Deep Video Quality Assessor: From Spatio-temporal Visual Sensitivity to A Convolutional Neural Aggregation Network

Woojae Kim[1], Jongyoo Kim[2], Sewoong Ahn[1],
Jinwoo Kim[1], and Sanghoon Lee[1(✉)]

[1]Department of Electrical and Electronic Engineering, Yonsei University
{wooyoa,anse3832,jw09191,slee}@yonsei.ac.kr
[2]Microsoft Research, Beijing, China
jongk@microsoft.com

Abstract. Incorporating spatio-temporal human visual perception into video quality assessment (VQA) remains a formidable issue. Previous statistical or computational models of spatio-temporal perception have limitations to be applied to the general VQA algorithms. In this paper, we propose a novel full-reference (FR) VQA framework named Deep Video Quality Assessor (DeepVQA) to quantify the spatio-temporal visual perception via a convolutional neural network (CNN) and a convolutional neural aggregation network (CNAN). Our framework enables to figure out the spatio-temporal sensitivity behavior through learning in accordance with the subjective score. In addition, to manipulate the temporal variation of distortions, we propose a novel temporal pooling method using an attention model. In the experiment, we show DeepVQA remarkably achieves the state-of-the-art prediction accuracy of more than 0.9 correlation, which is ∼5% higher than those of conventional methods on the LIVE and CSIQ video databases.

Keywords: Video Quality Assessment, Visual Sensitivity, Convolutional Neural Network, Attention Mechanism, HVS, Temporal Pooling

1 Introduction

With the explosive demand for video streaming services, it is vital to provide videos with high quality under unpredictable network conditions. Accordingly, video quality prediction plays an essential role in providing satisfactory streaming services to users. Since the ultimate receiver of video contents is a human, it is essential to develop a model or methodology to pervade human perception into the design of video quality assessment (VQA).

In this paper, we seek to measure the video quality by modeling a mechanism of the human visual system (HVS) by using convolutional neural networks (CNNs). When the HVS perceives a video, the perceived quality is determined by the combination of the spatio-temporal characteristics and the spatial error signal. For example, a local distortion can be either emphasized or masked

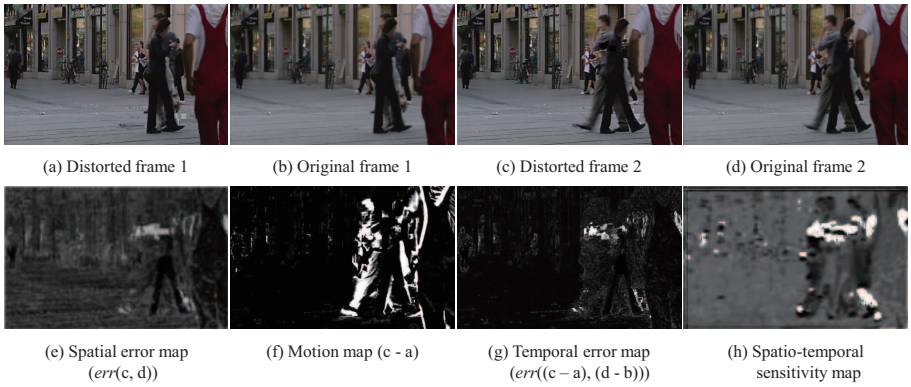

Fig. 1: Example of predicted sensitivity map: (a) and (c) are a set of consecutive distorted frames; (b) and (d) are the original frames of (a) and (c); (e) is the spatial error map of (c), calculated by an error function *err*(c, d); (f) is the motion map of distorted frame (c), calculated as subtraction of (c) and (a); (g) is the temporal error map between the distorted frames' motion map (f) and orginal motion map (d−b); (h) is the predicted spatio-temporal sensitivity map of the distorted frame (c).

by visual sensitivity depending on the spatio-temporal characteristics [1–3]. For image quality assessment (IQA), deep learning-based visual sensitivity was successfully applied to extract perceptual behavior on spatial characteristics [3]. In contrast, a video is a set of consecutive frames that contain various motion properties. The temporal variation of contents strongly affects the visual perception of the HVS, thus the problem is much more difficult than IQA. Moreover, several temporal quality pooling strategies have been attempted on VQA, but none of them could achieve high correlation as demonstrated for IQA, which still remains as a challenging issue to build a methodology to characterize the temporal human perception. In this respect, we explore a data-driven deep scheme to improve video quality remarkably from the two major motivations: *Temporal motion effect* and *Temporal memory for quality judgment*.

Temporal motion effect. Our major motivation comes from the combined masking effects caused by spatial and temporal characteristics of a video. Figs. 1 (a)-(d) show a set of consecutive distorted frames and their originals and Figs. 1 (e)-(g) show key examples of the spatial error map, a motion map, and a temporal error map of the distorted frame in (c). Each map will be explained in detail in Section 3.2. Being seen as a snapshot, several blocking artifacts induced by wireless network distortion are noticeable around pedestrians as shown in (a). However, they are hardly observable if they are shown in a playing video. This is due to a temporal masking effect which explains the phenomenon that the changes in hue, luminance, and size are less visible to humans when there exist large motions [4]. On the other hand, when a severe error in the motion map occurs as demonstrated in Fig. 1 (g), spatial errors become more visible to humans,

which is known as a mosquito noise in video processing studies [5,6]. Owing to these complex interactions between the spatial errors and motions, conventional IQA methods usually result in inaccurate predictions of the perceptual quality of distorted videos. In the meantime, among VQA studies, many attempts have been made to address the above phenomena by modeling the spatio-temporal sensitivity of the HVS [7–10]. However, these studies yielded limited performances because it is formidable to design a general purpose model considering both spatial and temporal behaviors of the HVS. Therefore, we propose a top-down approach where we establish the relationship between the distortions and perceptual scores first, then it is followed by pixel-wise sensitivities considering both spatial and temporal factors. Fig. 1 (h) is an example of the predicted spatio-temporal sensitivity map by ours. The dark regions such as the pedestrians are predicted less sensitively by the strong motion in Fig. 1 (f), while the bright regions have high weights by the temporal error component in Fig. 1 (g).

Temporal memory for quality judgment. In addition, as our second motivation, we explore the retrospective quality judgment patterns of humans given the quality scores of the frames in a video, which is demonstrated in Fig. 2. If there exist severely distorted frames in a video (Video B), humans generally determine that it has lower quality than a video having uniform quality distribution (Video A) even though both of them have the same average quality. Accordingly, a simple statistical temporal pooling does not work well in VQA [1,11,12]. Therefore, there has been a demand for an advanced temporal pooling strategy which reflects humans' retrospective decision behavior on video quality.

Our framework, which we call as Deep Video Quality Assessor (DeepVQA), fully utilizes the advantages of a convolutional neural network. To predict the spatio-temporal sensitivity map, a fully convolutional model is employed to extract useful information regarding visual perception which is embedded in a VQA database. Moreover, we additionally develop a novel pooling algorithm by borrowing an idea from an 'attention mechanism', where a neural network model focuses on only specific parts of an input [13–15]. To weight the predicted quality score of each frame adaptively, the proposed scheme uses a convolution operation, which we named a convolutional neural aggregation network (CNAN). Rather than taking a single frame quality score, our pooling method considers the distribution of predicted scores. Our contributions are summarized as follows:

1. The spatio-temporal sensitivity map is predicted through self-training without any prior knowledge of the HVS. In addition, a temporal pooling method is adaptively performed by utilizing the CNAN network.
2. Since the spatio-temporal sensitivity map and temporal pooling weight are derived as intermediate results, it is able to infer and visualize an important cue of human perception based on the correlation between the subjective and objective scores from the reverse engineering perspective, which is totally different from modeling based conventional methods.
3. Through achieving the state-of-the-art performance via end-to-end optimization, the human perception can be more clearly verified by the CNN/Attention based full reference (FR) VQA framework.

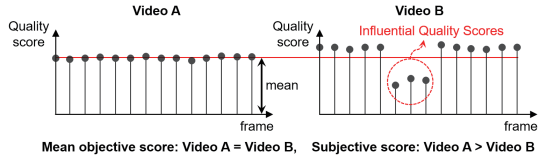

Fig. 2: Example of temporal quality variation and its effect on quality judgment.

2 Related Works

2.1 Spatio-temporal Visual Sensitivity

Numerous VQA models have been developed with respect to the human visual sensitivity. From these works, masking effects have been explained by a spatio-temporal contrast sensitivity function (CSF) [16–18]. According to the spatio-temporal CSF which resembles a band-pass filter, humans are not sensitive to signals with very low or high frequencies. Therefore, if strong contrast or motions exist, distortions are less noticeable in accordance with the masking effects [4,19, 20]. Based on these observations, various VQA methods have been developed. Saad *et al.* [7] used motion coherency and ego-motion as features that affect temporal masking. Mittal *et al.* [21] introduced a natural video statistics (NVS) theory, which is based on experimental results that pixel distributions can affect the visual sensitivity. However, there is a limitation in reflecting the complicated behavior of the HVS into the visual sensitivity models by these prior knowledge. Therefore, we design a learning-based model that learns human visual sensitivity autonomously from visual cues that affect the HVS.

Recently, there have been attempts to learn visual sensitivity by using deep-learning in I/VQA [3,22,23]. However, they did not consider motion properties when they extracted quality features. Therefore, a limitation still exists in predicting the effect of large motion variance.

2.2 Temporal Pooling

Temporal quality pooling methods have been studied in the VQA field. As mentioned, the simple strategy of taking the average has been employed in many VQA algorithms [24–26]. Other studies have analyzed the score distribution and adaptively pooled the temporal scores from the HVS perspective [12]. However, since these naive pooling strategies utilize only limited temporal features, it is difficult to generalize to practical videos.

Recently, the attention mechanism has been developed in machine learning field [13,15]. Attention mechanisms in neural networks are based on the visual attention in the HVS. The attention-based method essentially allows the model to focus on specific regions and adjust focus over the temporal axis. Motivated by this, there was a study to solve temporal pooling through attention feature embedding [14]. However, since it adaptively embeds a weight vector to each independent score feature vector, it is difficult to effectively utilize the scheme for

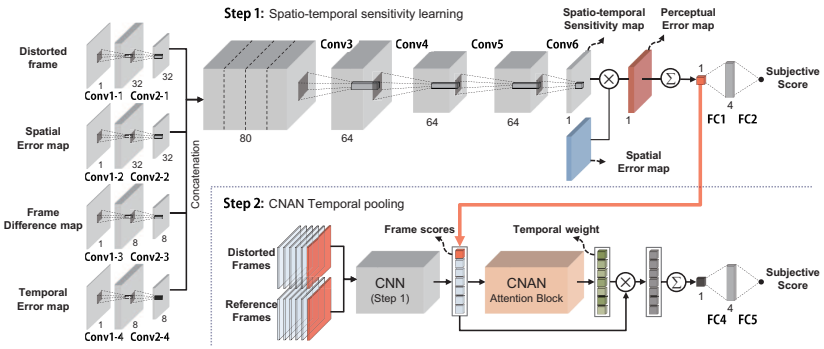

Fig. 3: Architecture of DeepVQA. The model takes a distorted frame, the spatial error map, the frame difference map and the temporal error map as input. Step 1: The CNN model is regressed onto a subjective score by the average pooling. Step 2: the overall frame scores are pooled using the CNAN and regressed onto the subjective score.

video temporal quality pooling due to the lack of consideration of the temporal score context. Instead, we use the convolution operation to detect specific patterns of score distribution, so it adaptively weights and combines the temporal scores as shown in Fig. 2.

3 DeepVQA Framework

3.1 Architecture

Visual sensitivity indicates which area of a given spatial error signal is perceived more sensitively to the HVS. The most intuitive way to learn visual sensitivity is to extract the weight map for a given spatial error map. As mentioned in Section 1, the visual sensitivity of a video content is determined by the spatial and temporal factors. Hence, by putting the sufficient information containing these factors as inputs, the model is able to learn visual sensitivity that reflects spatial and temporal masking effects. The proposed framework is depicted in Fig. 3. In our method, the spatio-temporal sensitivity map is first learned in step 1, then, a sensitivity weighted error map for each frame is temporally aggregated by the CNAN in step 2. As shown in Fig. 3, the CNAN takes a set of video frame scores in step 1 and computes a single pooled score as output.

A deep CNN with 3×3 filters is used for step 1 inspired by the recent CNN based work [3] for IQA. To generate the spatio-temporal sensitivity map without losing the location information, the model contains only convolutional layers. In the beginning, the distorted frames and spatial error maps are fed to spatial sensitivity representation. In addition, the model takes the frame difference and temporal error maps into account for a temporal factor. Each set of the input

maps goes through independent convolutional layers, and feature maps are concatenated after the second convolutional layer. Zero-padding is applied before each convolution to preserve the feature map size. Two stridden convolutions are used for subsampling. Therefore the size of the final output is 1/4 compared to that of the original frame, and the ground-truth spatial error maps are downscaled to 1/4 correspondingly. At the end of the model in step 1, two fully connected layers are used to regress features onto the subjective scores.

In step 2, the proposed CNAN is trained using the pre-trained CNN model in step 1 and regressed onto the subjective video score as shown in Fig. 3. Once each feature is derived from the previous CNN independently, they are fed into the CNAN. By the CNAN, an aggregated score yields the final score. After then, two fully connected layers are used to regress the final score.

Frame Normalization. From the HVS perspective, each input in Fig. 3 is preprocessed to make the necessary properties stand out. Since the CSF shows a band-pass filter shape peaking at around 4 cycles per degree, and sensitivity drops rapidly at low-frequency [27]. Therefore, the distorted frames are simply normalized by subtracting the lowpass filtered frames from its grey scaled frames (range in [0, 1]). The normalized frames are denoted by $\hat{I}_r^t$ and $\hat{I}_d^t$ for given distorted I_d^t and reference I_r^t frames where t is frame index.

Patch-based Video Learning. In the previous deep-learning based IQA works, a patch-based approach was successfully applied [3,23,28–30]. In our model, each video frame is split into patches, and then all the sensitivity patches in one frame are extracted. Next, these are used to reconstruct the sensitivity map as shown in Fig. 3. To avoid the overlapped regions of the predicted perceptual error map, the step of the sliding window is determined as $step_{patch} = size_{patch} - (N_{ign} \times 2 \times R)$, where N_{ign} is the number of ignored pixels, and R is the size ratio of the input and the perceptual error map. In the experiment, the ignored pixel N_{ign} was setted 4, and the patch size $size_{patch}$ was 112 112. To train the model, one video was split into multiple patches, which were then used as one training sample. In step 1, 12 frames per video were uniformly sampled, and 120 frames were used to train the model in step 2.

3.2 Spatio-temporal Sensitivity Learning

The goal of spatio-temporal sensitivity learning is to derive an importance of each pixel for a given error map. To achieve this, we utilize the distorted frame and spatial error map as the spatial factors. Also, the frame difference and temporal error maps are used as the temporal factor. We define a spatial error map $\boldsymbol{e}_{\mathrm{s}}^t$ as a normalized log difference, as in [3],

$$\boldsymbol{e}_{\mathrm{s}}^t = \frac{\log(1/((\hat{I}_r^t - \hat{I}_d^t)^2 + \epsilon/255^2)}{\log(255^2/\epsilon)}, \tag{1}$$

where $\epsilon = 1$ for the experiment. To represent motion map, the frame difference is calculated along the consecutive frames. Since each video contains different frames per second (fps), the frame difference map considering fps variation is

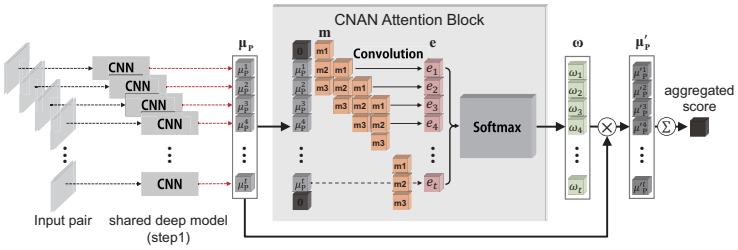

Fig. 4: Architecture of convolutional neural aggregation network.

simply defined as $\boldsymbol{f}_d^t = |I_d^{t+\delta} - I_d^t|$ where $\delta = \lfloor fps/25 \rfloor$. In a similar way, the temporal error map, which is the difference between the motion information of the distorted and reference frames, is defined as $\boldsymbol{e}_{\mathrm{T}}^t = |\boldsymbol{f}_d^t - \boldsymbol{f}_r^t|$, where $\boldsymbol{f}_r^t$ is the frame difference of the reference frame. Then, the spatio-temporal sensitivity map $\mathbf{s}^t$ is obtained from the CNN model of step 1 as

$$\mathbf{s}^t = CNN_{s1}(\hat{I}_d^t, \boldsymbol{e}_{\mathrm{S}}^t, \boldsymbol{f}_d^t, \boldsymbol{e}_{\mathrm{T}}^t; \theta_{s1}), \tag{2}$$

where CNN_{s1} is the CNN model of step 1 with parameters θ_{s1}. To calculate a global score of each frame, the perceptual error map is defined by $\mathbf{p}^t = \mathbf{s}^t \bigodot \mathbf{e}_{\mathrm{S}}^t$, where $\bigodot$ is element-wise product.

Because we use zero-padding before each convolution, we ignore border pixels which tend to be zero. Each four rows and columns for each border are excluded in the experiment. Therefore, the spatial score $\mu_{\mathbf{p}}^t$ is derived by averaging the cropped perceptual error map $\mathbf{p}^t$ as

$$\mu_{\mathbf{p}}^t = \frac{1}{(H-8)\cdot(W-8)} \sum_{(i,j)\in\Omega} \mathbf{p}^t, \tag{3}$$

where H and W are the height and width of $\mathbf{p}^t$, (i, j) is a pixel index in cropped region Ω. Then, the score in step 1 is obtained by average pooling over spatial scores as $\mu_{s1} = \sum_t \mu_{\mathbf{p}}^t$. The pooled score is, then, fed into two fully connected layers to rescale the prediction. Then the final objective loss function is defined by a weighted summation of loss function and the regularization term as

$$\mathcal{L}_{step1}(\hat{\mathbf{I}}_d, \mathbf{e}_{\mathrm{S}}, \mathbf{f}_d, \mathbf{e}_{\mathrm{T}}; \theta_{s1}, \phi_1) = ||f_{\phi_1}(\mu_{s1}) - \mathbf{s}_{sub}||_2^2 + \lambda_1 TV + \lambda_2 L_2,$$

where $\hat{\mathbf{I}}_d, \mathbf{e}_{\mathrm{S}}, \mathbf{f}_d, \mathbf{e}_{\mathrm{T}}$ are sequences of each input, $f(\cdot)$ is a regression function with parameters ϕ_1 and $\mathbf{s}_{sub}$ is the ground-truth subjective score of the distorted video. In addition, a total variation (TV) and L_2 norm of the parameters are used to relieve high-frequency noise in the spatio-temporal sensitivity map and to avoid overfitting [3]. λ_1 and λ_2 are their weight parameters, respectively.

3.3 Convolutional Neural Aggregation Network

In step 1, the average of the perceptual error maps over spatial and temporal axes is regressed to a global video score. As mentioned, simply applying a mean

pooling results in inaccurate predictions. To deal with this problem, we conduct temporal pooling for each frame's predicted score using the CNAN in step 2.

The memory attention mechanism has been successfully applied in various applications to pool spatial or temporal data [13–15]. Likewise, the CNAN is designed to predict human patterns of score judgment over all the frames scores. The basic idea is to use a convolutional neural model to learn external memories through a differentiable addressing/attention scheme. Then the learned memories adaptively weight and combine scores across all frames.

Fig. 4 shows the architecture of the CNAN for temporal pooling. The overall frame scores from step 1 are represented by a single vector $\boldsymbol{\mu}_{\mathbf{p}}$. We then define a set of corresponding significance $\mathbf{e}$ in the attention block using the memory kernel $\mathbf{m}$. To generate the significance $\mathbf{e}$, one dimensional convolution is performed on the given $\boldsymbol{\mu}_{\mathbf{p}}$ using the memory kernel $\mathbf{m}$. In other words, the significance is designed to learn a specific pattern of score variation during a certain filter length. This operation can be described as a simple convolution $\mathbf{e} = \mathbf{m} * \boldsymbol{\mu}_{\mathbf{p}}$. To maintain the dimension of weights equal to $\boldsymbol{\mu}_{\mathbf{p}}$, we padded zeros to the border of input $\boldsymbol{\mu}_{\mathbf{p}}$. They are then passed to the softmax operator to generate positive temporal weights ω_t with $\sum_t \omega_t = 1$ as

$$\omega_t = \frac{\exp(e_t)}{\sum_j \exp(e_j)}. \tag{4}$$

Finally, the temporal weight ω_t derived from the attention block, is applied to the origin score vector to generate the final aggregated video score as $\mu_{s2} = \sum_t \omega_t \mu_{\mathbf{p}}^t$. Therefore, the objective function in step 2 is represented as

$$\mathcal{L}_{step2}(\hat{\mathbf{I}}_d, \mathbf{e}_{\mathrm{S}}, \mathbf{f}_d, \mathbf{e}_{\mathrm{T}}; \theta_{s1}, \phi_2) = ||f_{\phi_2}(\mu_{s2}) - \mathbf{s}_{sub}||_2^2 \tag{5}$$

where, $f_{\phi_2}(\cdot)$ represents a nonlinear regression function with parameters ϕ_2, and θ_{s1} refers to parameters in step 1.

4 Experimental Results

Since our goal is to learn spatio-temporal sensitivity and to aggregate frame scores via the CNAN, we chose the baseline model which takes only two spatial inputs (DeepQA [3]). Moreover, to study the effect of the temporal input, two simpler models of DeepVQA without CNAN are defined. First, DeepVQA-*3ch* takes only two spatial inputs and the frame difference map. Second, DeepVQA-*4ch* takes all input maps. For both models, average pooling was conducted as described in step 1. We indicate the complete model as DeepVQA-*CNAN*.

4.1 Dataset

To evaluate the proposed algorithm, two different VQA databases were used: LIVE VQA [11], and CSIQ [31] databases. The LIVE VQA database contains 10 references and 150 distorted videos with four distortion types: wireless, IP,

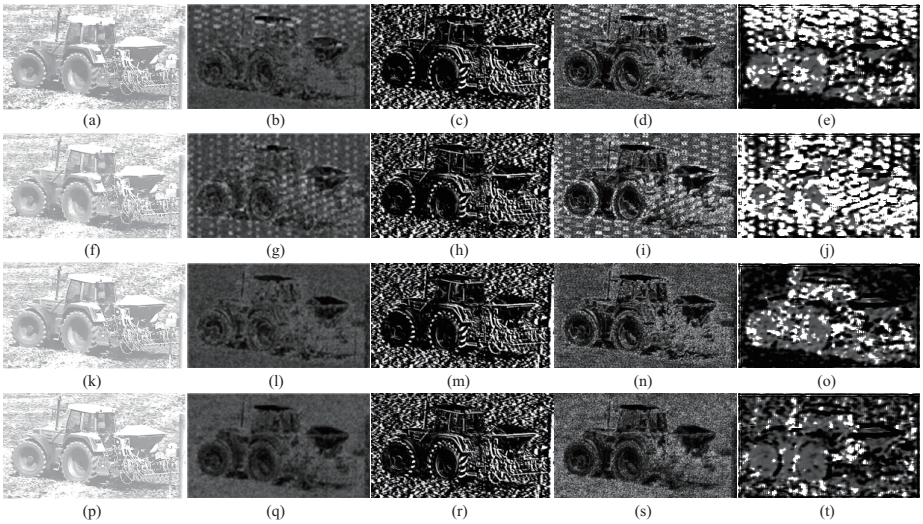

Fig. 5: Examples of the predicted sensitivity maps; (a), (f), (k), and (p) are distorted frames with Wireless, IP, H.264 compression, and MPEG-2 compression; (b), (g), (l), and (q) are the objective error maps; (c), (h), (m), and (r) are the frame difference maps; (d), (i), (n), and (s) are the temporal error maps; (e), (j), (o), and (t) are the predicted spatio-temporal sensitivity maps.

H.264 compression and MPEG-2 compression distortions. The CSIQ database includes 12 references and 216 distorted videos with six distortion types: motion JPEG (MJPEG), H.264, HEVC, wavelet compression using SNOW codec, packet-loss in a simulated wireless network and additive white Gaussian noise (AWGN). In the experiment, the ground-truth subjective scores were rescaled to the range [0, 1]. For differential mean opinion score (DMOS) values, their scale was reversed so that the larger values indicate perceptually better videos. Following the recommendation from the video quality experts group [32], we evaluated the performance of the proposed algorithm using two standard measures, i.e., Spearman's rank order correlation coefficient (SROCC) and Pearson's linear correlation coefficient (PLCC).

4.2 Spatio-temporal Sensitivity Prediction

To study the relevance of trained DeepVQA-*4ch* to the HVS, the predicted spatio-temporal sensitivity maps are shown in Fig. 5. Here, DeepVQA-*4ch* was trained with λ_1=0.02, λ_2=0.005. An example frames with four types of artifacts (wireless, IP, H.264 and MPEG-2) are represented in Figs. 5 (a), (f), (k) and (p). Figs. 5 (b), (g), (l) and (q) are the spatial error maps, (c), (h), (m) and (r) are the frame difference maps, (d), (i), (n) and (s) are the temporal error maps, and (e), (j), (o) and (t) are the predicted sensitivity maps. In Fig. 5, darker regions indicate that pixel values are low. In case of wireless and IP distortions,

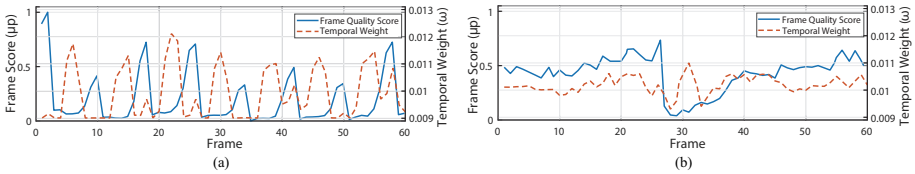

Fig. 6: Examples of frame quality scores $\mu_{\mathbf{p}}$ and its temporal weight ω from the CNAN. (a) shows first 60 frames of "st02_25fps.yuv" in the LIVE video database; (b) shows first 60 frames of "mc13_50fps.yuv" in the LIVE video database.

temporal errors ((d) and (i)) are large in overall areas. Since humans are very sensitive to this motion variation cues, predicted sensitivity values ((e) and (j)) are high in all areas. Conversely, for H.264 and M-JPEG2 distortions, temporal errors ((n) and (s)) are relatively lower than those of wireless and IP distortions. In this case, the frame difference map which contains the motion information is a dominant factor in predicting the sensitivity map. In Fig. 5, a foreground object is being tracked in a video. Therefore, the motion maps ((m) and (r)) in the background region have higher values than those of the object. Finally, the value of background regions in the predicted sensitivity maps ((o) and (t)) is relatively low. These results are consistent with the previous studies on the temporal masking effect, which cannot be obtained only by considering spatial masking effect. Therefore, it can be concluded that the temporal information, as well as spatial error, is important to quantify the visual quality of videos.

4.3 CNAN Temporal Pooling

To evaluate the CNAN, we analyzed the relationship between the temporal pooling weight ω computed in the attention block and the normalized spatial score $\mu_{\mathbf{p}}$ computed in step 1. Here, the size of kernel $\mathbf{m}$ was set to 21×1 experimentally. Figs. 6 (a) and (b) show two predicted temporal score distributions of $\mu_{\mathbf{p}}$ (straight line) and its temporal weights ω (dotted line). In Fig. 6 (a), the scores tend to rise or fall sharply at about every 5 frames. Conversely, the temporal weight has a higher value when the predicted score is low. This is because, as mentioned in Section 1, the human rating is highly affected by negative peak experiences than the overall average quality [7,12]. Therefore, it is obvious that the learned model mimics the temporal pooling mechanism of a human.

Fig. 6 (b) shows that the scores are uniformly distributed except for the middle region. As explained before, the CNAN shows the filter response for a particular pattern by memory kernel $\mathbf{m}$. Thus, the filter response of a monotonous input signal also tends to be monotonous. However, at near the 30^{th} frame, the temporal pooling weight ω is large when the frame score abruptly changes. Therefore, the CNAN enables to reflect the behavior of scoring appropriately and leads to a performance improvement in Tables 3 and 4.

Table 1: Comparison of computational cost and median SROCC according to the number of sampled frames in LIVE database.

# of sampled frame	6	12	48	120
SROCC (120 epochs)	0.8787	0.8812	0.8772	0.8704
Computational time (1 epoch)	69s	201s	796s	1452s

Table 2: Cross dataset comparison on the LIVE video database (SROCC).

Models	Wireless	IP	H.264	MPEG-2	ALL
DeepVQA-*4ch*	0.8134	0.8023	0.8726	0.8439	0.8322
DeepVQA-*CNAN*	0.8211	0.8214	0.8748	0.8624	0.8437

4.4 Number of Frames vs. Computation Cost

The number of video frames used for training the DeepVQA model has a great impact on a computational cost. As shown in Fig. 6, although the quality scores vary for each frame, the distribution shows certain patterns. Therefore, it is feasible to predict a quality score by using only a few sampled frames. To study the computational cost, we measure the performance according to the sampling rate.

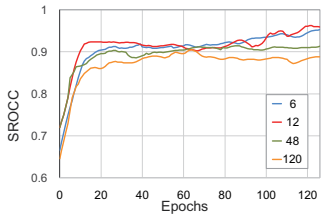

Fig. 7: Comparison of SROCC curves according to the number of sampled frames (6, 12, 48 and 120 frames).

For the simulation, a machine powered by a Titan X and equipped with the Theano. SROCC over 130 epochs with the 4 subset frames (6, 12, 48 and 120) is depicted in Fig. 7. When the number of sampled frames was 12, the SROCC was slightly higher than those of the other cases with a faster convergence speed. However, when the sampled frame was 120, the model suffered overfitting after 70 epochs, showing performance degradation. As shown in Table 1, DeepVQA obviously shows higher performance and lower execution time when using a video subset which contains a small number of frames.

4.5 Ablation Study

We verify the ablation of each input map and CNAN in our framework. To evaluate the ablation set, we tested each model (DeepQA (*2ch* [3]), DeepVQA-*3ch*, DeepVQA-*4ch* and DeepVQA-*CNAN*) on the LIVE and CSIQ databases. The experimental settings will be explained in Section 4.6 and the comparison results are tabulated in Tables 3 and 4. DeepQA [3] using only the distorted frame and spatial error map yielded lower performance than DeepVQA-*3ch* and *4ch*. Since DeepQA only infers the visual sensitivity of the spatial masking effect, it is strongly influenced by the spatial error signals. However, the performances of DeepVQA-*3ch* and *4ch* which were designed to infer the temporal motion effects were gradually improved. Moreover, the DeepVQA model combined with CNAN showed the highest performance since it considers the human patterns of quality judgment.

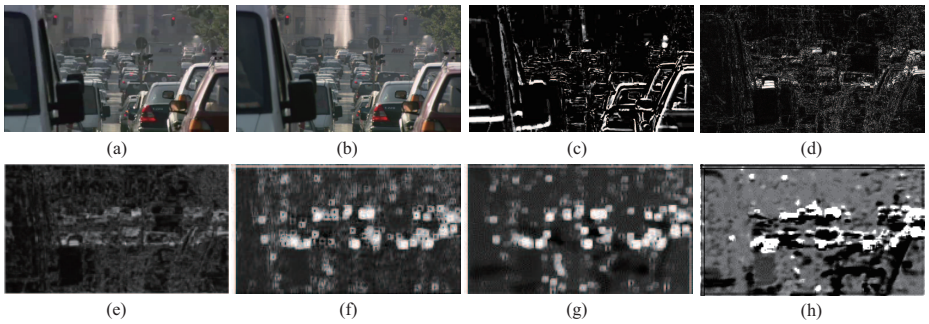

Fig. 8: Examples of the predicted sensitivity maps with different channel inputs: (a) is the distorted frame; (b) is the original frame; (c) is the frame difference map; (d) is the temporal error map; (e) is the spatial error map (f)-(h) are its predicted sensitivity maps from DeepQA (*2ch*) [3], DeepVQA-*3ch* and DeepVQA-*4ch*, respectively.

To study the effect of each channel input, we visualized the spatio-temporal sensitivity maps over different channel input. Fig. 8 shows the predicted sensitivity map with different channel inputs. Figs. 8 (a), (b) and (e) are the distorted frame, its original and the spatial error map, respectively. In the case of Fig. 8 (f), the local region of the sensitivity map looks similar to the spatial blocking artifact. However, when the frame difference map (Fig. 8 (c)) is added in the model as Fig. 8 (g), the sensitivity is decreased for the regions with strong motions (darker region) as we expected. Finally, as Fig. 8 (h), when all the four inputs including the temporal error map (Fig. 8 (d)) are used, the sensitivity map is learned to consider all of the motion effects as described in Section 1. In addition, as the number of channels increases, the predicted sensitivity map tends to be smoother, which agrees with the HVS well [3].

4.6 Performance Comparison

To evaluate the performances, we compared DeepVQA with state-of-the-art I/VQA methods on the LIVE and CSIQ databases. We first randomly divided the reference videos into two subsets (80% for training and 20% for testing) and their corresponding distorted videos were divided in the same way so that there was no overlap between the two sets. DeepVQA was trained in a non-distortion-specific way so that all the distortion types were used simultaneously. The training stage of step 1 (step 2) iterated 300 (20) epochs, then a model with the lowest validation error was chosen over the epochs. The accuracy of step 1 mostly saturated after 200 epochs as shown in Fig. 7. The correlation coefficients of the testing model are the median values of 20 repeated experiments while dividing the training and testing sets randomly in order to eliminate the performance bias. DeepVQA-*3ch*, DeepVQA-*4ch* and DeepVQA-*CNAN* were compared to FR I/VQA models: PSNR, SSIM [33], VIF [34], ST-MAD [35],

Table 3: Median PLCC and SROCC comparison on the LIVE VQA Database. *Italics* indicate full-reference (FR) methods.

Metrics	PLCC					SROCC				
	Wireless	IP	H.264	MPEG-2	**ALL**	Wireless	IP	H.264	MPEG-2	**ALL**
PSNR	0.7274	0.6395	0.7359	0.6545	0.7499	0.7381	0.6000	0.7143	0.6327	0.6958
SSIM [33]	0.7969	0.8269	0.7110	0.7849	0.7883	0.7381	0.7751	0.6905	0.7846	0.7211
VIF [34]	0.7473	0.6925	0.6983	0.7504	0.7601	0.7143	0.6000	0.5476	0.7319	0.6861
STMAD [35]	**0.8887**	**0.8956**	0.9209	0.8992	0.8774	0.8257	0.7721	0.9323	0.8733	0.8301
ViS3 [36]	0.8597	0.8576	0.7809	0.7650	0.8251	0.8257	0.7712	0.7657	0.7962	0.8156
MOVIE [25]	0.8392	0.7612	0.7902	0.7578	0.8112	0.8113	0.7154	0.7644	0.7821	0.7895
V-BLIINDS [7]	**0.9357**	**0.9291**	0.9032	0.8757	0.8433	0.8462	0.7829	0.8590	0.9371	0.8323
SACONVA [26]	0.8455	0.8280	0.9116	0.8778	0.8714	**0.8504**	0.8018	**0.9168**	0.8614	0.8569
DeepQA [3]	0.8070	0.8790	0.8820	0.8830	0.8692	0.8290	0.7120	0.8600	0.8940	0.8678
DeepVQA-3ch	0.8723	0.8661	**0.9254**	**0.9222**	0.8754	0.8376	**0.8615**	0.9014	**0.9543**	0.8723
DeepVQA-4ch	0.8867	0.8826	**0.9357**	**0.9416**	**0.8813**	**0.8494**	**0.8716**	**0.9193**	**0.9664**	**0.8913**
DeepVQA-VQPooling	-	-	-	-	**0.8912**	-	-	-	-	**0.8987**
DeepVQA-CNAN	**0.8979**	**0.8937**	**0.9421**	**0.9443**	**0.8952**	**0.8674**	**0.8820**	**0.9200**	**0.9729**	**0.9152**

Table 4: Median PLCC and SROCC comparison on the CSIQ VQA Database. *Italics* indicate full-reference (FR) methods.

Metrics	PLCC							SROCC						
	H.264	PLoss	MJPEG	Wavelet	AWGN	HEVC	**ALL**	H.264	PLoss	MJPEG	Wavelet	AWGN	HEVC	**ALL**
PSNR	0.9208	0.8246	0.6705	**0.9235**	0.9321	0.9237	0.7137	0.8810	0.7857	0.6190	0.8810	0.8333	0.8571	0.7040
SSIM [33]	0.9527	0.8471	0.8047	0.8907	**0.9748**	**0.9652**	0.7627	0.9286	0.8333	0.6905	0.8095	**0.9286**	0.9148	0.7616
VIF [34]	0.9505	**0.9212**	0.9114	**0.9241**	**0.9604**	**0.9624**	0.7282	0.9048	0.8571	0.8095	0.8571	0.8810	0.9012	0.7256
STMAD [35]	**0.9619**	0.8793	0.8957	0.8765	0.8931	0.9274	0.8254	0.9286	0.8333	0.8333	0.8095	0.8095	0.8810	0.8221
ViS3 [36]	0.9356	0.8299	0.8110	**0.9303**	0.9373	**0.9677**	0.8100	0.9286	0.8095	0.7857	0.9048	0.8571	0.9025	0.8028
MOVIE [25]	0.9035	0.8821	0.8792	0.8981	0.8562	0.9372	0.7886	0.8972	0.8861	**0.8874**	0.9012	0.8392	0.9331	0.8124
V-BLIINDS [7]	0.9413	0.7681	0.8536	0.9039	0.9318	0.9214	0.8494	0.9048	0.7481	0.8333	0.8571	**0.9048**	0.8810	0.8586
SACONVA [26]	0.9133	0.8115	0.8565	0.8529	0.9028	0.9068	0.8668	0.9048	0.7840	0.7857	0.8333	0.8810	0.8333	0.8637
DeepQA [3]	0.8753	0.8456	0.8460	0.9103	**0.9423**	0.9213	0.8723	0.8921	0.9013	**0.8623**	0.8010	**0.9021**	0.9566	0.8752
DeepVQA-3ch	0.9398	0.9009	**0.9159**	0.8621	0.8090	0.8756	0.8827	**0.9622**	**0.9501**	0.8103	**0.9134**	0.8145	**0.9718**	0.8854
DeepVQA-4ch	**0.9579**	**0.9241**	**0.9375**	0.8856	0.8271	0.8894	**0.9013**	**0.9732**	**0.9662**	0.8390	**0.9344**	0.8314	**0.9925**	**0.9043**
DeepVQA-4ch-VQPooling	-	-	-	-	-	-	**0.9057**	-	-	-	-	-	-	**0.9067**
DeepVQA-4ch-CNAN	**0.9633**	**0.9335**	**0.9401**	0.8853	0.8153	0.8897	**0.9135**	**0.9777**	**0.9672**	**0.8510**	**0.9243**	0.8106	**0.9950**	**0.9123**

ViS3 [36], MOVIE [25] and DeepQA [3]. For IQA metrics (PSNR, SSIM, VIF and DeepQA), we took an average pooling for each frame score to get a video score. In addition, the no-reference (NR) VQA models were benchmarked: V-BLIINDS [7], SACONVA [26]. To verify the temporal pooling performance, we further compare the existing temporal pooling method: VQPooling [12].

Tables 3 and 4 show the PLCC and SROCC comparisons for individual distortion types on the LIVE and CSIQ databases. The last column in each table reports overall SROCC and PLCC for all the distortion types, and the top three models for each criterion are shown in bold. Since our proposed model is a non-distortion specific model, the model should work well for overall performance when various distortion types coexist in the dataset. In our experiment, the highest SROCC and PLCC of overall distortion types were achieved by DeepVQA-*CNAN* in all the databases. In addition, DeepVQA-*CNAN* are generally competitive in most distortion types, even when each type of distortion is evaluated separately. Because the most of the distortion types in LIVE and CSIQ is distorted by video compression, which cause local blocking artifacts, there are

many temporal errors in the databases. For this reason, the spatio-temporal sensitivity map is excessively activated in the large-scale block distortion type such as Fig. 5 (j). Therefore, DeepVQA achieved relatively low performance in face of Wireless and IP distortions which include a large size of blocking artifacts. As shown in Table 4, since AWGN causes only spatial distortion, it shows a relatively low performance compared to the other types having blocking artifacts. Nevertheless, DeepVQA achieved a competitive and consistent accuracy across all the databases. Also, comparing the DeepVQA-*4ch* and DeepQA, we can infer that using the temporal inputs helps the model to extract useful features leading to an increase in an accuracy. Furthermore, VQPooling (DeepVQA-*VQPooling*) showed a slight improvement compared to DeepVQA-*4ch*, but CNAN showed approximately ∼2% improvement. Therefore, it can be concluded that temporal pooling via the CNAN improves performance the overall prediction.

4.7 Cross Dataset Test

To test the generalization capability of DeepVQA, the model was trained using the subset of the CSIQ video database, and tested on the LIVE video database. Since the CSIQ video database contains broader kinds of distortion types, we selected four distortion types (H.264, MJPEG, PLoss, and HEVC) which are similar in the LIVE database. The results are shown in Table 2, where both DeepVQA and DeepVQA-CNAN show nice performances. We can conclude that this models do not depend on the databases.

5 Conclusion

In this paper, we proposed a novel FR-VQA framework using a CNN and a CNAN. By learning a human visual behavior in conjunction with spatial and temporal effects, it turned out the proposed model is able to learn the spatio-temporal sensitivity from a human perception point of view. Moreover, the temporal pooling technique using the CNAN predicted the temporal scoring behavior of humans. Through the rigorous simulations, we demonstrated that the predicted sensitivity maps agree with the HVS. The spatio-temporal sensitivity maps were robustly predicted against the various motion and distortion types. In addition, DeepVQA achieved the state-of-the-art correlations on LIVE and CSIQ databases. In the future, we plan to advance the proposed framework to NR-VQA, which is one of the most challenging problems.

Acknowledgment. This work was supported by Institute for Information & communications Technology Promotion through the Korea Government (MSIP) (No. 2016-0-00204, Development of mobile GPU hardware for photo-realistic real-time virtual reality)

References

1. Ninassi, A., Le Meur, O., Le Callet, P., Barba, D.: Considering temporal variations of spatial visual distortions in video quality assessment. IEEE Journal of Selected Topics in Signal Processing **3**(2) (2009) 253–265
2. Bovik, A.C.: Automatic prediction of perceptual image and video quality. Proceedings of the IEEE **101**(9) (2013) 2008–2024
3. Kim, J., Lee, S.: Deep learning of human visual sensitivity in image quality assessment framework. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.(CVPR). (2017)
4. Suchow, J.W., Alvarez, G.A.: Motion silences awareness of visual change. Current Biology **21**(2) (2011) 140–143
5. Fenimore, C., Libert, J.M., Roitman, P.: Mosquito noise in mpeg-compressed video: test patterns and metrics. In: Proceedings-SPIE The International Society For Optical Enginnering, International Society for Optical Engineering; 1999 (2000) 604–612
6. Jacquin, A., Okada, H., Crouch, P.: Content-adaptive postfiltering for very low bit rate video. In: Data Compression Conference, 1997. DCC'97. Proceedings, IEEE (1997) 111–120
7. Saad, M.A., Bovik, A.C., Charrier, C.: Blind prediction of natural video quality. IEEE Transactions on Image Processing **23**(3) (2014) 1352–1365
8. Manasa, K., Channappayya, S.S.: An optical flow-based full reference video quality assessment algorithm. IEEE Transactions on Image Processing **25**(6) (2016) 2480–2492
9. Kim, T., Lee, S., Bovik, A.C.: Transfer function model of physiological mechanisms underlying temporal visual discomfort experienced when viewing stereoscopic 3d images. IEEE Transactions on Image Processing **24**(11) (2015) 4335–4347
10. Kim, J., Zeng, H., Ghadiyaram, D., Lee, S., Zhang, L., Bovik, A.C.: Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment. IEEE Signal Processing Magazine **34**(6) (2017) 130–141
11. Seshadrinathan, K., Soundararajan, R., Bovik, A.C., Cormack, L.K.: Study of subjective and objective quality assessment of video. IEEE Transactions on Image Processing **19**(6) (2010) 1427–1441
12. Park, J., Seshadrinathan, K., Lee, S., Bovik, A.C.: Video quality pooling adaptive to perceptual distortion severity. IEEE Transactions on Image Processing **22**(2) (2013) 610–620
13. Vinyals, O., Bengio, S., Kudlur, M.: Order matters: Sequence to sequence for sets. arXiv preprint arXiv:1511.06391 (2015)
14. Yang, J., Ren, P., Zhang, D., Chen, D., Wen, F., Li, H., Hua, G., Yang, J., Li, H., Dai, Y., et al.: Neural aggregation network for video face recognition. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.(CVPR). 2492–2495
15. Graves, A., Wayne, G., Danihelka, I.: Neural turing machines. arXiv preprint arXiv:1410.5401 (2014)
16. Robson, J.: Spatial and temporal contrast-sensitivity functions of the visual system. Josa **56**(8) (1966) 1141–1142
17. Lee, S., Pattichis, M.S., Bovik, A.C.: Foveated video quality assessment. IEEE Transactions on Multimedia **4**(1) (2002) 129–132
18. Lee, S., Pattichis, M.S., Bovik, A.C.: Foveated video compression with optimal rate control. IEEE Transactions on Image Processing **10**(7) (2001) 977–992

19. Legge, G.E., Foley, J.M.: Contrast masking in human vision. Josa **70**(12) (1980) 1458–1471
20. Kim, H., Lee, S., Bovik, A.C.: Saliency prediction on stereoscopic videos. IEEE Transactions on Image Processing **23**(4) (2014) 1476–1490
21. Mittal, A., Saad, M.A., Bovik, A.C.: A completely blind video integrity oracle. IEEE Transactions on Image Processing **25**(1) (2016) 289–300
22. Le Callet, P., Viard-Gaudin, C., Barba, D.: A convolutional neural network approach for objective video quality assessment. IEEE Transactions on Neural Networks **17**(5) (2006) 1316–1327
23. Kim, J., Nguyen, A.D., Lee, S.: Deep CNN-based blind image quality predictor. IEEE Transactions on neural networks and learning systems (99) (2018) 1–14
24. Chandler, D.M., Hemami, S.S.: Vsnr: A wavelet-based visual signal-to-noise ratio for natural images. IEEE Transactions on image processing **16**(9) (2007) 2284–2298
25. Seshadrinathan, K., Bovik, A.C.: Motion tuned spatio-temporal quality assessment of natural videos. IEEE Transactions on image processing **19**(2) (2010) 335–350
26. Li, Y., Po, L.M., Cheung, C.H., Xu, X., Feng, L., Yuan, F., Cheung, K.W.: No-reference video quality assessment with 3d shearlet transform and convolutional neural networks. IEEE Transactions on Circuits and Systems for Video Technology **26**(6) (2016) 1044–1057
27. Daly, S.J.: Visible differences predictor: an algorithm for the assessment of image fidelity. In: Human Vision, Visual Processing, and Digital Display III. Volume 1666., International Society for Optics and Photonics (1992) 2–16
28. Kim, J., Lee, S.: Fully deep blind image quality predictor. IEEE Journal of selected topics in signal processing **11**(1) (2017) 206–220
29. Oh, H., Ahn, S., Kim, J., Lee, S.: Blind deep S3D image quality evaluation via local to global feature aggregation. IEEE Transactions on Image Processing **26**(10) (2017) 4923–4936
30. Ye, P., Kumar, J., Kang, L., Doermann, D.: Unsupervised feature learning framework for no-reference image quality assessment. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.(CVPR), IEEE (2012) 1098–1105
31. : Laboratory of computational perception & image quality, oklahoma state university, csiq video database. [online]. available: http://vision.okstate.edu/?loc=stmad
32. VQEG: Final report from the video quality experts group on the validation of objective models of video quality assessment, phase ii
33. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on image processing **13**(4) (2004) 600–612
34. Sheikh, H.R., Bovik, A.C.: Image information and visual quality. IEEE Transactions on image processing **15**(2) (2006) 430–444
35. Vu, P.V., Vu, C.T., Chandler, D.M.: A spatiotemporal most-apparent-distortion model for video quality assessment. In: Image Processing (ICIP), 2011 18th IEEE International Conference on, IEEE (2011) 2505–2508
36. Vu, P.V., Chandler, D.M.: Vis3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices. Journal of Electronic Imaging **23**(1) (2014) 013016–013016