)-VQA [4]. Among these methods, the NR-VQA method can be used in various applications, such as real-time VQA and automated quality control [5], because it estimates the video quality by accessing only the distorted video without requiring any information of the reference video. Because of this advantage, the NR-VQA method has been actively studied.

A video is a set of consecutive frames that contain various motion properties. Not only videos from other genres with little motion activity, but also sports videos with dynamic motion activities have substantial natural spatiotemporal correlations. Therefore, the perception of motion information plays an important role in the perception of videos including various motion properties, and this motion information can be extracted by a system that responds to the oriented spatiotemporal energy [6]. When distortions are introduced into natural videos, the property of the distorted video will become different from that of the natural video; the property of a distorted video deviates from that predicted by natural scene statistics models [7]. These deviations can be applied as an indicator of video quality; however, only a few existing VQA algorithms explicitly detect motion characteristics and directly use motion information. Seshadrinathan and Bovik [8] presented an FR-VQA method to evaluate dynamic video fidelity by integrating both spatial and temporal aspects of distortion assessment. Their method delivered scores that correlate quite closely with human subjective judgment. Saad and Bovik [9] introduced a motion-aware NR-VQA method using machine learning. Their approach relies on a natural scene statistics model of video scenes in the discrete cosine transform domain as well as a temporal model of motion coherency occurring in the scene.

Recently, there have been attempts to learn visual motion sensitivity using deep learning [10] for NR-VQA. In NR-VQA using deep learning, Li and others [11] proposed a simple and efficient method to extract mean spatiotemporal features from video blocks. In their work, the mean spatiotemporal features were simply extracted using a three-dimensional shearlet transform (3DST), which can efficiently capture anisotropic features in multidimensional data. To train a statistical regression model on the mean spatiotemporal features for each quality label class, a convolutional neural network (CNN) is applied. The performance of this method is similar to those of current state-of-the-art FR-VQA methods and general-purpose NR-VQA algorithms. However, this method does not reflect the fact that video distortion caused by fluctuations in network conditions may not be the same in every frame. To improve quality evaluation, a method capable of predicting video quality in a frame or block unit from a given video quality label for regression learning and reflecting it in the quality evaluation of the entire video is required. To predict the quality score frame-by-frame, Xu and others [12] proposed an NR-VQA method via feature learning. In their study, frame-level features were first extracted by unsupervised feature learning and used to train a linear support vector regression (SVR) model. The final score of a single video was obtained by combining the frame-level scores using temporal pooling. For non-intrusive speech quality assessment, Quality-Net [13] was designed to automatically learn reasonable frame-level quality, even if the quality label in the training data is utterance-wise.

Motivated by previous results, we present an NR-VQA for dynamic sports videos based on a spatiotemporal motion model.

The main contributions of this study are as follows: (a) We apply 3DST to efficiently extract spatiotemporal features, which capture the dynamic motion scene statistics from the incoming video blocks and use them to estimate the quality of each video block. (b) We use a concatenation of a deep residual bidirectional gated recurrent neural network (DRBGRNN) and logistic regression to learn the spatiotemporal correlation more robustly for video sequences in detail and predict the improved perceptual quality score. (c) We use a conditional constraint method in the video block quality assessment to automatically learn reasonable block-level quality from distorted video blocks. Subsequently, we obtain an improved video quality score by combining the block-wise scores through a global average.

The rest of this paper is organized as follows. Section 2 describes the proposed method, Section 3 presents the experimental results, and Section 4 provides conclusions.

2 | PROPOSED NO-REFERENCE QUALITY ASSESSMENT OF SPORTS VIDEOS

In this study, we use a regression model-based no-reference video quality metric that is trained to learn the mapping relationship between the specific feature and subjective quality scores and predict the quality score in the testing phase.

Figure 1 shows the process flow of the proposed NR-VQA approach, which is composed of a training stage and a test stage.

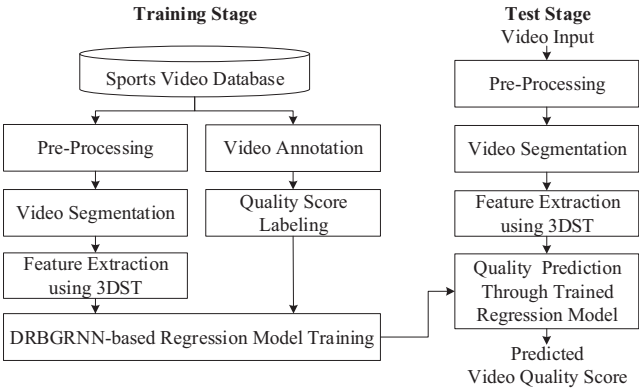

FIGURE 1 Flowchart of the proposed NR-VQA system

First, in the training phase, each video in the training video database contains a video signal and a corresponding subjective quality rating score. The measured subjective quality score is labeled as the target quality score of the corresponding video to learn the mapping relationship between the specific feature and the subjective quality scores.

In the pre-processing stage, the consecutive video frames of each video are converted to grayscale and resized to a fixed size suitable for video quality analysis.

Second, each video of the video database is split into several blocks to achieve a multidimensional and multidirectional analysis. From each segmented multidimensional video block, we efficiently extract shearlet-based spatiotemporal features with anisotropic properties using 3DST.

Third, the spatiotemporal features extracted from the video blocks are concatenated and used for training the DRBGRNN. During the training, the DRBGRNN automatically learns the reasonable block-level quality from distorted video blocks using a conditional constraint on video block quality assessment. With this concept, a mapping relationship between the specific features and subjective quality scores can be reasonably learned in block units.

In the test stage, we perform NR-VQA by inputting the shearlet-based spatiotemporal features from the test videos into the concatenation of logistic regression and trained DRBGRNN.

2.1 | Spatiotemporal feature extraction based on the 3D shearlet transform

Sports videos contain dynamic natural scenes with a high degree of motion activities. As shown in Figure 2, in sports videos, such as from soccer or basketball events, the video scenes are largely made up of dynamic and natural motions exhibited by numerous players. The human vision system actively seeks salient regions and movements in video sequences. Indeed, if a moving target or player is observed in one frame of a video sequence, it is highly likely that it will also appear in the next

FIGURE 2 Examples of soccer and basketball videos

frame, so dynamic natural information is an important factor. Since the object of the video changes its position over time, dynamic characteristics including the movement of the object should be considered in addition to static characteristics such as the edges of the object. To deal with the data in both space and time, a spatiotemporal feature extraction technique is applied. In particular, 3DST has the advantage of extracting the apparent movement of objects, surfaces, and edges in a visual scene containing dynamic motions exhibited by numerous players. For this reason, we apply 3DST to efficiently extract spatiotemporal features, capturing dynamic motion scene statistics from incoming video blocks.

The spatiotemporal feature extraction process is divided into four steps: splitting of a video into blocks, calculation of 3D shearlet coefficients based on 3DST, mean pooling of shearlet coefficients, and normalization. First, the consecutive video frames of each video are converted to grayscale, resized to 624 × 360, and cropped into a 416 × 240 center patch owing to the fixed input size. Thereafter, each video is split into several video blocks, as shown in Figure 3. Each video block has a size of 416 × 240 × 80 (horizontal × vertical × temporal). A 50% overlap is used while dividing the video. The size of this block is selected through experiments for video quality evaluation that include various distortions. At this time, the blur effect/noise in the spatial area of the original video was not ignored or underestimated.

Second, 3DST is applied to each video block. The 3DST is a directional transform approach derived from the theory of shearlets. A 3D shearlet considers the motion feature in the temporal axis, such that the decomposed coefficients can sufficiently approximate the spatiotemporal features of video sequences in different scales. Our 3DST can be efficiently computed by the discrete Fourier transform. The Fourier-frequency space of the video block is decomposed

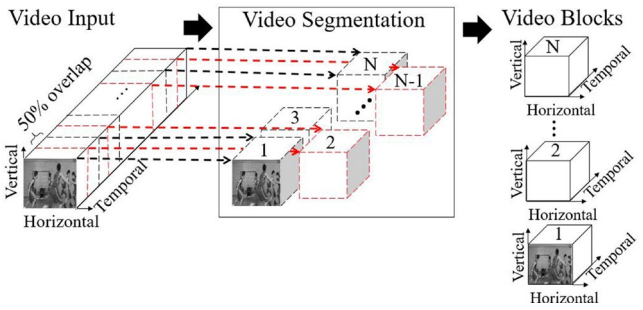

FIGURE 3 Splitting of a video into blocks

into four different subfamilies. The first family takes care of the low-frequency cube Φ. The other three families are associated with the high-frequency domain. Each of them corresponds to a pyramid whose symmetry axis is one of any cartesian axes ξ_1, ξ_2, and ξ_3 in the Fourier-frequency domain, as shown in Figure 4. Thus, three pyramidal regions P_1, P_2, and P_3 are obtained by partitioning the high-frequency space. In other words, the 3DST is constructed by a shearlet system associated with the pyramidal regions, as shown in Figure 3.

The 3DST of a video block is given by.

$$\begin{aligned} &\mathrm{SH}(\phi, p_1, p_2, p_3; \alpha, c) \\ &= \Phi(\phi; c_1) \cup P_1(p_1; \alpha, c) \cup P_2(p_2; \alpha, c) \cup P_3(p_3; \alpha, c), \end{aligned} \tag{1}$$

where Φ denotes the low-pass components; P_1, P_2, and P_3 are the first, second, and third pyramid regions from the high-pass components depending on the direction, respectively; α and c are an anisotropy parameter and a positional element, respectively; ϕ denotes a scaling function associated with the low-frequency cube; and p_1, p_2, and p_3 denote shearlets associated with three pyramidal regions.

Third, 3D shearlet filters at a particular scale, direction, and time are constructed by applying a combination of multi-scale decomposition and a direction filtering step to each of the three pyramidal regions, as depicted in Figure 4. The low-frequency cubes are excluded from the calculation of the 3D shearlet filter. The multi-scale decomposition is first implemented using the Laplacian pyramid algorithm. The directional components are then obtained using shearing matrices to control the orientations in the pseudospherical domain.

The 3D shearlet filter for the first pyramid region is defined as.

$$\begin{aligned} P_1(p_1; \alpha, c) = \Big\{ & p_{1j,k,m} = 2^{\frac{\alpha_j + 1}{4} j} p_1(\mathbf{S}_k \mathbf{A}_{\alpha, 2^j} \cdot - \mathbf{M}_c m) \\ & : j \geq 0, |k| \leq \left\lceil 2^{\frac{j(\alpha_j - 1)}{2}} \right\rceil, m \in Z^2 \Big\}, \end{aligned} \tag{2}$$

using

$$\mathbf{A}_{\alpha, 2^j} = \begin{pmatrix} 2^j & 0 & 0 \\ 0 & 2^{\alpha j/2} & 0 \\ 0 & 0 & 2^{\alpha j/2} \end{pmatrix}, \quad \mathbf{S}_k = \begin{pmatrix} 1 & k_1 & k_2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{3}$$

$$\mathbf{M}_c = diag(c_1, c_2, c_2),$$

where j, k, and m denote the scale, shearing, and translating parameters, respectively; $\mathbf{A}_{\alpha, 2}{}^j$, $\mathbf{S}_k$, and $\mathbf{M}_c$ represent the scaling, shearing, and translating matrices.

We decompose each pyramidal region into four scales. The smaller the scale parameter j, the lower the resolution, so only clear contour information, such as edges, exists. On the contrary, the greater the j value, the higher the resolution, so it contains even more detailed information. Nine directionalities are applied to each scale region. Each direction proceeds from one side of the pyramid to the opposite side through the center of the pyramid, as shown in Figure 5. Therefore, the first pyramid region outputs 36 shearlet coefficients by analyzing the video block through four scales and nine directionalities. The second and third pyramidal regions of the video block are also decomposed into four scales and nine directionalities. However, some regions are overlapped among the three pyramids. Therefore, we omitted certain shearlets lying on the borders of the second and third pyramids, thus extracting 52 3D shearlet filters with respect to the multiple scales and multiple directions from the input video block through 3DST.

Fourth, we perform pointwise multiplication (denoted by .*) of the Fourier-frequency coefficients with each 3D shearlet filter defined in the pyramid region in the frequency domain. The outputs are 3D fused coefficients. Inverse 3DST is applied on the fused coefficients to reconstruct the image sequence.

Figure 6 shows the process of obtaining the 3D vector of the shearlet coefficients through one 3D shearlet filter ($j = 1$, first directional filer in Figure 5) defined in the first pyramid region.

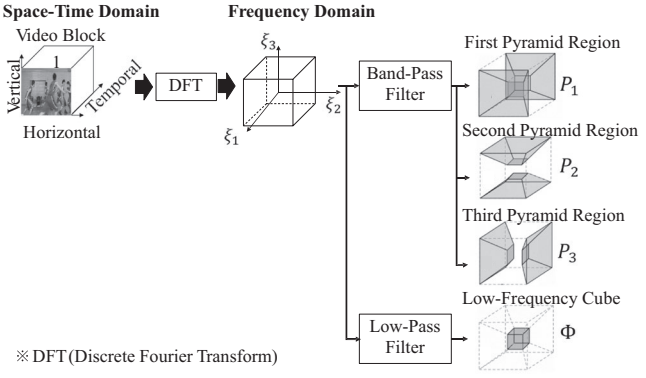

FIGURE 4 Partition of Fourier-frequency space

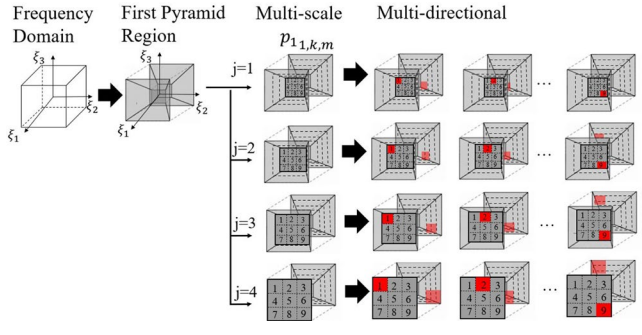

FIGURE 5 Tiling of the frequency domain induced by 3D shearlets inside the first pyramidal region

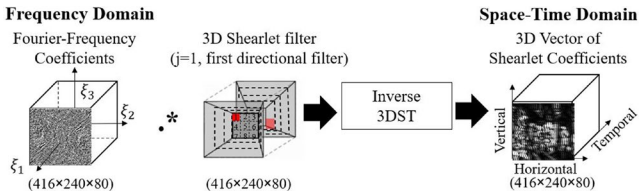

FIGURE 6 Image sequence reconstruction through the inverse 3D shearlet transform

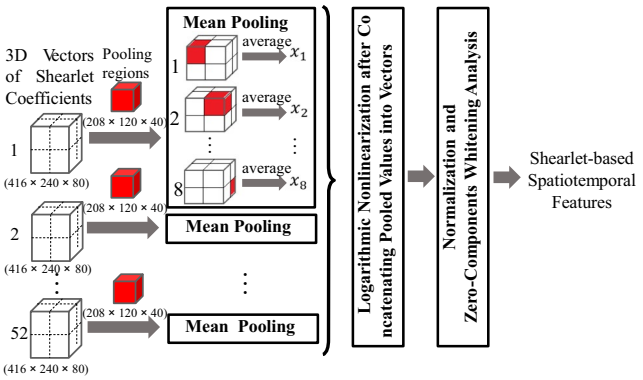

FIGURE 7 Shearlet-based spatiotemporal feature extraction process from the 3D vector of shearlet coefficients

Fifth, the average pooling is applied to each 3D vector of the shearlet coefficients at the particular scale, direction, and time. Each pooling region has a size of 208 × 120 × 40 and is marked with a red block in Figure 7. Eight pooled values are outputted from one 3D shearlet coefficient through the average pooling. The same procedure is applied to the 52 shearlet coefficients, outputting 416 pooled values. After the mean pooling of the 3D vector of the shearlet coefficients in each pooling region, the pooled values are concatenated as a vector, and every element in this vector is subject to a logarithmic nonlinearity.

Finally, to obtain the spatiotemporal motion features, the logarithmic vectors are normalized by subtracting the mean and dividing by the standard deviation of its elements.

2.2 | Logistic regression concatenated with DRBGRNN

In the evaluation of the quality of sports video composed of dynamic scenes, the human vision system is concentrated on the movement of players, making it essential to consider spatiotemporal motion information for a successful NR-VQA. To this end, as the first step in the main phase, we use 3DST to efficiently extract the spatiotemporal features. The second step is to incorporate the extracted features into the regression framework to make an appropriate NR-VQA. Several deep learning methods have been investigated for NR-VQA, for example, CNNs [14], recurrent neural networks (RNNs) [15], long short-term memory (LSTM) [16], and residual neural networks (ResNet) [17]. Among the various methods, the general-purpose NR-VQA method named SACONVA motivates us in this study. In this approach, the primary mean spatiotemporal features are extracted by 3DST, and CNN and logistic regression are concatenated to exaggerate the discriminative parts of the primary features and predict a perceptual quality score. Experimental results have shown that SACONVA strongly correlates with human perception and is very similar to state-of-the-art NR-VQA methods. To improve the performance of NR-VQA for sports videos with dynamic motion scenes using spatiotemporal features, we proposed a concatenation of logistic regression and DRBGRNN. The proposed DRBGRNN has a structure that combines gated recurrent neural networks (GRNNs) [18] and a residual framework to learn long-term dependencies and ensure the validity of information transmission through bidirectional cells and residual connections.

In Figure 8, the DRBGRNN is depicted as an unfolded form, in which information flows in the horizontal direction (temporal dimension) as well as in the vertical direction (depth dimension). Excluding the input and output layers, there are 13 residual layers with residual connections inside. Moreover, each residual layer consists of two bidirectional layers, three activity functions, and batch normalization (BN)

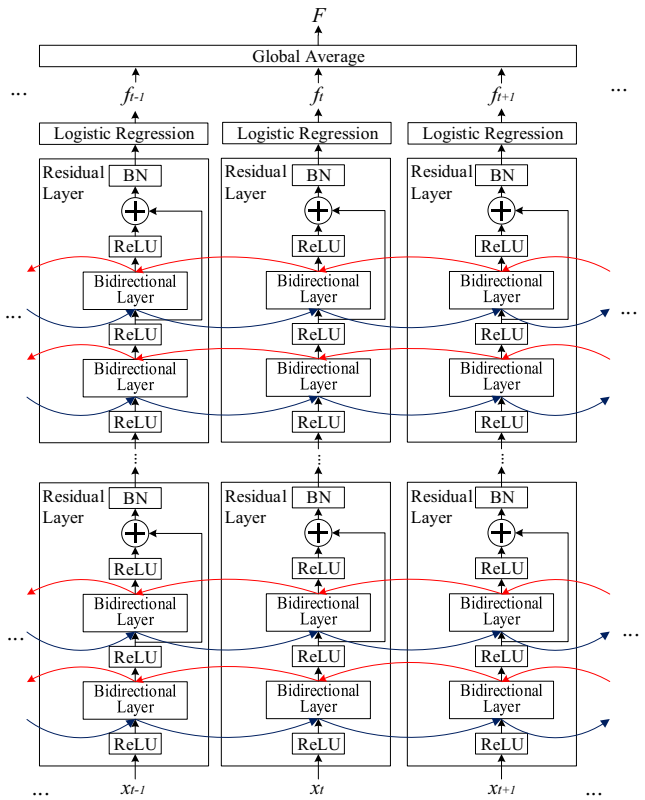

FIGURE 8 Block diagram of logistic regression concatenated with an unfolded DRBGRNN

step. Further, each bidirectional layer contains two gated recurrent unit (GRU) [18] cells, and the output is the result of their concatenation.

The GRNN maintains the solution of the LSTM to the long-term dependency problem while simplifying the structure of the complex LSTM. In other words, the information of the previous input value is continuously stored in the GRNN structure and is outputted and reflected at the required time. In addition, GRU can train faster and requires less amount of data to generalize, all while using only two gates (an update gate and a reset gate), whereas the LSTM uses three gates to control the information flow of the internal cell unit. Here, the update gate decides the amount of previous memory that should be retained, while the reset gate controls the combining mode of new input with the previous value. With these GRNNs, the temporal correlation can be learned from dynamic motion scenes.

The design of the bidirectional GRNN reflects the additional considerations required for GRNNs. It extracts powerful representations that preserve past and future information by taking full advantage of the forward and backward passes from time-series dynamic video sequences.

Figure 9 shows the bidirectional GRNN. In our bidirectional architecture, the forward GRUs are calculated by past states along the positive time axis while the backward GRUs are computed by future states along the reverse time axis. The output of the bidirectional layer is represented by the following equation from $t = 1$ to T:

$$h_t = \text{Concat}\left[\vec{h}_t, \overleftarrow{h}_t\right], \tag{4}$$

using

$$\vec{h}_t = g\left(\mathbf{U}_{\vec{h}} x_t + \mathbf{W}_{\vec{h}} \vec{h}_{t-1} + b_{\vec{h}}\right), \tag{5}$$

$$\overleftarrow{h}_t = g\left(\mathbf{U}_{\overleftarrow{h}} x_t + \mathbf{W}_{\overleftarrow{h}} \overleftarrow{h}_{t-1} + b_{\overleftarrow{h}}\right), \tag{6}$$

where $\vec{h}$ and $\overleftarrow{h}$ represent the forward sequences through the left-GRU layer and backward sequences through the right-GRU layer, respectively; x_t, $\mathbf{U}$, $\mathbf{W}$, b, and g denote the current input, weight matrix from the input layer to the hidden layer, weight matrix from the hidden layer to the hidden layer, bias vector, and hidden layer function, respectively.

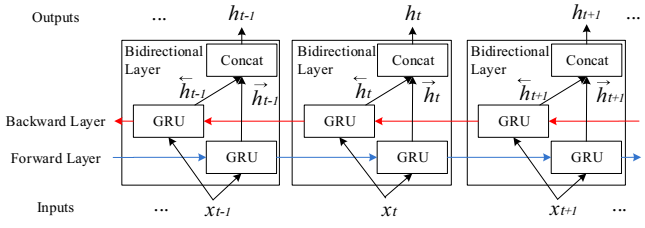

FIGURE 9 Unfolded architecture of bidirectional GRNN with three consecutive steps

For the design of the DRBGRNN, the bidirectional GRNN is combined with a residual learning framework, which is designed with two main considerations: (a) The residual learning framework learns the differences between the input and output, making it easy to optimize the deep neural network. It will be sensitive to small changes in the input values; (b) The residual connections over the bidirectional layer help significantly improve the training speed by restraining the gradient vanishing and exploding problem in deep networks. With the above considerations, the proposed DRBGRNN learns the spatiotemporal correlation more robustly from video sequences based on the residual framework. The output layer of the DRBGRNN is then concatenated with logistic regression for quality estimation tasks.

2.3 | Conditional constraint on block quality assessment

A digital video consists of consecutive frames. After it is acquired through a capture device, it is accompanied by distortion while being processed, compressed, stored, transmitted, or reproduced. Sometimes, such a distortion is not stationary, or the degree of video distortion is not the same across frames. In this case, it is not appropriate to directly assign the video-level quality label to every individual frame within the input video for the VQA. Just as how human beings can perceive the distortion of the video quality of each frame or scene and readily evaluate the video quality without any reference information, a method that can mimic the human visual perception process is required for VQA.

The proposed regression-based NR-VQA method aims at achieving the objective VQA using only test videos for estimating the perceptual video quality without making any reference to the original video. To this end, the regression model is trained to learn the mapping relationship between video samples and subjective quality scores and to generate its prediction-based quality metric. However, the subjective quality evaluation scores of the videos given for generating the regression model are only overall evaluation scores for one video and do not include the detailed quality evaluation scores for each video frame or each video block. In other words, the intermediate evaluation process is not reflected in the generated regression model. If the NR-VQA method is designed with a structure that reflects the intermediate evaluation process, we believe that a more accurate and effective evaluation can be achieved. In such a structure, if noise or video distortion occurs in one video block, its quality score should be decreased accordingly. Subsequently, the final estimated video-level quality score is then obtained by combining the block-wise scores through a global average. Based on

this concept, we incorporate a conditional block-wise constraint in the objective function of logistic regression concatenated with DRBGRNN.

Figure 10 shows the overall block diagram of the proposed conditional constraint on the video block quality assessment.

First, the shearlet-based spatiotemporal features extracted from one video block are inputted to the DRBGRNN with a quality label. Here, the quality label is the true quality score of the input video.

Second, the weighting factor of the objective function is calculated using (7):

$$\alpha(Q) = 10^{(Q - Q_{\mathrm{MAX}})}, \tag{7}$$

where Q and Q_{MAX} are the true and maximum quality scores, respectively.

Moreover, the shearlet-based spatiotemporal features inputted to the DRBGRNN flow forward from the input layer to the output layer through several hidden layers, thereby calculating the estimated quality score of the output layer.

Third, the estimated quality score of the actual output layer is compared with true quality score (as the target value) to calculate the error value of the objective function. The output layer is a logistic regression concatenated with DRBGRNN, and its objective function incorporates a conditional block-wise constraint to automatically learn a reasonable block-level quality despite the video-level-wise quality label in the training data. Accordingly, the error of the objective function O, which proceeds in the direction of minimizing errors to train the model, is defined as

$$O = \frac{1}{S}\sum_{s=1}^{S}\left[(Q_s - F_s)^2 + \alpha(Q_s)\sum_{t=1}^{T}(Q_s - f_{s,t})^2\right], \tag{8}$$

using

$$F_s = \frac{1}{T}\sum_{t=1}^{T} f_{s,t}, \tag{9}$$

where S, F_s, T, and $f_{s,t}$ denote the total number of training videos, the estimated quality scores of the s-th video, the number of input video blocks, and the estimated block quality of the t-th video block of video s, respectively.

The first term in (8) only focuses on the accuracy of video-level quality and is not concerned with the distribution of block-level quality. On the other hand, the second term in (8) forces the block-level quality to follow a uniform distribution. That is, the weighting factor α of (7) is calculated according to the given video-level quality label, and the block-level errors are reflected in the overall error in proportion to the weighting factor α. Assuming that the maximum quality score is 5, when the quality label of the given video is 5, α is calculated as 1, and when the quality label of the given video is 1, α has a value of 0.000 1. The higher the quality score, the higher the value of the weighting factor. That is, errors in a video block segmented from a video with a high-quality label are reflected significantly in the objective function, and errors in a video block segmented from a video with a low-quality label are reflected only to a small extent. In this way, when noise or video distortion occurs in one block, reasonable block-level quality estimation is performed in which the quality score is lowered accordingly. This constraint also explicitly guides the DRBGRNN to differentiate clean frames from degraded frames.

Fourth, the total mean error of the objective function is moved backward from the output layer to the input layer through the hidden layers. Through this process, the connection strength between the output and hidden layers is updated, and the weight of the input layer is also updated. When the difference between the output and target values converges within a specified range, backpropagation is terminated, and the updated weights are stored. These weights consist of various combinations of connections, where the learning model is stored.

This approach makes it possible to automatically learn a reasonable block-level quality even when providing only the

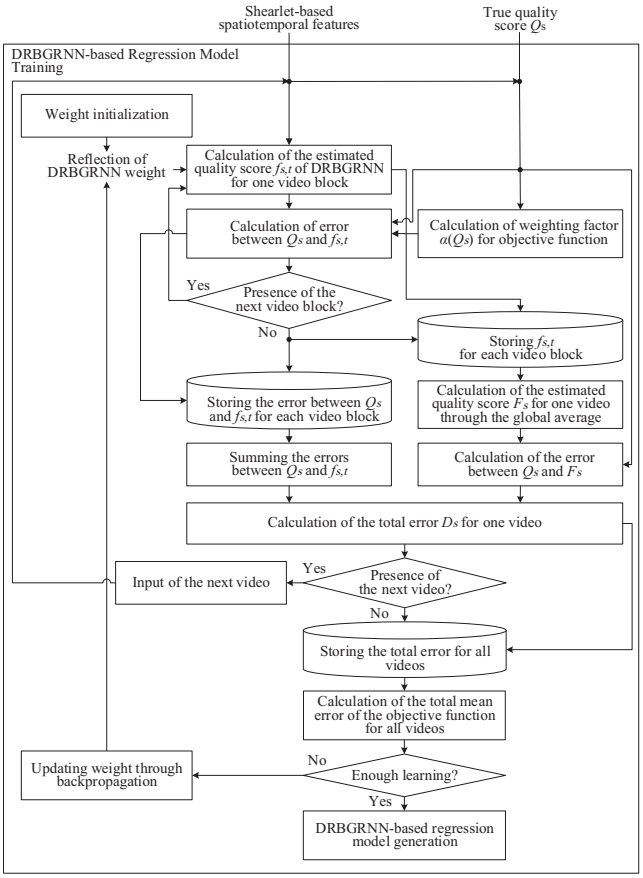

FIGURE 10 Block diagram of the conditional constraint on the video block quality assessment

quality label for each video of the training video database, without detailed quality labels.

3 | EXPERIMENTS AND RESULTS

In this section, the performance of the proposed method is evaluated for the NR-VQA of sports videos with dynamic scenes. The subsections present the dataset used for the evaluation, experimental setup and a description of the performances obtained with the proposed approach.

3.1 | Testbed infrastructure and evaluation dataset

We believe that motion information with spatiotemporal characteristics plays an important role in human perception of video. Considering this, we designed an NR-VQA method that can effectively predict video quality through a trained regression model by extracting the spatiotemporal features and applying them to the DRBGRNN architecture. Specifically, to construct a sports video database with dynamic motion scenes and to evaluate the performance of this method on sports videos, we set up a testbed. The testbed consists of a real-time transport protocol (RTP)-video server, RTP-video client, and network traffic emulator. The RTP-video server streams videos to the RTP-video client on demand. Between server and client, a network traffic emulator is installed. In the network traffic emulator module, a traffic generator is used to simulate wireless local area network connections of different traffic loads with packet loss. On reception of the impaired video with the designed traffic load, the client performs NR-VQA.

The original video set used for constructing an experimental database through the testbed consisted of 60 10-s videos at 30 frames per second (fps) from the sports video database. We considered two distortions, which can be made in the compression and transmission stages. To prepare distorted video data with compression, the original video data were compressed using the H.264 lossy video compression codec at four different bitrates: 30, 2, 1, and 0.8 Mbps. If the bit rate is 30 Mbps, it is similar to the original video. However, the distortion gradually becomes apparent below the bit rate of 2 Mbps. To consider the transmission error sequences, the packet loss rates of 2%, 4%, 6%, and 8% were applied using a network traffic emulator. The packet loss rate is calculated as the percentage of lost RTP packets with respect to the total number of RTP packets.

To effectively evaluate the performance of the proposed method and other VQA methods, the following three video quality databases are used, which contain several types of distortions.

- The sports video database (SVD): The sports video database consists of 60 reference videos and 480 distorted videos. All videos are in the YUV420 format with a resolution of 960 × 540. The subjective quality rating score of each video is given as a mean opinion score (MOS) rated by 15 participants aged between 18 and 40 years. In the MOS tests, after viewing each video, the participants rated the MOS score on a five-level scale (5 – Excellent, 4 – Good, 3 – Fair, 2 – Poor, and 1 – Bad).
- The categorical subjective image quality (CSIQ) video quality database [19]: The CSIQ video quality database consists of 12 reference videos and 216 distorted videos. All videos are in the raw YUV420 format with a resolution of 832 × 480 pixels and each video is 10 s long. The videos in the database have different frame rates, ranging from 24 to 60 fps. Each reference video has 18 distorted versions with six types of distortions. The evaluation was performed by 35 test subjects using a different MOS ranging from 1 to 100. This video database contains various movements of people as well as sports events.
- The KoNViD-1k video quality database [20]: The KoNViD-1k video quality database consists of 1200 public-domain video sequences. All the videos are in the MP4 format with a spatial resolution of 960 × 540, and each video is 8 s long. The KoNViD-1k videos are encoded at three predominant frame rates: 24, 25, and 30 fps. The subjective video quality score of each video is given as a MOS evaluated by 642 participants through crowdsourcing. This video database contains not only people, but also various objects, and contains numerous static scenes.

In each train-test iteration for the experiment, we randomly selected and applied 80% of the distorted videos as the training dataset and used the remaining for the test dataset.

3.2 | Performance comparison

We compared the performance of the proposed method with those of several NR-VQA methods designed with different neural network architectures and different spatial and temporal features. The NR-VQA methods used for comparison are as follows:

- CF-SVR (codec features and SVR) [21]: Along with features based on the spatial perceptual information measure and the temporal perceptual information measure, peak signal-to-noise ratio estimation from MPEG-2 and H.264/AVC analysis was mapped to a perceptual measure of video quality by SVR.
- VF-NN (video-level features and neural network) [22]: Six frame-level features were extracted from the discrete cosine transform coefficients of each decoded frame to quantify the distortion of natural scenes. All frame-level features for all frames were transformed to corresponding video-level features via temporal pooling and inputted into

a trained multilayer neural network to provide a predicted quality score for the video sequence.

- 3S-CNN (3D shearlet transform and CNNs): Mean spatiotemporal features were extracted by 3DST as the average of the spatiotemporal features for each video block and were exaggerated by CNN to make them more discriminative. The perceptual quality score was given by a simple logistic regression.
- 3S-CNNc: Instead of calculating and using the mean spatiotemporal features, the spatiotemporal features extracted from the 3DST were applied to the CNN. The conditional constraints were incorporated into the objective function of the logistic regression with the CNN.
- 3S-RNet: Instead of using the CNN, a ResNet was adopted as a regression model with the mean spatiotemporal features. The ResNet had three residual layers and a dense layer. Each residual layer consisted of two residual blocks with three convolution filters. The output layer was a logistic regression concatenated with ResNet.
- 3S-RNetc: Instead of using the mean spatiotemporal features, the spatiotemporal features obtained from the 3DST were applied to ResNet. The conditional constraints were incorporated into the objective function of logistic regression.
- CNN-MR (CNN and multi-regression) [14]: The spatial features were captured at the frame level by 2D CNN, and motion information was extracted as temporal information at the sequence level. A multi-regression model was applied to comprehensively measure video quality.
- ILSTM (improved LSTM) [16]: To predict video quality, an LSTM was combined with a decision tree method and applied to regression analysis.
- ConvLSTM (CNN and LSTM): The frame-level deep features were extracted from the CNN and applied to the LSTM network containing LSTM layers and a fully connected layer. The CNN had 13 convolution layers.
- ConvGRU (CNN and gated recurrent units): Instead of adopting the LSTM network, a GRU network was used as the regression model with frame-level deep features. The GRU network consisted of a two-layer GRU, and a temporal pooling was embedded as an output layer.

To evaluate the performance of NR-VQA, performance metrics such as the Spearman rank-order correlation coefficient (SROCC) and linear correlation coefficient (LCC) were employed.

- The MOS metric used in KoNViD-1k ranged over values of 1–5, while the MOS metric used in the CSIQ database ranged over values of 1–100. To facilitate a fair comparison, the range of all MOS metrics was normalized to the range 0–1 before computing LCC and SROCC. Details are described extensively in [11].
- The LCC measures the strength of a linear correlation between the estimated video quality score variable and the subjective video quality score variable. The LCC can range from +1 to −1, where +1 indicates the best quality estimation, while −1 indicates the worst quality estimation.
- The SROCC measures a single relationship regardless of the linearity and represents the LCC between the ranked variables. The SROCC has a high value if the two variables considered have similar ranks. In the absence of repeated score ranks, a complete SROCC of 1 or −1 occurs when each of the ranked variables is a perfect monotonic function of the other.

4 | RESULTS

Table 1 presents the experimental results of the proposed method, along with those of the NR-VQA methods with different neural network architectures and features, on three video quality databases: SVD, CSIQ, and KoNViD-1k.

According to the experimental results of all the comparison methods on the SVD, the proposed method, 3S-RBGc, achieved the best performance in terms of the LCC (0.8768) and SROCC (0.8861). ConvGRU exhibited a poorer performance than 3S-RBGc but slightly better performance than ConvLSTM. The results of 3S-RBG are poorer than those of ConvLSTM, but slightly better than the performance of 3S-RNetc in terms of the LCC and slightly poorer in terms of SROCC. This is because frame-level features cover more detail than spatiotemporal features, even though the structure of the DRBGRNN can effectively handle spatiotemporal features. More importantly, the conditional constraint method aids in better evaluating the video quality score. The 3S-RNet achieves

TABLE 1 Results of experiments on sports video databases

	SVD		CSIQ		KoNViD-1k	
Method	**LCC**	**SROCC**	**LCC**	**SROCC**	**LCC**	**SROCC**
CF-SVR	0.7753	0.7845	0.8256	0.8349	0.7364	0.7221
VF-NN	0.7826	0.7937	0.8386	0.8497	0.7137	0.7113
3S-CNN	0.8031	0.8158	0.8557	0.8524	0.7513	0.7426
3S-CNNc	0.8167	0.8254	0.8645	0.8731	0.7641	0.7617
3S-RNet	0.8312	0.8462	0.8658	0.8756	0.7825	0.7723
3S-RNetc	0.8537	0.8654	0.8723	0.8827	0.8038	0.7949
3S-RBG	0.8561	0.8641	0.8754	0.8812	0.8035	0.7947
CNN-MR	0.7843	0.7921	0.8341	0.8439	0.7435	0.7225
ILSTM	0.7762	0.7856	0.8301	0.8431	0.7248	0.7147
ConvLSTM	0.8753	0.8683	0.8765	0.8823	0.8087	0.7932
ConvGRU	0.8755	0.8734	0.8784	0.8863	**0.8243**	**0.8176**
3S-RBGc	**0.8768**	**0.8861**	**0.8876**	**0.8972**	0.8101	0.8023

TABLE 2 Comparison of LCC and SROCC for different combinations of the number of residual layers and number of directions

Number of residual layers	Number of directions					
	9		25		81	
	LCC	SROCC	LCC	SROCC	LCC	SROCC
5	0.7573	0.7624	0.7476	0.7621	0.7131	0.7123
9	0.8101	0.8197	0.8082	0.8150	0.7747	0.7708
12	0.8685	0.8613	0.8499	0.8528	0.8210	0.8194
13	0.8768	0.8861	0.8615	0.8768	0.8312	0.8564
14	0.8704	0.8607	0.8396	0.8503	0.8025	0.8014

better results than 3S-CNNc and is less efficient than the 3S-RNetc. This indicates that, even if the same 3DST-based spatiotemporal feature extraction method is applied, ResNet provides further improved performance than CNN by learning the spatiotemporal features of the input video block in detail. In the other four methods, 3S-CNN outperforms VF-NN, CNN-MR, ILSTM, and CF-SVR, with the worst performance provided by CF-SVR. From this, we can conclude that the logistic regression model based on the CNN outperforms SVR, NN, and MR. In addition, it was confirmed that the spatiotemporal features extracted using 3DST in each video block are more effective in regression learning than independently extracted spatial and temporal features.

To find the optimized parameters of the proposed method, comparison experiments were performed in terms of the number of residual layers of the DRBGRNN and the number of directions in the 3DST fixed at four scales. Table 2 lists the experimental results for SVD. From the results, the best performance can be obtained using 13 residual layers and nine directions, in terms of the LCC and SROCC. When the number of directions is nine, the LCC and SROCC increase with the number of residual layers as long as the number of layers is less than 13. When using more than 13 layers, the values decrease. If the number of residual layers is fixed at 13, performance is degraded when the number of directions the performance increases.

In terms of LCC and SROCC for CSIQ and KoNViD-1k, we also achieved the best performance using 13 residual layers and nine orientations.

The results for the commonly used CSIQ show that best performance is obtained with 3S-RBGc in terms of both LCC and SROCC. The ConvGRU provides better results than 3S-RBG and 3S-RNetc, but slightly poorer results than those obtained using the proposed method, 3S-RBGc. Performance priorities are almost entirely similar to the results for SVD. The reason for this is that CSIQ consists of videos that contain numerous sports events and movements of people.

In the results for KoNViD-1k with many static scenes, ConvGRU slightly outperforms the proposed method. These results indicate that a spatiotemporal feature-based logistic regression model of the proposed method is more effective for visual tracking and motion salience of dynamically moving objects than videos composed of static scenes.

5 | CONCLUSION

In this study, we presented an NR-VQA method that can effectively predict the video quality through a trained regression model by extracting spatiotemporal features and applying them to the DRBGRNN architecture based on a conditional video block-wise constraint. We trained and validated the proposed method on the three most relevant VQA datasets. The experimental results showed that the proposed method outperforms the state-of-the-art methods in evaluating the quality of sports video datasets with dynamic motion scenes; however, the performance was limited on video datasets with many static scenes, such as KoNViD-1k.

As part of future work, we will develop more advanced spatiotemporal feature extraction methods to improve the performance of our method not only for sports videos, but also for video datasets containing static motion scenes. In addition, for predicting the video quality under a wider range of conditions, we will construct a new hybrid metric that combines the simplicity of the NR algorithm with the accuracy of the video quality metric.

ORCID

Hyoung-Gook Kim https://orcid.org/0000-0002-1518-4100
Seung-Su Shin https://orcid.org/0000-0002-1904-1780

REFERENCES

1. Z. Akhtar and T. H. Falk, *Audio-visual multimedia quality assessment: A comprehensive survey*, IEEE Access **5** (2017), 21090–21117.
2. K. Manasa and S. S. Channappayya, *An optical flow-based full reference video quality assessment algorithm*, IEEE Trans. Image Process. **25** (2016), 2480–2492.
3. R. Soundararajan and A. C. Bovik, *Video quality assessment by reduced reference spatio-temporal entropic differencing*, IEEE Trans. Circuits Syst. Video Technol. **23** (2012), 684–694.
4. C. Keimel, T. Oelbaum, and K. Diepold, *No-reference video quality evaluation for high-definition video*, in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (Taipei, Taiwan), Apr. 2009, pp. 1145–1148.
5. H. Mao, R. Netravali, and M. Alizadeh, *Neural adaptive video streaming with pensieve*, in Proc. Conf. ACM Special Interest Group Data Commun. (New York, NY, USA), Aug. 2017, pp. 197–210.
6. P. Yan and X. Mou, *Video quality assessment based on correlation between spatiotemporal motion energies*, in Proc. Appl.

Digit. Image Process. XXXIX (San Diego, CA, USA), Sept. 2016, 997130: 1–12.
7. T. R. Goodall, A. C. Bovik, and N. G. Paulter, *Tasking on natural statistics of infrared images*, IEEE Trans. Image Process. **25** (2016), 65–79.
8. K. Seshadrinathan and A. C. Bovik, *Motion tuned spatio-temporal quality assessment of natural videos*, IEEE Trans. Image Process. **19** (2010), 335–350.
9. M. A. Saad and A. C. Bovik, *Blind quality assessment of videos using a model of natural scene statistics and motion coherency*, in Proc. Conf. Rec. Asilomar Conf. Signals Syst. Comput. (Pacific Grove, CA, USA), Nov. 2012, pp. 332–336.
10. M. T. Vega et al., *Deep learning for quality assessment in live video streaming*, IEEE Signal Process. Lett. **24** (2017), 736–740.
11. Y. Li et al., *No-reference video quality assessment with 3D shearlet transform and convolutional neural networks*, IEEE Trans. Circuits Syst. Video Technol. **26** (2016), 1044–1057.
12. J. Xu et al., *No-reference video quality assessment via feature learning*, in Proc. IEEE Int. Conf. Image Process. (Paris, France), Oct. 2014, pp. 491–495.
13. S. W. Fu et al., *Quality-net: An end-to-end non-intrusive speech quality assessment model based on BLSTM*, in Proc. Interspeech (Hyderabad, India), Setp. 2018, pp. 1873–1877.
14. C. Wang, L. Su, and Q. Huang, *CNN-MR for no reference video quality assessment*, in Proc. Int. Conf. Inf. Sci. Control Eng. (Changsha, China), July 2017, pp. 224–228.
15. C. G. Bampis et al., *Recurrent and dynamic models for predicting streaming video quality of experience*, IEEE Trans. Image Process. **27** (2018), 3316–3331.
16. Q. Bao, R. Huang, and X. Wei, *Video quality assessment based on the improved LSTM model,* in Image and Graphics, vol. 10667, Springer, Cham, Switzerland, 2017, pp. 313–324.
17. M. Zhang et al., *Deep residual-network-based quality assessment for SD-OCT retinal images: Preliminary study*, in Proc. Medical Imaging 2019: Image Percept., Obs. Perform., Technol. Assess. (San Diego, CA, USA), Mar. 2019, 095214: 1–6.
18. D. Nilsson and C. Sminchisescu, *Semantic video segmentation by gated recurrent flow propagation*, in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (Salt Lake City, UT, USA), June 2018, pp. 6819–6828.
19. P. V. Vu and D. M. Chandler, *ViS3: An algorithm for video quality assessment via analysis of spatial and spatiotemporal slices*, J. Electron. Imaging **23** (2014), no. 1, article no. 01316.
20. V. Hosu et al., *The konstanz natural video database (KoNViD-1k)*, in Proc. Int. Conf. Qual. Multimedia Exper. (QoMEX) (Erfurt, Germany), June 2017, pp. 1–6.
21. J. Søgaard, S. Forchhammer, and J. Korhonen, *No-reference video quality assessment using codec analysis*, IEEE Trans. Circuits Syst. Video Technol. **25** (2015), 1637–1650.
22. K. Zhu et al., *No-reference video quality assessment based on artifact measurement and statistical analysis*, IEEE Trans. Circuits Syst. Video Technol. **25** (2014), 533–546.

AUTHOR BIOGRAPHIES

Hyoung-Gook Kim received his Dr-Ing degree in electrical engineering and computer science from the Technical University of Berlin, Germany. From 1998 to 2005, he worked on mobile service robots at Daimler Benz, and speech recognition at Siemens, Berlin, Germany. Since 2007, he has been a professor in the Department of Electronics Convergence Engineering, Kwangwoon University, Seoul, Rep. of Korea. His research interests include multimedia signal processing and deep learning.

Seung-Su Shin received his BS degree in electronics convergence engineering from Kwangwoon University, Seoul, Rep. of Korea, in 2019. Currently, he is pursuing his MS degree in electronics convergence engineering from Kwangwoon University. His research interests include multimedia signal processing and deep learning.

Sang-Wook Kim received his MS degree in electrical and electronics engineering from the Korea Advanced Institute of Science and Technology, Daejeon, Rep. of Korea. From 1991 to 2016, he worked on multimedia and wearable device development at Samsung Electronics. Since 2017, he has been with the Department of Software, Chung-Ang University, Seoul, Rep. Korea. His main research interests are multimedia signal processing and human perceptions.

Gi Yong Lee received his BS degree in electronics convergence engineering from Kwangwoon University, Seoul, Rep. of Korea, in 2020. Currently, he is pursuing his MS degree in electronics convergence engineering from Kwangwoon University. His research interests include multimedia signal processing and deep learning.