일반논문 (Regular Paper)
방송공학회논문지 제28권 제6호, 2023년 11월 (JBE Vol.28, No.6, November 2023)
https://doi.org/10.5909/JBE.2023.28.6.753
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

초협대역 대역폭 조건에서 압축된 군용감시 비디오를 위한 완전참조 비디오 화질 평가 방법

A. F. M. Shahab Uddin[a)], 김 동 현[b)], Mst. Sirazam Monira[c)], 최 증 원[b)], 정 태 충[c)], 배 성 호[c)]

Full-Reference Video Quality Assessment for Military Surveillance Video Compressed with Ultra Narrow Bandwidth

A. F. M. Shahab Uddin[a)], Donghyun Kim[b)], Mst. Sirazam Monira[c)], Jeung Won Choi[b)], TaeChoong Chung[c)], and Sung-Ho Bae[c)]

요 약

비디오 품질 평가는 많은 이미지 처리 및 컴퓨터 비전 응용에 필수적이며, 특히 비디오가 전송될 때 서비스 품질을 관리하기 위해 비디오 품질평가가 필수적으로 사용된다. 대역폭이 극히 제한된 군사 통신과 같은 특수한 통신 네트워크가 존재하며, 이러한 초협대 대역폭 전송 시스템에서 전송되는 비디오의 특성은 일반적인 용도의 비디오와 차이가 있다. 특히 이러한 비디오는 특정 객체나 영역에 매우 집중하며, 이러한 특성이 고려된 초 협대역 전송 시스템 기반 객관적 비디오 품질 평가 방법을 개발해야 한다. 본 논문에서는 먼저 고도로 압축된 감시용 비디오와 해당 비디오들의 시각적 품질을 사람들이 평가한 비디오 품질 평가 데이터셋을 제안한다. 그런 다음 완전 참조 비디오 품질 평가를 수행하는 심층신경망을 제안한다. 패치별로 기준 품질 값이 없기 때문에, 제안 방법은 비디오 전체와 해당 주목도 맵을 합성곱 신경망에 전달하여 특징을 추출하고, 이를 회귀기에 전달하여 최종 품질 점수를 추정한다. 제안 방법은 전체 이미지 신호를 이용하여 중요한 특징을 추출하고 전체 품질을 예측하며, 결과적으로 패치별 특징 추출 후 오류 풀링 계층을 거쳐 전체 품질을 예측하는 것보다 더 높은 수용 영역(receptive field)을 확보할 수 있고, 모델에 오류 풀링 단계를 포함할 수 있다. 광범위한 실험 결과를 통해 제안 방법의 성능을 평가하여 높은 예측 성능을 보임을 확인하였다.

Abstract

Video quality assessment has become an essential step in many image processing and computer vision applications, especially when the task is to maintain the quality of service where the video is transmitted from source to destination. In real world, there are some specialized communication networks e.g., military communication, where the bandwidth is extremely limited. Also, the characteristics of the videos that are transmitted in such ultra-narrow band transmission systems, are usually different from general-purpose videos. Specifically, these videos are highly concentrated on some specific objects or regions. Consequently, the existing image or video quality assessment methods are not suitable to predict the quality of this kind of saliency videos and also there is no related dataset available. In this paper, we first propose a video quality assessment dataset that contains carefully collected extremely compressed surveillance videos along with their visual qualities that are evaluated by human. Then we also propose a deep neural network as a full-reference video quality assessment method. Since there are no patch-wise ground truth quality values, the proposed method passes the whole video and its corresponding saliency map to the convolutional neural network to extract the features which is then transferred to the regressor to estimate the final quality score. Instead of patch-wise feature extraction followed by an error pooling layer to predict the overall quality, the proposed method makes use of the full image signal to extract important features and predict the overall quality. This approach helps to achieve a higher receptive field and does not require any error pooling stage. The extensive experimental results validate the prediction performance of the proposed quality assessment method.

Keyword : Subjective Video Quality Assessment, Surveillance Video Quality, Ultra-Low Band Transmission System

Ⅰ. Introduction

Due to the availability of internet connectivity and the recent advancement of high-technology inexpensive devices (e.g., smartphones, tablets, digital cameras, cam-coders, and close-circuit cameras, etc.), video information has become one of the most important and frequently used contents in communication systems. Videos are being used in an enormous way e.g., for entertainment, education, surveillance, marketing, and so on. For past decades, because of the need for high-quality videos, the data rates of videos are getting higher.

However, sometimes it is limited for high-quality video data streaming to acquire enough data rate. Because this large amount of video content demands a huge portion of available bandwidth to be transmitted. However, the bandwidth is very limited and the most expensive resource in any communication system. Especially in a tactical (military) edge network, the edge users and many network subscribers induce a very low allocated bandwidth (i.e., low data rate) for video streaming. However, even with this 'narrow' bandwidth, the users still need 'sufficient' video quality. As a result, the videos should be compressed before transmission to reduce unnecessary bandwidth consumption. This compression is achieved by utilizing a video compression (lossy) technique, consisting of an encoder in the transmitting end that removes some size of the video; and a decoder in the receiving end that reverses the process to restore original information.

But as a side effect of the compression process, the quality of the video becomes degraded. In order to satisfy the end user, maintaining the quality of experience (QoE) is necessary. However, to control the QoE, we first need to know the level of distortion present in the received videos. The best way to estimate the quality of a video is to evaluate it by human observer since they are the ultimate receivers of those videos. However evaluating by humans (subjective evaluation) is an expensive, inconvenient, and time-consuming process. Also, this process is not suitable to be used in an automated system.

Therefore, the objective image quality assessment (IQA) or video quality assessment (VQA) method has become indispensable as an alternative and practical solution that aims to automatically assess the quality of images/videos by mimicking the principle of human visual system (HVS). Here we focus on VQA. In general, IQA/VQA methods can be classified into three categories: (i) full reference VQA (FR-VQA) that requires a distortion-free video that is considered to have the highest quality and used as a reference video; (ii) reduced reference VQA (RR-VQA) that utilizes the partially available information about the reference video; and (iii) no reference VQA (NR-VQA) that does not require any information about the reference video[65].

The first and most simple IQA was mean squared error (MSE) which has been considered as one of the most com-

a) Department of Computer Science and Engineering, Jashore University of Science and Technology, Bangladesh-7408.
b) 국방과학연구소(Agency for Defense Development)
c) 경희대학교(Department of Computer Science and Engineering, Kyung Hee University)
‡ Corresponding Author : 배성호(Sung-Ho Bae)
E-mail: shbae@khu.ac.kr
Tel: +82-31-201-2593
ORCID: https://orcid.org/0000-0003-2677-3186
※ This research was supported in part by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2023-RS-2023-00258649) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation). and in part by the MSIT(Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2023-RS-2023-00259004) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation)", and in part by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2022-00155911, Artificial Intelligence Convergence Innovation Human Resources Development (Kyung Hee University)).
· Manuscript received August 2, 2023; Revised November 7, 2023; Accepted November 7, 2023.

monly used metrics. Since MSE is simple and has many mathematically desirable properties, it has also been used for optimization of image processing algorithms. However, it does not highly correlate with human evaluation[60],[66],[81]. Afterward, many computational IQA [3]-[5],[32],[36],[48],[58],[60]-[62],[72],[78],[80],[85] and VQA [1],[6],[33],[38],[49],[53],[59],[67],[68],[75],[76] methods have been proposed to achieve higher correlation with subjective quality scores. Among them structural similarity index (SSIM) [85] is considered as a milestone work in the IQA domain. It assumes that HVS importantly extracts structural information from image textures in visual perception and effectively uses the contrast/structural information of image signals to estimate the perceived visual quality. After the great success of SSIM, researchers are in search of effective features that can well characterize contrast and/or structural information in image textures to enhance the performance of IQA methods. IQA methods can be adopted to estimate video quality by applying for each frame and the predicted quality can be found by the average score. However, unlike image quality, video quality depends on several other factors e.g., local spatial frequency characteristics (static textures), types of motion (dynamic textures), and various types of artifacts.

As a result, VQA methods additionally utilize the temporal information that leads to more accurate estimates of visual quality[1],[6],[16],[31],[33],[37],[38],[43],[49],[51],[53],[54],[59],[67],[68],[75]-[77]. Recently, by following the tremendous success of deep convolutional neural networks (CNN) in the field of image processing and computer vision, several learning based IQA/VQA models have been introduced[8],[32],[36],[48]. The CNN-based models are trained to learn the visual sensitivity maps and apply weighted pooling to predict the quality scores. One severe problem of those methods is that they use small patches during training where the quality of the whole image/video is used as the quality of that small patch. It is undesirable because the quality of any local region or image patch depends on (i) the type of the region (i.e., homogeneous or complex) and (ii) the type of distortion (noise, blur, blocking artifacts, etc.)[85]. Moreover, no study has been performed (to the best of our knowledge) on the quality assessment of surveillance video streaming services with ultra-narrow band transmission systems and, unfortunately, there is no publicly available dataset related to this study.

Consequently, we systematically collect a surveillance dataset (more details about the dataset preparation can be found in Section III) and measure the mean opinion score (MOS) by human observers following the standard rules of subjective evaluation[19]. Since the surveillance videos are highly concentrated on the target objects and compressed with low bit rate encoder-decoder (CODEC) configuration, their characteristics are different from other videos. As a result, the existing IQA/VQA methods failed to achieve sufficient correlation with human observers. Motivated by the above-mentioned problems, we investigate a more appropriate data-driven VQA method by considering the surveillance video characteristics. Our main contributions are as follows:

We carefully collect a sufficient number of surveillance videos for a wide range of characteristics that are important for quality assessment.

Then we evaluate the quality of the compressed version of the collected videos by human observers following the standard rules provided by [19]. From the subjective scores, we calculate the MOS.

Finally, we propose a new FR-VQA method that employs a fully convolutional neural network with global average pooling to allow any size of input videos.

The proposed method utilizes the full image signal as input to the CNN instead of using patches. As a result, it offers the network a higher receptive field and does not require any pooling stage.

The new FR-VQA method achieves a significantly higher correlation with subjective MOS and thereby outperforms

existing IQA/VQA methods.

The rest of the paper is organized as follows: Section II reviews the related works; Section III explains the systematic way of dataset preparation; Section IV presents the proposed FR-VQA methods; Section V shows the experimental results and finally Section VI concludes the paper.

Ⅱ. Related Works

1. Deep Learning Based IQA/VQA

Due to the recent advancements and great success of deep learning, deep neural networks have become the dominant approach in many computer vision and image processing fields. CNNs have widely been used in image classification[14],[24],[41], object detection[35],[50], semantic segmentation[11],[20], natural scene understanding[71],[73], human pose estimation[57],[70] and so on. However, only a few IQA/VQA methods have adopted CNNs to assess the quality of images or videos. The first attempt was made by Kang et al.[21], where CNN was used to solve an NR-IQA problem instead of using any hand-designed features. The author used a patch-based training process by following the state-of-the- art training mechanism where each patch quality was needed. However, the IQA benchmark datasets do not provide patch-wise ground truth quality scores but a single visual quality value for the whole image. As a result, they assigned all patch quality with the ground truth quality score.

To remedy the patch quality problem, Kim et al.[22] proposed an FR-IQA method where the network takes the distorted image and an error map as input and estimates the sensitivity map which is then used in the weighted pooling stage. Bosse et al.[9] introduced an impressive method that can effectively be used for both the FR-IQA and NR-IQA problems. The network is first trained to predict the quality score and the weight for each patch. At the final stage, the overall quality score of the whole image is estimated by applying a weighted average pooling method. Prashnani et al.[46] reveal that pairwise learning helps to improve IQA performance and based on that they proposed PieAPP which offers much higher prediction accuracy.

To overcome the limited dataset problem, Liu et al.[36] exploited synthetically distorted images with various levels of distortion and used a Siamese network to rank the images based on their quality. And finally, this knowledge is transferred to a traditional CNN to estimate the quality of a single image. Another interesting approach for NR-IQA was proposed in[32]. To compensate for the absence of a reference image, Lin et al.[32] generated a hallucinated reference image and then the pair of the hallucinated reference and corresponding distorted image is passed to the regressor to estimate the quality. Since human is the ultimate receiver of any image or video, it is necessary for the IQA/VQA methods to consider the HVS properties during the quality prediction. As a result, Seo et al.[48] proposed an HVS inspired IQA network that incorporated two important HVS properties named visual saliency (VS) and just noticeable difference (JND).

On the other hand, the Video Quality Metric (VQM)[43] is an early VQA method, that employs losses in the spatial gradients of luminance, along with features based on the product of luminance contrast and motion. The MOVIE index[49] uses the idea of motion tuning by tracking perceptually relevant artifacts along motion trajectories to measure video quality. ST-MAD[59] index uses spatiotemporal video slices and applies the idea of “most apparent distortion”[25] to assess quality. SpEED-VQA[6] computes spatial and temporal entropic differences in a band-pass domain to measure quality deviation.

Recently, data-driven approaches have become increasingly popular because of their high prediction performance. For example, the Video Multi-method Fusion (VMAF)[31] model developed by Netflix is a widely used quality pre-

dictor built on existing IQA/VQA models on features that are combined using a Support Vector Regressor (SVR).

However, the existing methods are developed based on uncompressed images/videos that are not focused on any particular scene or object. As a result, they poorly performed on compressed saliency videos, which suggests the need for a dedicated VQA method.

2. Visual Saliency

The human visual system (HVS) has the capability of automatically focusing on the distinctive regions (called salient regions) in any visual scene. This powerful property of HVS has widely been studied by cognitive scientists. However, due to its excellence, it has recently been adopted in computer science to solve various computer vision problems. In computer vision, the visual saliency detection models aim to simulate this HVS attention mechanism. According to [7], image saliency detection methods can be categorized into bottom-up models which are stimulus-driven, and top-down models which are task-driven. Bottom-up models extract low-level vision features and use visual priors to describe the properties of salient objects.

Visual priors include contrast prior[12], background prior [10],[63],[86], compactness prior[84], etc. Further- more, frequency domain analysis[2], sparse representation[30], cellular automata[74], random walks[10], low-rank recovery[42], and Bayesian theory[28] are also used to achieve image saliency detection. On the other hand, top-down models take advantage of supervised learning strategy, especially, deep learning techniques including hierarchical deep networks[15],[34], multi-scale or multi-context deep network[29],[83], encoder-decoder networks[47],[82], etc.

Among several well-established models, some of the fast and well-known saliency detection algorithms are [2], [18], [39], [47] Hou et al.[18] proposed a spectral residual method that focuses on the background properties instead of the target object's property. This method analyzes the log spectrum of an image, extracts the spectral residual in the spectral domain, and then reconstructs the saliency map in the spatial domain. Achanta et al.[2] proposed a frequency-tuned approach that preserves the boundary information by retaining a sufficient amount of high-frequency contents and produces a full-resolution saliency map. Montabone & Soto introduced a method that was originally designed for fast human detection in a scene by proposing novel features derived from the visual saliency mechanism[39].

Later on, this feature extraction mechanism was generalized for other forms of saliency detection. Qin et al.[47] exploited the power of deep CNN and proposed a boundary-aware saliency detection network (BASNet) that consists of a densely supervised Encoder-Decoder network and a residual refinement module to produce saliency prediction and to perform saliency map refinement, respectively.

3. Quality Assessment Dataset

One of the most important and challenging tasks in IQA/VQA research is to collect the dataset. IQA/VQA dataset needs to provide the ground truth quality values of the collected images or videos. However, human beings are the ultimate receivers of those contents and as a result, it is necessary to evaluate those contents by humans, which is highly time-consuming and cumbersome. LIVE[85], TID2008[45], CSIQ[25], and TID2013[44] are the four benchmark datasets that are widely used to evaluate the IQA/VQA metrics. LIVE[85] consists of 779 distorted images which are generated from 29 reference images using 5 distortion types. 161 subjects evaluated the dataset. CSIQ[25] consists of 866 distorted images which are generated from 30 reference images using 6 distortion types and the dataset is evaluated by 35 subjects.

Another well-known dataset is proposed by Ponomarenko et al.[45] where there are 1700 distorted images which are generated from 25 reference images using 17 distortion types and evaluated by 838 subjects. After that, the author

of TID2008 extended the dataset by introducing 7 more distortion types, resulting in an increased number of distorted images. TID2013[44] consists of 3000 distorted images that are the largest number of images among existing IQA datasets. Those images are generated from 25 reference images using 24 distortion types and evaluated by 917 subjects. Consequently, TID2013[44] has become the most reliable dataset to evaluate IQA/VQA metrics.

However, there is no dataset available for highly concentrated videos that are compressed to use in an ultra-narrow band transmission system. Since, this kind of video usually focuses on a specific target object or region and requires a high compression rate, existing IQA/VQA metrics are not suitable to predict their visual quality. This phenomenon motivated us to collect a new VQA dataset that will help to develop a better quality assessment metric that can be used to automatically estimate the visual quality of ultra-narrow band transmission system contents.

Ⅲ. Dataset Introduction

In the evaluation of a video streaming system, not only IQA/VQA metric but also the selection of test sequences is important. If the type/quality of video sequences is very limited, then the performance evaluation results based on those test sequences cannot achieve good credibility (i.e., it cannot be trusted). The results only will be applicable in limited cases i.e., the videos that are very similar to the selected sequences. Therefore, we collect various types of surveillance video sequences and measure several characteristics including temporal variation, spatial variation, various sharpness metrics, etc., of those collected videos. Based on their characteristics, we shortlist the videos and categorize them to use as representatives of surveillance videos in ultra-narrow-band video transmission systems.

1. Dataset Selection

We collect candidate video sequences for performance evaluation from largely two sources: self-collected video sequences and standard video sequences. In the collection of self-collected videos, we use several types of cameras: IR camera (built by i3 Systems, resolution 320 × 480) and EO cameras: GoPro (built by GoPro Inc., resolution 1920 × 1080). Using these cameras, we capture tremendous battlefield-like surveillance raw video sets (since we cannot acquire “real” battlefield surveillance videos, we organize a pseudo-battlefield environment). Furthermore, to add more videos to our test sequence pool, we additionally collect video sequences from YouTube. In order to maintain the ‘generality of videos’, we collect standard videos provided by high-efficiency video coding (HEVC) and MPEG. Consequently, we have three groups of test sequences based on the source of videos: 1) Standard test sequences; 2) Collected-EO test sequences; and 3) Collected-IR test sequences.

The collected sequence also can be categorized based on pixel variation statistics. We categorize all the collected sequences into two groups: 1) stationary camera sequences; and 2) moving camera sequences. The reason behind categorizing the videos based on camera motion is that we are interested in two types of cameras: 1) a camera installed on a tripod - which will be similar to the stationary camera sequences; and 2) a camera mounted on a helmet - which will be similar to the moving camera sequences. Then the stationary camera sequences are resized to CIF (352 × 288) resolution, clipped to make 10 seconds long of 5 frames per second (fps), and re-encoded to YUV 4:2:0 format of 8-bit color depth. Similarly, the moving camera sequences are resized to QCIF (176 × 144) resolution, clipped to make 10 seconds long of 10 fps, and re-encoded to YUV 4:2:0 format of 8-bit color depth. Each resolution and fps specification for moving and stationary target sequences are driven by the system-level requirement of video transmission.

2. Description and Analysis of the Dataset

According to ITU recommendation, there should be a high variation in the test sequences when evaluating their perceptual quality, since the quality perception usually depends on their spatial-temporal characteristics[55]. Therefore, we analyze candidate test sequences using the following metrics: spatial variation (SV), temporal variation (TV), and sharpness. In

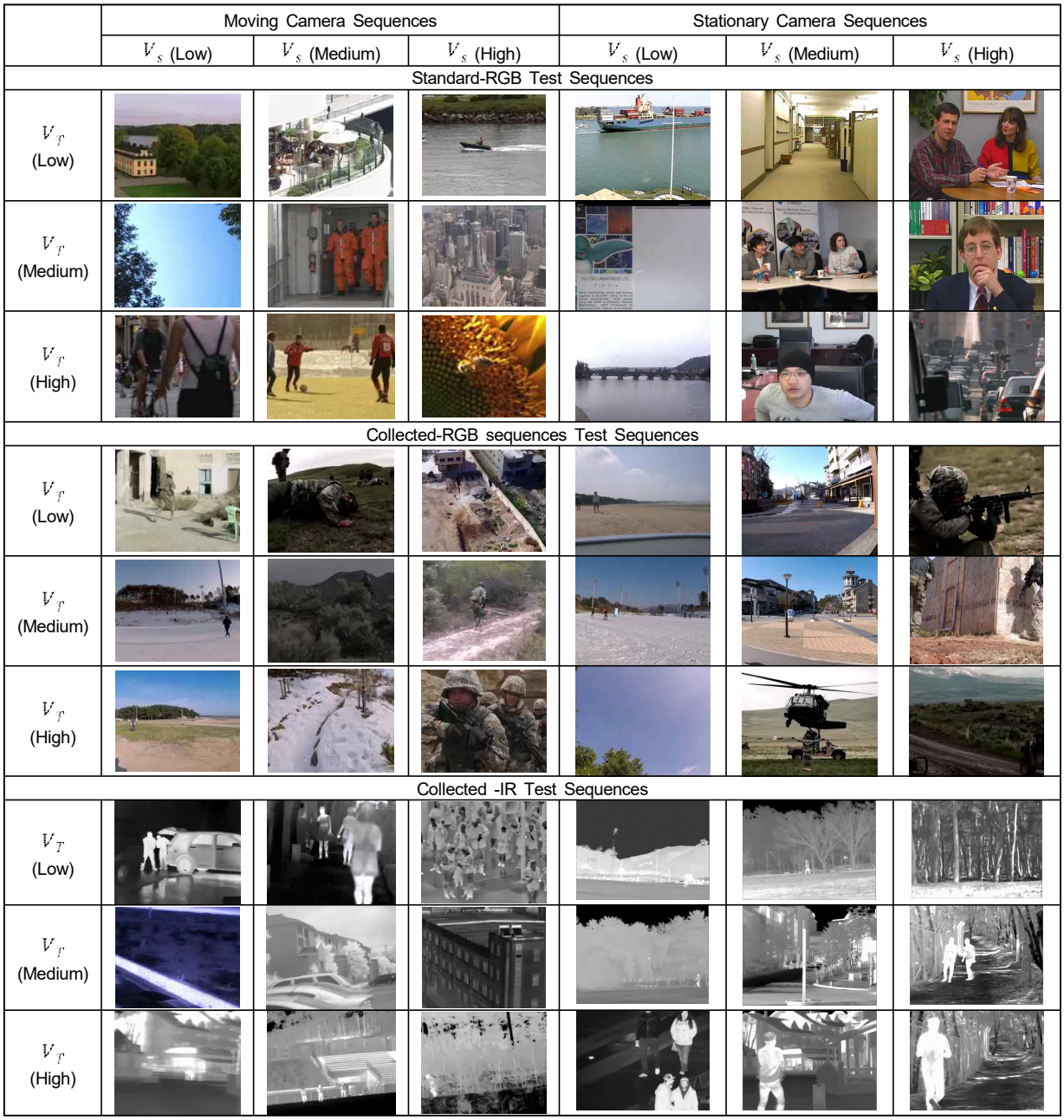

그림 1. 제안 데이터셋의 참조 영상. 데이터셋은 총 9개의 카테고리로 구성, 각 카테고리별로 6개의 비디오가 존재하여 총 54개의 비디오로 구성됨.
Fig. 1. Reference images of the proposed dataset. In each category there are six videos, and for nine categories there are total 54 videos.

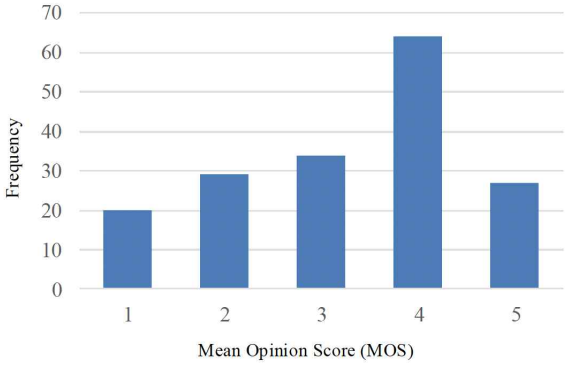

그림 2. 수집된 MOS 값의 히스토그램 분포
Fig. 2. Statistics of the collected MOS values i.e., the score wise histogram

spatial and temporal variation evaluation, the MPEG Edge Histogram Descriptor[69] for SV and the average magnitude of motion vectors[55] for TV are used, respectively. For the sharpness measure, we use cumulative probability of blur detection (CPBD) metric[40]. We first categorize the candidate sequences based on SV and TV, and verify the adequacy of selected sequences based on CPBD. As mentioned in the previous section, we categorize overall collected test sequences into 2 (stationary, moving) by 3 (standard, collected-EO, collected-IR) groups, which results in a total of 6 groups.

For each group and each test sequence, we measure the SV and TV characteristics and based on the measured SV and TV, the video sequences are classified into the following nine categories: LL, ML, LL, ML, HL, LM, MM, HM, LH, MH, HH, where the first letters indicate Low (L), Medium (M) and High (H) spatial complexity and the second letters indicate the temporal complexity in the same manner. After that, from each category, one sequence is selected as a representative of that class, i.e., nine representative videos for each group. For verification of representative video selection, we use several aforementioned sharpness metrics and check the distribution of the selected videos. Figure 3 shows the scatter plots of the selected videos from the perspective of spatial complexity and temporal

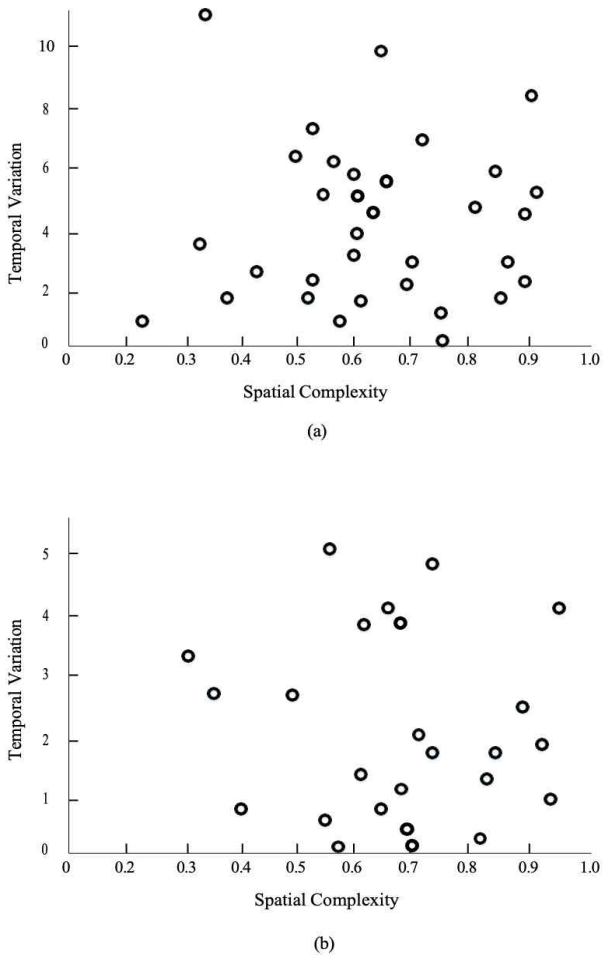

그림 3. 공간 및 시간 복잡도 관점에서 데이터셋의 산점도. (a) 이동 카메라 비디오 샘플의 분포, (b) 정지 카메라 비디오 샘플의 분포.
Fig. 3. Scatter plot of the selected videos in perspective of spatial complexity and temporal variation. (a) Distribution of the moving camera videos. (b) Distribution of the stationary camera videos.

variation. As shown in Figure 3, the videos are well distributed all over the range of spatial complexity and temporal variation.

Therefore, we finally get 6 groups with 9 representative test sequences in each, resulting in 54 selected test sequences. The first frames of each selected test sequence are shown in Figure 1. The selected ‘moving’ sequences show ‘moving camera’ characteristics as we intended, and the ‘stationary’ sequences show ‘stationary camera’ characteristics. It can be seen from

Figure 1 that most of the collected EO videos contain natural scenes (complex patterns) that do not present in the standard sequences. Usually, military operations take place near 'borders' i.e., where man-made things can hardly be found such as buildings, traffic lights, traffic signs, etc. Moreover, usually, military equipment and/or personnel are with camouflage, and therefore their appearances are very similar. As a result, the test sequences need to contain natural scenes. Figure 1 also shows that the IR video sequences have different characteristics from the EO sequences.

However, IR sequences are widely used for military and surveillance tasks, and consequently, we include IR video sequences in this study. Similar to the EO sequences, IR video sequences also contain a large portion of natural scenes.

3. Subjective Video Quality Assessment

Using the selected 54 test sequences, we conduct a subjective video quality assessment study. Subjective quality assessment in a contextual manner is one of the most fundamental and important for video quality assessment. In this paper we perform the subjective quality assessment experiments following the ITU-R. BT500-11[19]. A total of 16 persons participated which satisfied the recommendation[19]. The age range of the participants is between 22 to 35 years. The environmental conditions are satisfied according to the recommendation and we adopt double-stimulus impairment scale (DSIS) method.

We generate impaired sequences for three target bit-rates i.e., 20, 45, and 65 kilobits per second (kbps) for each reference video using the x265 encoding tool[56] where the GOP size is 10 and the number of consecutive B frames is four encoded with the two reference frames. The participants give a score for every impaired sequence within a range of 1 to 5 mainly considering the quality of the target object/region.

Based on the collected scores, we evaluate the Mean

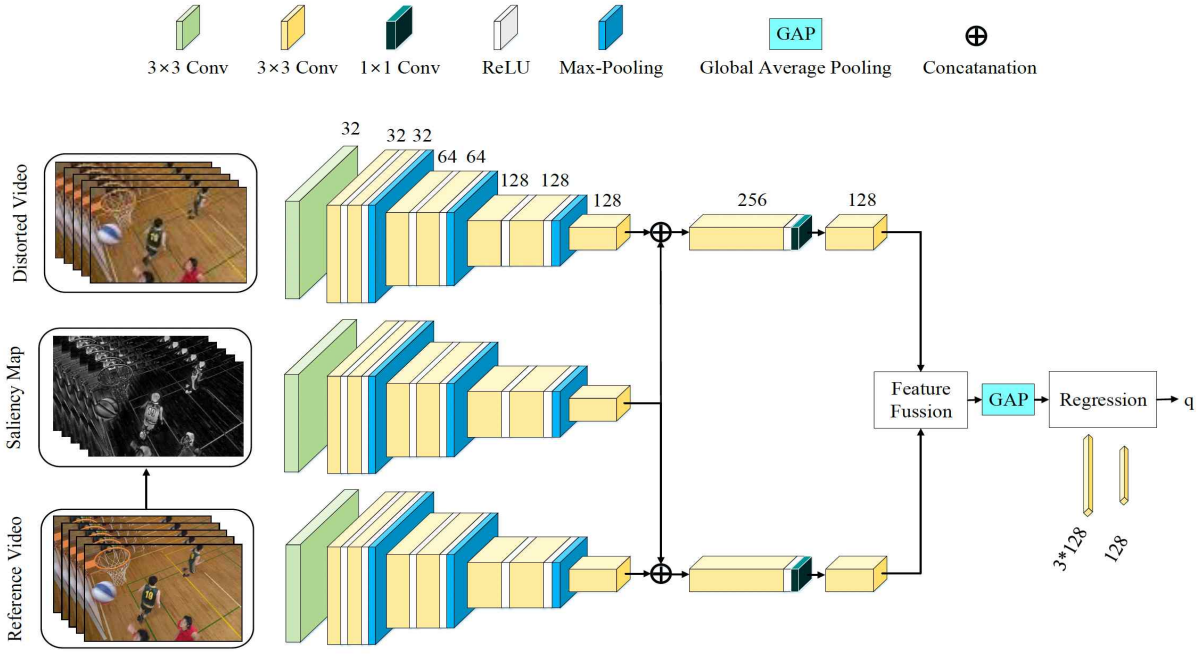

그림 4. 제안하는 심층신경망 구조.
Fig. 4. Graphical representation of the proposed network architecture.

Opinion Score (MOS) value. The outlier scores (which are unnatural compared to others) are distinguished based on z- score method[13], and discarded from MOS evaluation. The z-score measures the extent of deviation from the mean score with the following formula

$$z = \frac{x - \bar{x}}{\sigma}, \quad (1)$$

where $\bar{x}$ and σ are the mean and standard deviation of the data. The z-score ranges from -3σ and 3σ and if any score belongs outside of this range, it is considered as an outlier and removed from the data. After that, the MOS value is calculated as follows:

$$MOS = \frac{1}{N}\sum_{i=1}^{N} S_i, \quad (2)$$

where N is the number of subjects and S_i is the ith subject's score. Statistics about the collected MOS values are shown in Figure 2.

Ⅳ. Proposed Method

1. Data Preprocessing

Since the collected dataset contains videos of YUV color format i.e., the color space is defined by one luminance component (Y) and two chrominance components, called blue projection (U) and red projection (V), respectively, and also the videos are stored as a one-dimensional vector. To feed into a CNN, the videos need to be converted to RGB color space.

Furthermore, we have a quality score neither for each frame nor for any image patch, but rather for the whole video. As a result, we extract all the RGB frames from a YUV video and stack them as a single input, and use its subjective score as the target label for training our network. Since the video frames are in color format i.e., they are of three channels, the dimension becomes W × H × 3N, where W, H, and N represent the width, height, and number of frames in a video. It is worth noting that patch-based quality assessment methods assign each patch quality score similar to the quality of the whole image or video, which is undesired as described in Section I. Using the whole image frame instead of the image patch offers two benefits: it does not require any patch quality score and pooling stage at the end of the network (see Section IV-C for more details).

2. Visual Saliency Detection

We apply a well-recognized and widely accepted saliency detection algorithm proposed by Montabone & Soto[39] to estimate the saliency map. It is not restricted to[39], rather any saliency detection algorithm can be used here (see Section V-C for more details). Since, Montabone & Soto[39] is a state-of-the-art method and easily available as a OpenCV library, we incorporate this saliency detection in our method. First, we extract the saliency maps of the reference video frames and then stack them together, similar to the video input. Since the saliency maps are gray scale image, the dimension becomes W × H × N, where W, H and N represent the width, height and the number of frames in a video.

3. Network Architecture

Our proposed network consists of three main parts i.e., a video feature extraction part; a saliency feature extraction part and the final quality estimation part. We use a Siamese network architecture that has two branches of networks but with shared parameters to extract the features of the reference and distorted video. This kind of network is popular where the goal is to find out the similarities/differences between two different inputs. On the other hand, the saliency feature extraction module takes the corresponding saliency

map and extracts the important features. The joint effect of video and saliency feature extraction offers a more powerful feature ex- traction that enables the network to extract more meaningful, significant, and effective feature maps. Figure 4 shows the architecture of the proposed network. It is noted that one video clip consists of three channels and a saliency map which has a single channel.

Therefore, we apply a 3 × 3 convolution on the input video where the input channel is 3 and the output channel is 32. Similarly, for the input saliency map, we apply 3 × 3 convolution with input channel 1 and output channel 32. The features ($href$ and $hdist$) are extracted in a series of conv3-32, conv3-32, maxpool, conv3-64, conv3-64, maxpool, conv3-128, conv3-128, maxpool layers where conv3-64 means the convolution layer with total 64 filters sized of 3 × 3 and the maxpool layers perform with the stride value of 2. The extracted features are then passed to the feature fusion stage. The feature fusion first calculates the differences between reference and distorted features $(href - hdist)$ and then concatenates this difference with reference and distorted features as $cat((href - hdist), href, hdist)$.

This fused feature map now needs to be passed to the regression module. It is worth noting that the input videos are of different resolutions and the whole video (instead of patch) is considered as input. These various resolutions of input will produce feature maps with different sizes, which are not allowed to feed into the fully connected layer-based regression module. To address this problem, we apply global average pooling (GAP) on the fused features before feeding them to the fully connected layer for quality estimation. The regression module is made up of a sequence of two fully connected layers i.e., one FC-3*128 and one FC-128 layer. It is worth noting that the proposed model does not require any pooling stage. Because, unlike other IQA/VQA methods, we utilize the full image/video as input instead of a patch. Therefore, we directly estimate the overall quality of the given input rather than estimating each patch's quality.

This results in about 5.2 million trainable network parameters of the proposed CNN. All convolutional layers apply 3×3 pixel size convolution kernels and are activated through a rectified linear unit (ReLU) activation function g = max (0, x)[17]. To obtain an output of the same size as the input, convolutions are applied with zero padding. All max-pool layers have 2 × 2 pixel-sized kernels. Dropout regularization with a ratio of 0.5 is applied to the fully connected layers in order to prevent overfitting[52].

4. Training Details

The proposed networks are trained iteratively by the standard error back-propagation[26],[27] over a number of epochs, where one epoch is defined as the period during which all the samples from the training set have been used. It is noted that since the videos are of different sizes, batch-wise training is not possible. The learning rate for the optimization is controlled per parameter adaptively using the ADAM method[23] based on the variance of the gradient. Parameters of ADAM are chosen as recommended in [40] as $\beta 1 = 0.9$, $\beta 2 = 0.999$, $E = 10^{-8}$ and $\alpha = 10^{-4}$. The mean loss over all images during validation is computed in evaluation mode (i.e. dropout is replaced with scaling) after each epoch. The final model that is used in the evaluation is the one with the best validation loss.

Ⅴ. Experimental Results

1. Experimental Setup

The overall samples in the dataset are randomly split by 70% training and 30% test set. This split is done by reference video to make sure that no test video has been seen by the network during training. Results are reported based on 5 random splits. Models are trained for 500 epochs. In order to evaluate the proposed method, four well-known and widely accepted evaluation metrics are used i.e.,

Spearman Rank-Order Correlation coefficient (SROC), Kendall Rank-Order Correlation coefficients (KROC), Pearson Linear Correlation Coefficient (PLCC), and Root Mean Square Error (RMSE). The SROC, KROC, and PLCC are used to measure the monotonicity between MOS and their estimated values by the IQA/VQA methods, while RMSE is used to measure the prediction error. Usually, SROC and KROC are the representative performance measures in IQA/VQA methods. Due to the non-linear mapping function between the estimated IQA/VQA scores and their measured subjective visual quality scores, we adopt a logistic regression when measuring the PLCC and RMSE. For SROC, KROC, and PLCC, the higher the score the better the quality. While for RMSE, the lower the score the better the quality. It is worth noting that, in order to compare the existing IQA methods for VQA we apply the IQA method for each frame and then take the average to estimate the final quality score.

2. Performance Comparison

Table 1 compares the performances of existing FR-IQA methods with the proposed FR-VQA method. Here the per-formance is compared in terms of prediction accuracy.

표 1. 제안 VQA 방법과 기존 IQA 방법 간 성능비교
Table 1. Performance comparison of the proposed VQA method with state-of-the-art IQA methods.

Metric	SROCC	KROCC	PLCC	RMSE
PSNR	0.5730	0.4128	0.6188	0.7821
SSIM[85]	0.5730	0.4128	0.6188	0.7821
UQI[64]	0.6028	0.4369	0.6099	0.7888
RFSIM[79]	0.5667	0.4092	0.5959	0.7994
GMSD[72]	0.7446	0.5535	0.7868	0.6145
VSI[78]	0.6453	0.4676	0.6821	0.7279
SCQI[5]	0.6191	0.4574	0.6543	0.7528
VSSCQI[58]	0.6074	0.4469	0.6452	0.7601
WaDIQaM-FR[9]	0.6834	0.5053	0.7232	0.7501
Our	0.8670	0.6790	0.8713	0.4676

The best results are highlighted using black boldface text. It can be seen that all other methods except the proposed method achieved low prediction accuracy when applied to the collected saliency videos. This is because the characteristics of those videos are significantly different from the characteristics of existing benchmark datasets. Since the proposed method employs deep CNN to extract the meaningful features and also utilizes the visual saliency of highly important regions, it offers a higher prediction map of the distorted videos to help the model focus on accuracy in terms of SROCC, KROCC, PLCC, and RMSE. The prediction accuracy of the proposed method is significantly higher than all other methods in comparison. Compared to the nearest competitor GMSD[72], the proposed method achieves 0.12%, 0.13%, 0.09%, and 0.15% higher accuracy in terms of SROCC, KROCC, PLCC, and RMSE, respectively.

Table 2 compares the performances of existing FR-VQA methods with the proposed FR-VQA method. For comparison, we collected the most representative and publicly available FR-VQA methods, i.e., DVQA[54], GREED[37], and VMAF[31]. Here the performance is also compared in terms of prediction accuracy. All the methods that are compared here are developed for video quality estimation. According to the results, the existing VQA methods poorly estimate the video quality for the proposed dataset. Among them, DVQA[37] and GREED[37] offer poor prediction performance, and VMAF[31] offers rela-tively competitive pre-

표 2. 제안 VQA 방법과 기존 VQA 방법 간 성능비교
Table 2. Performance comparison of the proposed VQA method with state-of-the-art VQA methods.

Metric	SROCC	KROCC	PLCC	RMSE
DVQA[54]	0.5463	0.3844	0.5426	3.9130
GREED[37]	0.5420	0.3671	0.5215	4.2994
VMAF[31]	0.7680	0.5687	0.7933	0.8696
Our	0.8670	0.6790	0.8713	0.4676

higher accuracy in terms of SROCC, KROCC, PLCC, and RMSE, respectively.

diction performance. The proposed method outperforms all other methods in comparison. Compared to the nearest competitor VMAF[31], the proposed method achieves 0.10%, 0.11%, 0.08%, and 0.40%.

3. Effect of Various Saliency Detection Algorithms

Our network takes the VS map of the input video and those VS maps are generated by using a third-party VS detection method. Since the quality of the VS map varies from method to method, here we investigate the effect of incorporating different saliency detection methods in our proposed VQA. We use four well-recognized saliency detection algorithms i.e., Montabone & soto[39], Spectral Residual[18], Frequency Tuned[2], and BSANet[47] to extract the VS map of the training videos and use those VS maps during experiments. Figure 5 shows the performance comparison when the proposed VQA method is applied with VS maps extracted by various saliency detection algorithms. The results suggest that the effect of the quality of generated VS maps varies by a small point. And since the network has a sufficient number of learnable parameters, it can compensate the VS map quality variation. As a result, the network is capable of offering almost similar prediction accuracy despite of the Saliency detection algorithm used in the model. However, we have used[39] to extract the saliency map because it is well accepted and easily available as OpenCV library.

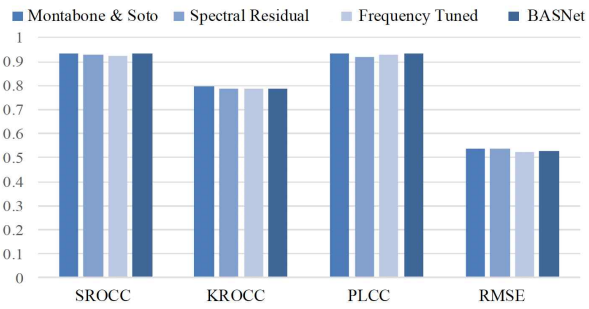

그림 5. 제안 VQA 방법에 적용되는 다양한 관심영역 검출 알고리즘 효과 비교
Fig. 5. Effects of various Saliency detection algorithms on the proposed VQA method

4. Investigating Possible Fusion Styles

As shown in Figure 4, the proposed network consists of two branches, one is to extract features of the input video and another one is to extract features of VS map of the corresponding input video. Those two branches of the network share identical configurations and the output of each branch is concatenated before passing to the regression network. However, this concatenation may take place in the early stage, middle stage, or later stage. To find out the optimum place to concatenate the features, we performed experiments with early stage fusion, middle stage fusion, and later stage fusion. The experimental results are shown in Table 3. It can be seen that the prediction accuracy is poor for early stage fusion. This happened since the characteristics of input video and its corresponding VS map are significantly different and when we first concatenate them and feed them into the network, the model fails to get meaningful representations. Middle and later stage fusion works as a feature attention mechanism at feature level. Because VS map highlights the important pixels of an image or video and mixing this information with input video features followed by convolution helps the model to focus on significant pixels, resulting in higher prediction accuracy. However, the later stage fusion has a better ability to extract important features since it contains multiple layers of convolutions and thereby offers the highest prediction accuracy.

표 3. 제안 방법의 퓨전 방식에 따른 성능 비교
Table 3. Performance comparison of the proposed method with different fusion styles

Metric	SROCC	KROCC	PLCC	RMSE
Early stage fusion	0.7384	0.5880	0.8102	0.6340
Middle stage fusion	0.7919	0.6127	0.8217	0.5561
Later stage fusion	0.8670	0.6790	0.8713	0.4676

Ⅵ. Conclusion

In this study, we first collected a sufficient number of videos that are highly concentrated on some specific objects or regions. The selection process follows standards so that the collected videos contain a wide range of characteristics that play significant roles during the visual quality evaluation. We then encoded the test video sequences according to the limited allocated resources and evaluated them by human subjects since they are the ultimate receivers. After that we reveal that the existing IQA or VQA methods poorly perform to estimate the quality of those encoded test sequences i.e., the correlation between the objective quality scores and the MOS is low. So, we propose a new FR-VQA method that predicts the visual quality by focusing on salient objects or regions of the input video. The proposed method utilizes the whole video rather than patches since there are no patch-wise ground truth quality values and further, this helps to discard the error pooling stage from the network. The proposed method outperforms the existing IQA or VQA methods in terms of the representative metrics i.e., SROCC, KROCC, PLCC, and RMSE.

참 고 문 헌 (References)

[1] M. A. Aabed, G. Kwon, and G. AlRegib, “Power of tempospatially unified spectral density for perceptual video quality assessment,” *In 2017 IEEE International Conference on Multimedia and Expo (ICME),* pages 1476 – 1481, 2017.

[2] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, “Frequency-tuned salient region detection,” *In 2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1597 – 1604, June 2009.

[3] S. Bae and M. Kim, “A novel dct-based jnd model for luminance adaptation effect in dct frequency,” *IEEE Signal Processing Letters*, Vol. 20, No. 9, pp. 893 – 896, 2013.
doi: https://doi.org/10.1109/LSP.2013.2272193.

[4] S. Bae and M. Kim, “A novel generalized dct-based jnd profile based on an elaborate cm-jnd model for variable block-sized transforms in monochrome images,” *IEEE Transactions on Image processing*, Vol. 23, No. 8, pp. 3227 – 3240, 2014.
doi: https://doi.org/10.1109/TIP.2014.2327808.

[5] S. Bae and M. Kim, “A novel image quality assessment with globally and locally consilient visual quality perception,” *IEEE Transactions on Image Processing*, Vol. 25, No. 5, pp. 2392 – 2406, 2016.
doi: https://doi.org/10.1109/TIP.2016.2545863.

[6] C. G. Bampis, P. Gupta, R. Soundararajan, and A. C. Bovik, “Speed-qa: Spatial efficient entropic differencing for image and video quality,” *IEEE Signal Processing Letters*, Vol. 24, No. 9, pp. 1333 – 1337, 2017.
doi: https://doi.org/10.1109/LSP.2017.2726542.

[7] A. Borji, M. Cheng, H. Jiang, and J. Li, “Salient object detection: A benchmark,” *IEEE Transactions on Image Processing*, Vol. 24, No. 12, pp. 5706 – 5722, 2015.
doi: https://doi.org/10.1109/TIP.2015.2487833.

[8] S. Bosse, D. Maniry, K. Müller, T. Wiegand, and W. Samek, “Deep neural networks for no-reference and full-reference image quality assessment,” *IEEE Transactions on Image Processing*, Vol. 27, No. 1, pp. 206 – 219, 2018.
doi: https://doi.org/10.1109/TIP.2017.2760518.

[9] Changyang Li, Yuchen Yuan, Weidong Cai, Yong Xia, and David Dagan Feng, “Robust saliency detection via regularized random walks ranking,” In 2015 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2710 – 2717, 2015.

[10] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 40, No. 4, pp. 834 – 848, April 2018.
doi: https://doi.org/10.1109/TPAMI.2017.2699184.

[11] M. Cheng, G. Zhang, N. J. Mitra, X. Huang, and S. Hu, “Global contrast based salient region detection,” *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 409 – 416, 2011.

[12] David Clark-Carter, “Z scores,” in “Wiley StatsRef: Statistics Reference Online,” 2014.
doi: https://doi.org/10.1002/9781118445112.stat06236.

[13] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770 – 778, June 2016.

[14] Shengfeng He, Rynson Lau, Wenxi Liu, Zhe Huang, and Qingxiong Yang, “Supercnn: A superpixelwise convolutional neural network for salient object detection,” *International Journal of Computer Vision*, Vol. 115, No. 03, pp. 330-344, 2015.
doi: https://doi.org/10.1007/s11263-015-0822-0.

[15] A.P. Hekstra, J.G. Beerends, D. Ledermann, F.E. de Caluwe, S. Kohler, R.H. Koenen, S. Rihs, M. Ehrsam, and D. Schlauss, “Pvqm – a perceptual video quality measure,” *Signal Processing: Image Communication*, Vol. 17, No. 10, pp. 781 – 798, 2002.
doi: https://doi.org/10.1016/S0923-5965(02)00056-5.

[16] Vinod Nair, Geoffrey E. Hinton, “Rectified linear units improve restricted boltzmann machines vinod nair,” *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, June 21-24, 2010, Haifa, Israel.

[17] X. Hou and L. Zhang, “Saliency detection: A spectral residual approach,” *In 2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1 – 8, June 2007.

[18] ITU-R BT.500-11, “Methodology for the subjective assessment of the

quality of tv pictures," *International Telecommunications Union*, June 2012.

[19] Long Jonathan, Shelhamer Evan, and Darrell Trevor, "Fully convolutional networks for semantic segmentation," In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431 – 3440, June 2015.

[20] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," In 2014 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1733 – 1740, 2014.

[21] J. Kim and S. Lee, "Deep learning of human visual sensitivity in image quality assessment framework," In 2017 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1969 – 1977, 2017.

[22] Diederik Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 2014.

[23] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, "Imagenet classification with deep convolutional neural networks," *In Neural Information Processing Systems (NeurIPS)*, pages 1097 – 1105, January 2012.

[24] Eric Larson and Damon Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, Vol. 19, No. 1, pp. 011006, 2010.
doi: http://dx.doi.org/10.1117/1.3267105.

[25] Yann Lecun, Leon Bottou, Y. Bengio, and Patrick Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE,* Vol. 86, No. 11, pp.2278 – 2324, 1998.
doi: https://doi.org/10.1109/5.726791.

[26] Yann LeCun, Léon Bottou, Genevieve B. Orr, and Klaus-Robert Müller, "Efficient backprop," *In Neural Networks: Tricks of the Trade*, 1524, 9-50, 1998.
doi: https://doi.org/10.1007/3-540-49430-8_2.

[27] J. Lei, B. Wang, Y. Fang, W. Lin, P. L. Callet, N. Ling, and C. Hou, "A universal framework for salient object detection," *IEEE Transactions on Multimedia*, Vol. 18, No. 9, pp. 1783 – 1795, 2016.
doi: https://doi.org/10.1109/TMM.2016.2592325.

[28] G. Li and Y. Yu, "Deep contrast learning for salient object detection," In 2016 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 478 – 487, 2016.

[29] X. Li, H. Lu, L. Zhang, X. Ruan, and M. Yang, "Saliency detection via dense and sparse reconstruction," *In 2013 IEEE International Conference on Computer Vision,* pages 2976 – 2983, 2013.

[30] Zhi Li, Anne Aaron, Ioannis Katsavounidis, Anush Moorthy, and Megha Manohara, "Toward a practical perceptual video quality metric," *The Netflix Tech Blog*, Vol. 6, No. 2, 2016.

[31] Kwan-Yee Lin and Guanxiang Wang, "Hallucinated-IQA: No-reference image quality assessment via adversarial learning," *CoRR, abs/1804.01681,* 2018.

[32] Weisi Lin and C. C. Jay Kuo, "Perceptual visual quality metrics: A survey," *Journal of Visual Communication and Image Representation*, Vol. 22, No. 4, pp. 297 – 312, 2011.
doi: https://doi.org/10.1016/j.jvcir.2011.01.005.

[33] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," *In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pages 678 – 686, 2016.

[34] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C, "Berg. SSD: Single shot multibox detector," *Lecture Notes in Computer Science*, page 21 – 37, 2016.

[35] Xialei Liu, Joost van de Weijer, and Andrew D. Bagdanov, "RankIQA: Learning from rankings for no-reference image quality assessment," *CoRR, abs/1707.08347*, 2017.

[36] Pavan C Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C Bovik, "St-GREED: Space-time generalized entropic differences for frame rate dependent video quality prediction," *arXiv preprint arXiv:2010.13715*, 2020.

[37] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity oracle," *IEEE Transactions on Image Processing,* Vol. 25, No. 1, pp. 289 – 300, 2016.
doi: https://doi.org/10.1109/TIP.2015.2502725.

[38] Sebastian Montabone and Alvaro Soto, "Human detection using a mobile platform and novel features derived from a visual saliency mechanism," *Image and Vision Computing*, Vol. 28, No. 3, pp. 391 – 402, 2010.
doi: https://doi.org/10.1016/j.imavis.2009.06.006.

[39] Niranjan Narvekar and Lina Karam, "A no-reference image blur metric based on the cumulative probability of blur detection (CPBD)," *IEEE Transactions on Image Processing*, Vol. 20, No. 9, pp. 2678 – 2683, 2011.
doi. https://doi.org/10.1109/TIP.2011.2131660.

[40] Russakovsky Olga, Deng Jia, Su Hao, Krause Jonathan, Satheesh Sanjeev, Ma Sean, Huang Zhiheng, Karpathy Andrej, Khosla Aditya, Bernstein Michael, C. Berg Alexander, and Fei-Fei Li, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, Vol. 115, No. 3, pp. 211 – 252, 2015.
doi: https://doi.org/10.1007/s11263-015-0816-y.

[41] H. Peng, B. Li, H. Ling, W. Hu, W. Xiong, and S. J. Maybank, "Salient object detection via structured matrix decomposition," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 39, No. 4, pp. 818 – 832, 2017.
doi: https://doi.org/10.1109/TPAMI.2016.2562626.

[42] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transactions on Broadcasting,* Vol. 50, No. 3, pp. 312 – 322, 2004.
doi: https://doi.org/10.1109/TBC.2004.834028.

[43] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C. J. Kuo, "Color image database TID 2013: Peculiarities and preliminary results," *In European Workshop on Visual Information Processing (EUVIP),* pages 106 – 111, 2013.

[44] Nikolay Ponomarenko, Vladimir Lukin, Alexander Zelensky, Karen Egiazarian, Marco Carli, and Federica Battisti, TID 2008 - a database for evaluation of full-reference visual quality assessment metrics," *Advances of Modern Radioelectronics*, Vol. 10, pp. 30 – 45, 2009.

[45] Ekta Prashnani, Hong Cai, Yasamin Mostofi, and Pradeep Sen, "PieAPP: Perceptual image-error assessment through pairwise preference," *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1808 – 1817, 2018.

[46] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood De-hghan, and Martin Jagersand, "BASNet: Boundary-aware salient object detection," *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[47] Soomin Seo, Sehwan Ki, and Munchurl Kim, "Deep HVS-IQANet: Human visual system inspired deep image quality assessment networks," *CoRR, abs/1902.05316*, 2019.

[48] Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Transactions on Image Processing*, Vol. 19, No. 2, pp. 335 – 350, 2010.
doi: https://doi.org/10.1109/TIP.2009.2034992.

[49] Ren Shaoqing, He Kaiming, Girshick Ross, and Sun Jian, "Faster RCNN: Towards real-time object detection with region proposal networks," *In Neural Information Processing Systems (NeurIPS),* 2015.

[50] R. Soundararajan and A. C. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Transactions on Circuits and Systems for Video Technology,* Vol. 23, No. 4, pp. 684 – 694, 2013.
doi: https://doi.org/10.1109/TCSVT.2012.2214933.

[51] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, Vol. 15, No. 56, pp. 1929 – 1958, 2014.
doi: http://jmlr.org/papers/v15/srivastava14a.html.

[52] Peining Tao and Ahmet Eskicioglu, "Video quality assesment using m-svd," *Proceedings of SPIE - The International Society for Optical Engineering*, 6494, 01 2007.

[53] Qi Zheng and Zhengzhong Tu and Pavan C. Madhusudana and Xiaoyang Zeng and Alan C. Bovik and Yibo Fan, "FAVER: Blind Quality Prediction of Variable Frame Rate Videos", *arXiv preprint arXiv: 2201.01492*, 2022.

[54] ITU-R BT.1210, "Test Materials to Be Used in Subjective Assessment of picture quality," *International Telecommunications Union*, 02 2004.

[55] Suramya Tomar, "Converting video formats with ffmpeg," *Linux Journal*, Vol. 2006, No. 146, 2006.
doi: https://dl.acm.org/doi/10.5555/1134782.1134792.

[56] Alexander Toshev and Christian Szegedy, "DeepPose: Human pose estimation via deep neural networks," *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653 – 1660, 2014.

[57] A. F. M. Shahab. Uddin, T. C. Chung, and S. Bae, "Visual saliency based structural contrast quality index," *Electronics Letters*, vol. 55, No. 4, pp. 194 – 196, 2019.
doi: https://doi.org/10.1049/el.2018.6435.

[58] P. V. Vu, C. T. Vu, and D. M. Chandler, "A spatiotemporal most-apparent-distortion model for video quality assessment," *In 2011 18th IEEE International Conference on Image Processing,* pages 2505 – 2508, 2011.

[59] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *IEEE Signal Processing Magazine*, Vol. 26, No. 1, pp. 98 – 117, 2009.
doi: https://doi.org/10.1109/MSP.2008.930649.

[60] Z. Wang and Q. Li, "Information content weighting for perceptual image quality assessment," *IEEE Transactions on Image Processing*, Vol. 20, No. 5, pp. 1185 – 1198, 2011.
doi: https://doi.org/10.1109/TIP.2010.2092435.

[61] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," *In The Thrity-Seventh Asilomar Conference on Signals, Systems Computers*, 2003.
doi: https://doi.org/10.1109/ACSSC.2003.1292216.

[62] Z. Wang, D. Xiang, S. Hou, and F. Wu, "Background-driven salient object detection," *IEEE Transactions on Multimedia*, Vol. 19, No. 4, pp. 750 – 762, 2017.
doi: https://doi.org/10.1109/TMM.2016.2636739.

[63] Zhou Wang and Alan C Bovik, "A universal image quality index," *IEEE signal processing letters*, Vol. 9, No. 3, 81 – 84, 2002.
doi: https://doi.org/10.1109/97.995823.

[64] Zhou Wang and Alan C Bovik, "Modern image quality assessment," *Synthesis Lectures on Image, Video, and Multimedia Processing,* Vol. 2, No. 1, pp. 1 – 156, 2006.
doi: http://dx.doi.org/10.2200/S00010ED1V01Y200508IVM003

[65] Zhou Wang, Ligang Lu, and Alan C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Processing: Image Communication*, Vol. 19, No. 2, pp. 121 – 132, 2004.
doi: https://doi.org/10.1016/S0923-5965(03)00076-6.

[66] Zhenyu Wei and King Ngan, "Spatio-temporal just noticeable distortion profile for grey scale image/video in DCT domain," *IEEE Transactions on Circuits and Systems for Video Technology,* Vol. 19, No. 3, pp. 337 – 346, 2009.
doi: https://doi.org/10.1109/TCSVT.2009.2013518.

[67] Chee Won, Dong Park, and Soo-Jun Park, "Efficient use of MPEG-7 edge histogram descriptor," *Etri Journal,* Vol. 24, No. 1, pp. 23 – 30, 2002.
doi: http://dx.doi.org/10.4218/etrij.02.0102.0103

[68] Bin Xiao, Haiping Wu, and Yichen Wei, "Simple baselines for human pose estimation and tracking," *In Proceedings of the European conference on computer vision (ECCV)*, pages 466 – 481, 2018.

[69] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun, "Unified perceptual parsing for scene understanding," *In Proceedings of the European Conference on Computer Vision (ECCV)*, pages 418 – 434, 2018.

[70] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Transactions on Image Processing,* Vol. 23, No. 2, pp. 684 – 695, 2014.
doi: https://doi.org/10.1109/TIP.2013.2293423.

[71] S. Yang, W. Wang, C. Liu, and W. Deng, "Scene understanding in deep learning-based end-to-end controllers for autonomous vehicles," *IEEE Transactions on Systems, Man, and Cybernetics: Systems,* Vol. 49, No. 1, pp. 53 – 63, Jan 2019.
doi: https://doi.org/10.1109/TSMC.2018.2868372.

[72] Yao Qin, Huchuan Lu, Yiqun Xu, and He Wang, "Saliency detection via cellular automata," *In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pages 110 – 119, 2015.

[73] F. Zhang and D. R. Bull, "A perception-based hybrid model for video quality assessment," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 26, No. 6, pp. 1017 – 1028, 2016.
doi: https://doi.org/10.1109/TCSVT.2015.2428551.

[74] Fan Zhang and David Bull, "Quality assessment methods for perceptual video compression," *arXiv preprint arXiv: 2106.08124,* 2021. https://arxiv.org/abs/2106.08124.
[75] Fan Zhang and David R. Bull, "Chapter 7 - Measuring video quality," *Academic Press Library in signal Processing*, Vol. 5, pp. 227 – 249. Elsevier, 2014.
[76] L. Zhang, Y. Shen, and H. Li, "VSI: A visual saliency-induced index for perceptual image quality assessment," *IEEE Transactions on Image Processing,* Vol. 23, No. 10, pp. 4270 – 4281, 2014. doi: https://doi.org/10.1109/TIP.2014.2346028.
[77] L. Zhang, L. Zhang, and X. Mou, "RFSIM: A feature based image quality assessment metric using Riesz transforms," *In 2010 IEEE International Conference on Image Processing,* pages 321 – 324, 2010. doi: https://doi.org/10.1109/ICIP.2010.5649275.
[78] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, Vol. 20, No. 8, pp. 2378 – 2386, 2011. doi: https://doi.org/10.1109/TIP.2011.2109730.
[79] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "A comprehensive evaluation of full reference image quality assessment algorithms," *In 2012 19th IEEE International Conference on Image Processing,* pages 1477 – 1480, 2012.
[80] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Baocai Yin, "Learning uncertain convolutional features for accurate saliency detection," *In 2017 IEEE International Conference on Computer Vision (ICCV),* pages 212 – 221, 2017.
[81] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," *In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pages 1265 – 1274, 2015.
[82] L. Zhou, Z. Yang, Q. Yuan, Z. Zhou, and D. Hu, "Salient region detection via integrating diffusion-based compactness and local contrast," *IEEE Transactions on Image Processing*, Vol. 24, No. 11, pp. 3308 – 3320, 2015. doi: https://doi.org/10.1109/TIP.2015.2438546.
[83] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing,* Vol. 13, No. 4, pp. 600 – 612, 2004. doi: https://doi.org/10.1109/TIP.2003.819861.
[84] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," In 2014 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2814 – 2821, 2014.
[85] Z. Sinno and A.C. Bovik, "Large-Scale Study of Perceptual Video Quality," *IEEE Transactions on Image Processing*, vol. 28, no. 2, pp. 612-627, February 2019.

저 자 소 개

A. F. M. Shahab Uddin

- 2015년 : 방글라데시 Islamic University, Information and Communication Engineering 학사
- 2016년 : 방글라데시 Islamic University, Information and Communication Engineering 석사
- 2017년 9월 ~ 현재 : 경희대학교 박사과정
- ORCID : https://orcid.org/0000-0003-1074-0515
- 주관심분야 : 이미지 품질 측정, 지각적 이미지 처리, 이미지 처리에서의 딥 러닝 및 역 문제

김 동 현

- 2009년 2월 : 연세대학교 전기전자공학부 학사
- 2011년 2월 : 연세대학교 전기전자공학부 석사
- 2011년 6월 ~ 현재 : 국방과학연구소 제2기술연구본부
- ORCID : https://orcid.org/0000-0002-2136-5944
- 주관심분야 : 전술통신, 영상전송시스템, 데이터링크

Mst. Sirazam Monira

- 2015년 : 방글라데시 Islamic University, Information and Communication Engineering 학사
- 2016년 : 방글라데시 Islamic University, Information and Communication Engineering 석사
- 2017년 9월 ~ 현재 : 경희대학교 석사과정
- ORCID : https://orcid.org/0000-0001-6932-5557
- 주관심분야 : 이미지 품질 측정, 딥 러닝 기반 이미지 처리, 데이터 확장

저 자 소 개

최 증 원

- 1989년 2월 : 충남대학교 계산통계학과 학사
- 1993년 8월 : 충남대학교 계산통계학과(전산학) 석사
- 1997년 8월 : 충남대학교 전산학과 박사
- 1997년 7월 ~ 현재 : 국방과학연구소 수석연구원
- 2013년 9월 ~ 현재 : 과학기술연합대학원대학교 부교수
- ORCID : https://orcid.org/0000-0002-3642-2323
- 주관심분야 : 전술통신, 위성통신, 인지무선통신, 바이오통신, 정보융합 등

정 태 충

- 1980년 : 서울대학교 전자공학과 학사
- 1982년 : KAIST 컴퓨터과학 석사
- 1987년 : KAIST 컴퓨터과학 박사
- 1988년 ~ 현재 : 경희대학교 컴퓨터공학과 교수
- ORCID : https://orcid.org/0000-0001-7387-5113
- 주관심분야 : 기계학습, 메타서치, 로보틱스

배 성 호

- 2011년 : 경희대학교 전자공학과 학사
- 2012년 : KAIST 전기및전자공학 석사
- 2016년 : KAIST 전기및전자공학 박사
- 2016년 ~ 2017년: MIT 컴퓨터과학 박사후과정
- 2017년 ~ 현재 : 경희대학교 컴퓨터공학과 조교수
- ORCID : https://orcid.org/0000-0003-2677-3186
- 주관심분야 : 심층신경망 모델 압축/해석/탐색, 비디오코딩, 이미지 처리