Turkish Journal of Electrical Engineering and Computer Sciences

Volume 27 | Number 5 Article 9

1-1-2019

A no-reference framework for evaluating video quality streamed through wireless network

MUHAMMAD UZAIR

ROBERT D. DONY

MOHSIN JAMIL

KHAWAJA BILAL AHMAD MAHMOOD

MUHAMMAD NASIR KHAN

Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2019) 27: 3383 – 3399

doi:10.3906/elk-1811-173

A no-reference framework for evaluating video quality streamed through wireless network

Muhammad UZAIR[1,*], Robert D. DONY[2], Mohsin JAMIL[3,4] Bilal A. KHAWAJA[3,5], , Muhammad Nasir KHAN[6]

[1]Faculty of Engineering, Islamic University of Madinah, Madinah, KSA
[2]University of Guelph, Guelph, Ontario, Canada
[3]Faculty of Engineering, Islamic University of Madinah, Madinah, KSA,
[4]Department of Robotics and AI, National University of Sciences and Technology, Islamabad, Pakistan
[5]Department of Electronic and Power Engineering, National University of Sciences and Technology, Karachi, Pakistan
[6]Department of Electrical Engineering, The University of Lahore, Lahore, Pakistan

Received: 27.11.2018 • **Accepted/Published Online:** 26.03.2019 • **Final Version:** 18.09.2019

Abstract: In this work, a no-reference framework is proposed for the video quality estimation streamed through the wireless network. The work presents a comprehensive survey of the existing full reference (FR), reduced reference (RR), and no-reference (NR) algorithms. A comparison has been made among existing algorithms, i.e. in terms of subjective correlation and feasibility to use these algorithms in wireless architecture, to describe the necessity of the proposed framework to overcome the limitations of the existing algorithms. A brief summary of our previously published algorithms, i.e. NR blockiness, NR blur, NR network, NR just noticeable distortion, and RR, has also been presented. These algorithms have also been used as function modules in the proposed framework. The proposed framework is able to measure the video quality by taking into account major spatial, temporal, network impairments, and human visual system effects for a comprehensive quality evaluation. The proposed framework is able to measure the video quality compressed by different codecs, i.e. MPEG x / H.264x, Motion JPEG/Motion, and JPEG2000, etc. The framework is able to work with two different kinds of received data, i.e. bit streams and decoded pixels. The framework is an integration of the RR and NR method, and can work in three different modes depending on the availability of the RR data, i.e. 1) only RR measurement, 2) hybrid of RR and NR measurement, and 3) only NR estimation. In addition, any individual function block, i.e. blurring, can also be used independently for particular specific distortion. A new subjective video quality database containing compressed and distorted videos (due to channel induced distortions) is also developed to test the proposed framework. The framework has also been tested on publicly available LIVE Video Quality Database. Overall test results show that our framework demonstrates a strong correlation with subjective evaluation of the two separate video databases as compared with other existing algorithms. The proposed framework also shows good results while working only in NR mode as compared with existing RR and FR algorithms. The proposed framework is more scalable and feasible to use in any kind of available network bandwidth as compared with other algorithms, as it can be used in different modes by using different function modules.

Key words: H.264, framework, network video, quality assessment

*Correspondence: uzair91@hotmail.com

1. Introduction

The development and growth of digital video systems (becoming an essential part of today's lifestyle) are replacing most of the analog video services, i.e. HDTV, high-definition video conferencing, etc. [1]. Video quality evaluation is very important for the development of the digital video systems. Digital video data suffers various kinds of distortions during acquisition, compression, transmission, decoding and reproduction [2]. Compression techniques used in the digital video system introduce many visual artifacts, i.e. blocking, blurring, etc., in videos which degrade its quality. Similarly, video data transmitted over wireless networks may be received incorrectly due to transmission impairments, i.e., packet losses, jitter, flickering, etc. Furthermore, human visual system (HVS) effects play an important role in determining the quality of the received video. Therefore, all of the video processing, i.e. compression, transmission and postprocessing, etc., produce distortions/artifacts in the reconstructed video. The nature of the artifacts also depends on the compression methods, e.g., MPEG x, H.26x, and on the position of the error and decoder implementation [3]. It is therefore very important for a video communication system to measure the received video quality for further processing, i.e. maintenance, enhancement. Therefore, an image and video quality metric is required, which can assess the video quality by measuring all kinds of distortions in a real time scenario [2]. Video quality can be measured by using either subjective or objective approach. Full reference (FR), reduced reference (RR), and no reference (NR) are three kinds of objective methods of evaluating video quality.

1.1. Challenges and limitations

Many quality metrics are presented by the research community, but these metrics have many limitations as described below:

- Existing metrics are generally distortion-specific, and are not able to measure different kinds of distortion.
- Existing FR metrics are not applicable in wireless transmission due to the bandwidth constraints.
- Existing HVS-based metrics are very complex.
- Existing RR and NR metrics show good results, but fail to capture network and temporal losses efficiently.
- Existing compression distortions measuring metrics do not measure network, temporal, and HVS effects.

1.2. Research motivation

There is a need for a quality metric/framework that should be able to cover the above mentioned limitations. The goal of this research is to develop a quality framework which should have the following characteristics:

- It should be inclusive rather than specific to a particular kind of distortion.
- In order to work in real time, it should not require original frames for comparison at any stage.
- The metric should be able to measure major compression, network, temporal, and HVS distortions.
- The metric should be able to predict accurately and efficiently as compared to existing metrics.
- It should correlate strongly with subjective scores as compared with existing metrics, and it should also be feasible according to the demand and application.

In this work, we first present a detailed survey and comparison of the existing FR, RR, and NR algorithms in terms of their performance and feasibility in the wireless environment. Based on the survey, a hybrid (RR+NR) framework is proposed to predict the video quality for different kinds of distortions. The proposed framework can work in three different modes. The proposed framework also uses two different kinds of data to predict quality. For transmission distortion, the framework works only with the received bit stream. For compression-induced artifacts, i.e. blocking, blurring, temporal, RR, and just noticeable distortion (JND), the framework works on the decoded video. A video database is also developed in this work to measure the performance of the proposed framework. The framework was tested with developed database, and was also further verified with the H.264 compressed database provided by the LIVE Video Quality Database. The results show that the proposed framework demonstrates a strong correlation with subjective evaluation of the two separate video databases.

The remainder of this paper is organized as follows: Section 2 presents a detailed review of the previously presented objective quality metrics and their performance and feasibility over wireless networks. Section 3 presents the proposed system framework for video quality evaluation. Section 4 presents a brief summary of our previously published works which are used in the proposed framework. Section 4 also presents a new NR algorithm for the temporal distortion measurement. Section 5 describes the details of the new developed video database for this work. An overall final quality metric is also defined in this section. Section 6 presents the conclusion and future work.

2. Related work

This section presents a literature survey of the existing objective quality metrics, i.e. FR, RR and NR.

2.1. HVS-based metrics

Many quality metrics have been presented in the literature by using the anatomy and psycho-physical features of the HVS. HVS-based models can be divided into two groups: single-channel models, and multiple-channel models. Single-channel models process all inputs in the same way, i.e. by taking the HVS as a single spatial filter. Schade designed the first vision-based model on the assumption that the cortical representation is a shift invariant transformation of the retinal image [4]. Mannos and Sakrison designed an HVS-based model for luminance images [5]. Multichannel models divide the image into multiple channels, where each channel is sensitive to different spatial frequency and orientation. In another approach, three-dimensional wavelet filters [6] were used to estimate the video quality. Although the above-mentioned approaches show good correlation with subjective data, but all of these approaches are rather complex and computationally intensive. Therefore, an algorithm/model is really necessary to predict the quality for any kind of distortion. The model should also be simple and computationally efficient.

2.2. Full reference methods

FR metrics can be divided into simple pixel-based and similarity metrics.

2.2.1. Pixel-based metrics

Pixel-based metrics predict video quality by directly working with the pixels of the original and received images. The mean square error and peak signal-to-noise ratio (PSNR) are two pixel-based quality metrics. These metrics are simple, but they do not correlate well with the subjective evaluation. The structural similarity index

measurement (SSIM) is another FR method which measures the quality by evaluating structural distortion. The multistructural similarity index measurement (MSSIM) [7] metric has also been developed providing more flexibility as compared to the SSIM by incorporating the variety of image resolutions and view conditions.

2.2.2. Similarity metrics

Similarity metrics use different mathematical techniques to convert the intensity value of an image into another format by performing their calculation from pixel to block level. These new transformed values are compared between the original and degraded frames. One of the techniques [8] uses difference of the singular value decomposition (SVD), while another uses projected value [9], and the approach defined in [10] uses radon transform measurement to evaluate the quality. Although these FR algorithms show strong correlation with subjective data, they do not take HVS effects into account. Moreover, FR methods are not feasible for real-time applications due to the current bandwidth constraint wireless environment.

2.3. Reduced reference metrics

Webster et al. [11] presented the first RR quality metrics in which spatial, and temporal information features are extracted from the reference and transmitted data over a reduced bandwidth channel to evaluate the video quality. An RR hybrid image quality metric (HIQM) metric has also been developed in [12], which evaluates the quality by calculating the weighted sum of the different extracted features on the transmitter and receiver sides. Similarly, an RR objective quality metric [13] extracts features such as mean and standard deviation from the processed spatial-temporal (S-T) regions of the input, and output video streams. Although the proposed reduced reference methods show very good correlation with the subjective evaluation, but still these methods are not feasible for current bandwidth constraint wireless environment, and are feasible only when at least some information can be transmitted through the wireless medium. In order to overcome the limitations of the existing algorithms, we proposed an RR metric in one of our previous works [14]. Our proposed metric showed a strong correlation with subjective evaluation.

2.4. No-reference methods

In this method, the original image is not available for comparison. Most of these methods are designed for the assessment of distortion which comes from the discrete cosine transform (DCT)-based compression.

2.4.1. Blocking metrics

Many FR, RR, and NR algorithms presented in the literature define blocking distortion within the image/video. These methods use different approaches to evaluate the blocking distortion, i.e. error differences between the original and distorted images [15], DC coefficients of the DCT [16], by assigning weights to the HVS effects, and then measuring the difference of the intensity values between the columns and rows in each block [17], etc. Most of these metrics take into account additional features such as HVS effects, flatness measurement, and zero crossing to estimate blockiness. It makes them computationally less efficient, and/or they are FR-based. In order to overcome the existing limitations, we proposed an NR algorithm based on SVD in our previous work [18]. The proposed algorithm does not need any other measurements such as HVS, zero crossing, etc. and the results showed that our proposed algorithm outperformed other metrics.

2.4.2. Blurring metrics

Blurring is the next most important compression artifact after blockiness, especially with low compressed bit rates. FR, RR, and NR algorithms presented in the literature use many features to estimate blur, i.e. variance [19], and first- (gradient) and second-order (Laplacian) derivatives [20], etc. Most of these metrics do not take into account the important influence of specific image content, and HVS effects on the actual visibility of the artifacts. Therefore, these metrics are not able to predict the quality efficiently. An algorithm was also proposed in our previous work to overcome limitations in existing metrics [21]. In our proposed algorithm, the edges were detected by the threshold set by the HVS effects, and SVD was used to measure the spread of the edges in the spatial domain. Test results showed that our proposed algorithm has a strong correlation with subjective evaluation [21].

2.5. Network loss induced metrics

Many FR, RR, and NR algorithms were previously presented to estimate the network losses. In [22], the author uses the motion intensity and packet loss effect to measure the network losses. These informations are extracted from the video stream packet's content. The proposed algorithm does not show a strong correlation with subjective data. No-reference video quality monitoring metric was presented in [23] to estimate the video quality distorted by the packet loss for H.264/AVC compliant coded video. The algorithm measured quality efficiently at the macroblock, frame, and sequence levels. However, the metric can work with I and P frames only. Similarly, an NR algorithm was presented in [24] to detect the packet losses in transmitted video by working on the decoded pixel values. This algorithm is capable of processing up to 25 frames per second of Full HD video, and shows good correlation with subjective assessment. In [25], the author measures network losses by using user data field of the video. This approach is less intrusive, does not need to inject extra probing stream, and can also provide the packet loss detailed information of all frames. In order to overcome all these limitations, an algorithm was proposed in our previous work to estimate the transmission losses [26]. Our proposed NR algorithm is capable of measuring network losses for video encoded for all frames, i.e., I, P, and B frames. The proposed algorithm has low computational requirements, and can measure distortion up to 20% PLR. The algorithm works on the spatio-temporal dynamics of the video, and simulation results proved that metric correlates well with the subjective evaluation.

2.6. JND measurement metrics

JND models can be pixel-based and DCT-based models. For the pixel-based JND models, Chou and Li [27] presented an algorithm to measure JND by using luminance, and texture masking information. Research has shown that estimating the contrast masking while keeping in mind the difference of the texture, and edge region is very important. Therefore, Anmin et al. [28] decomposed an image into EM (edge masking), and TM (texture masking) structures to measure the JND in a pixel based approach. For the subband/DCT-based JND models, Ahumada and Peterson [29] developed a well-cited JND model in the DCT domain by measuring spatial contrast sensitivity function (CSF) for every DCT component to evaluate the JND threshold. Pixel- and DCT-based JND models do not measure texture masking, except that in [28]. Moreover, all JND models do not measure temporal effects of the HVS, except that in [30]. Moreover, the spatial/temporal CSF are not measured by the pixel-based JND models. For a complete JND model for videos, the temporal characteristics of HVS must be measured with spatial CSF. The model in [30] measures temporal effects in the pixel domain.

Kelly [31] developed a spatio-temporal CSF model based on retinally stabilized travelling wave stimuli, which was enhanced by Dally [32] by adding the eye movements. However, all of these existing models ignore the effects of foveal vision, which is very important as the size of the images and videos are increasing, i.e. high definition-images and videos. Keeping the limitations of the existing metrics in view, an algorithm was proposed in our previous work to estimate JND in pixel domain [33]. The proposed algorithm takes into account all major effects to measure JND, i.e. luminance adaptation, contrast masking, CSF, eye movement, and the foveal effect. The results showed that the proposed algorithm outperformed the existing algorithms.

2.7. Temporal metrics

Many FR, RR, and NR algorithms that measure temporal distortions are also presented in the literature. In [34], a Flicker Sensitive-MOtion-based Video Integrity Evaluation model is proposed to measure the temporal distortion. The model integrates the well-known MOVIE Index with a new perceptual flicker visibility index. The model uses the responses of neurons in primary visual cortex to measure flicker. The proposed model shows almost same results as existing algorithms. In [35], the author presents a full reference method to measure temporal distortion based on space time texture using motion tuning strategy. The test results show that the presented method correlates highly with the subjective quality and has a high computational efficiency. Similarly, a model is presented in [36] to enhance the SpatioTemporal model (VMAF(ST-VMAF)) proposed by Netflix. The model proves the improved performance on many subjective video databases. All of the existing metrics are good for measuring temporal distortion, but mostly use additional information to enhance their measuring capabilities. In the present study (Section 4), we also propose an NR algorithm to measure the temporal distortion. This algorithm is also used as a function module in the proposed framework.

2.8. Metrics based on neural networks approach

Similarly neural network approach is also used to measure quality in an NR way. In [37], the author uses deep convolutional neural networks (CNN) approach to measure the quality by integrating the feature learning and regression into one optimization process. Similarly, the author uses machine learning approach in [38] to combine a simple NR metrics approach to derive a predictive NR assessment metric. The algorithm obtained a correlation of over 97% correlation, but the algorithm has been tested up to the packet loss rate of 10% only. In another approach [39], an NR deep blind video quality assessment approach is used by considering various spatial and temporal cues obtained by using the deep CNN approach, and temporal cues features are obtained from spatial cues. However, the algorithm does not show very good results as compared with existing algorithm. In [40], another deep CNN-based approach is presented by using ten convolutional layers, five pooling layers for feature extraction, and two fully connected layers for regression. The algorithm does not require any other additional information and shows good correlation with subjective data. In [41], an NR framework is proposed based on the 3D shearlet transform and CNN. Spatiotemporal features are extracted by using 3D shearlet transform, and then CNN and logistic regression are used to predict a perceptual quality score.

The existing NR algorithms, i.e. blocking, blurring, etc., predict the distortion efficiently, and can also be used in a bandwidth constraint wireless environment. However, many of these metrics are distortion-specific, or only able to work with images, and are not able to predict video quality efficiently. Therefore, the existing NR metrics also create a necessity to develop other metrics/framework, which should be suitable in a bandwidth constraint wireless environment, and should be able to predict the quality comprehensively, and efficiently, i.e. taking into account all kinds of distortions.

2.9. Comparison between metrics

We have presented a brief overview of different quality metrics. There is no common ground among these metrics, and it is difficult to make a proper performance comparison among them. However, we present a comparison in Table 1 among these metrics based on their performance (correlation) with subjective scores, and their suitability for wireless network architecture.

Table 1. Comparison of different metrics.

Approach	Performance	Feasibility in wireless architecture
Subjective evaluation	Very Good	Low (Complex and Expensive, Time consuming)
FR (Pixel-based, e.g., PSNR)	Low	Low (Restricted due to FR condition)
FR (Struct. inform.-based, e.g., SSIM)	Good	low (restricted due to FR condition)
HVS-based	Average	Low (Complex, comput. high, no network loss estimation)
RR (Spatial/temporal artifacts-based)	Good	medium/high (less feasible as compared to NR)
NR (Spatial, temp., and network-based)	Good	medium/high (distortion specific)
Data hiding (watermarking)	Average	low/medium (additional overhead, network lost prob.)
Network Impairments (QoS measurement)	Average	medium (no compression artifacts meas.)

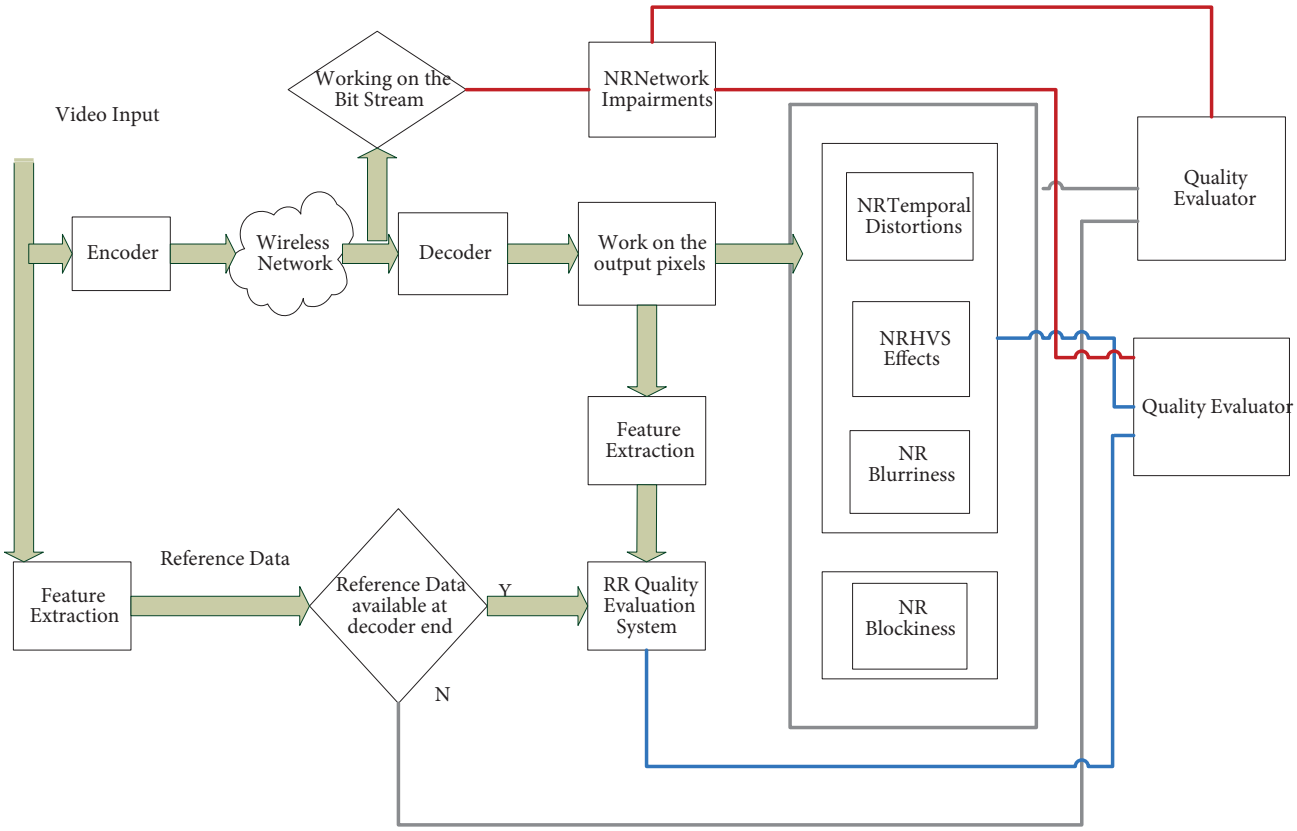
3. Proposed framework

The literature survey has shown that a model/framework is required, which should be able to measure all kinds of distortions, and should provide a complete end-to-end quality evaluation. The true quality cannot be predicted by just measuring one type of distortion. Therefore, a framework is proposed by an integration of RR and NR methods by taking into account major spatial, temporal, and network impairments along with HVS effects. This framework is able to measure video quality compressed by different codecs such as MPEG x, H.264x, and Motion JPEG/Motion JPEG2000. Figure 1a shows the framework and its function modules. The framework is able to work with two different types of received data. The transmission distortion has been estimated from the data received through the bit stream only, while all other distortions have been estimated by working on the decoded video at the output of the decoder. The framework can work in three modes depending on the availability of the RR data as described below.

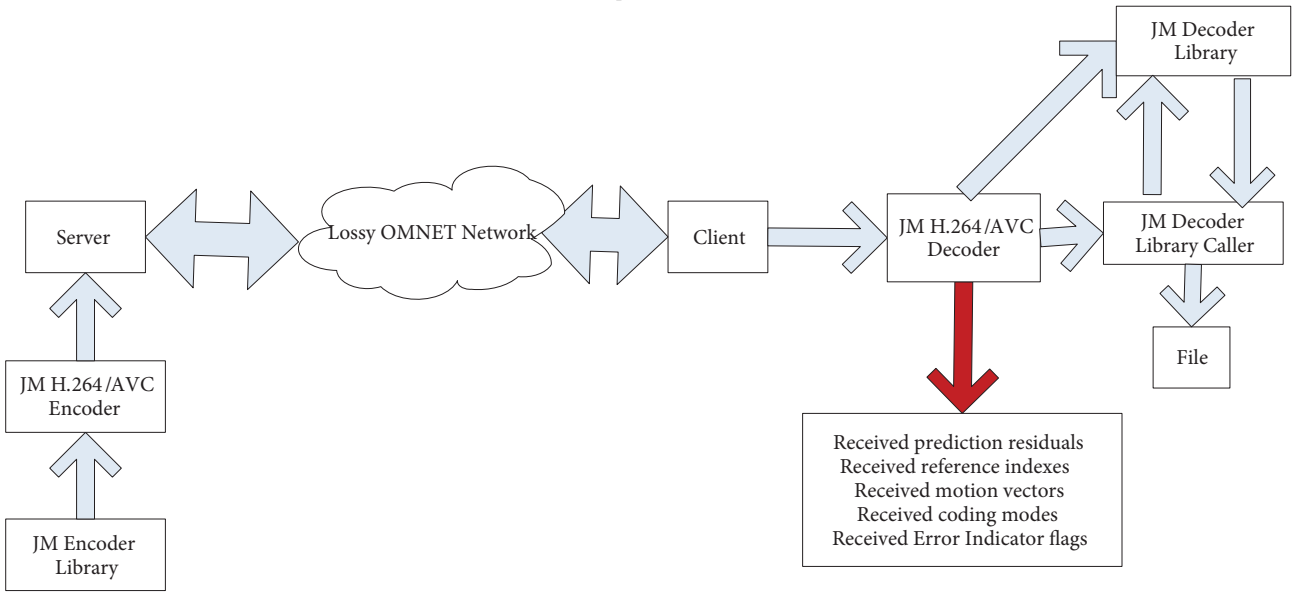
1. Working with RR measurements only.
2. Integration of the RR and NR (network, temporal, blurriness, and HVS) functional blocks. In this case, the model does not need to measure spatial distortions (blockiness).
3. Only NR estimation by using network, temporal, blurriness, blockiness, and HVS functional blocks. Any individual functional block, e.g., blurring, can also be used independently for specific distortion.

3.1. Experimental setup

In this work, OMNeT++ 4.1 network simulator [42] is used to integrate the JM16.2 H.264/AVC Reference CODEC [43] for AVC simulations. Actual H.264/AVC traffic, i.e. video packets, are encapsulated in the OMNeT++ packets instead of simulated OMNeT++ packets for creating a real-time simulated scenario. Figure 1b shows the experimental setup as explained in [26].



(a) Proposed framework



(b) Experimental set-up

Figure 1. Framework and experimental set-up

4. Distortion measurements

In order to measure the quality, we measured network distortion, JND, blocking distortion, blurring distortion, and we also measured the quality by using reduced reference approach. All of these distortion algorithms (function modules) are our previously published work. The network distortion was measured using an NR approach by working on the received bit stream only [26]. The proposed algorithm was capable of working on all I, P, and B frames. The JND was also measured using an NR approach [33]. A pixel-based JND model was proposed which takes into account all the major effects, including foveal effects for a complete spatio-temporal CSF estimation in pixel domain. The blocking, and blurring distortions were measured using an NR approach [18, 21]. Similarly, the quality was evaluated by using the RR method [14]. The reader can review these papers for further details. The next section describes a newly proposed NR temporal distortion algorithm for this work. This algorithm is also used as a function module in the proposed framework.

4.1. NR temporal distortion measurement

Moving artifacts are related to the motion compensation effects, and are responsible for causing image persistence. Therefore, an observer is not able to observe a visual artifact because of the previous frame persistence over current frame. Existing temporal distortion measuring metrics generally use motion vectors, and motion compensated information of the P and B frames. But these frames can also contain distortion due to the temporal error propagation. In order to overcome this limitation, an NR temporal estimation algorithm is proposed by using only I frames as a reference, as I frames are not affected by the temporal propagation, i.e. movement artifacts. The algorithm uses incoming I frames in each group of pictures (GOP) as reference, and following P, and B frames are treated as distorted. The MSSIM is a full reference algorithm [44]. This algorithm is applied in an NR approach in this work, i.e. an MSSIM value is obtained for each of the P and B frames by taking the I frame as reference, and P and B frames as distorted [44]. Similarly, an MSSIM value is also measured for each of the I frame, i.e. a previous I frame is treated as reference, and the following I frame is treated as distorted in two consecutive GOPs. An average is taken for the whole video sequence which provides us an estimate of the temporal distortion. The next section describes the subjective database and simulation results.

5. Subjective data analysis and final quality evaluation

This section describes the subjective study, which was conducted to create a test database to evaluate the framework. This procedure is also known as subjective assessment, in which viewers provide us a score defining the quality of the received video. A human study of the distorted videos is conducted to create a database containing compressed as well as distorted videos due to the channel-induced distortions, i.e. creating a wide range of quality videos in order to have a good perceptual variation to test the efficiency of the framework.

5.1. Source and test sequences

The source videos which we used to formulate the database is in raw, uncompressed, and progressive scan YUV420 format. Twelve video sequences are used, and these video sequences are chosen based on the variety of the content, spatial, and motion information. A set of 120 distorted sequences was created by using different bit-rates, and packet loss rates for each reference sequence. Twelve distorted videos were produced for each of the "City", "Crew", "Foreman", "Akiyo", "Hall", and "Coastguard" video sequence. Eight distorted video sequences were produced for each of the "Race", "Fries", "Rugby", "Car phone", "Mobile", and "Mother" video sequence. The packet losses are modeled using two state Gilbert model. The main profile is used for encoding,

and the videos are encoded with I, P, and B frames with two B frames in a GOP. Table 2 shows the parameter values.

Table 2. Video sequences encoded with I, P, and B frames

Sequence	Resolution	Data rate	fps	Total frames	PLR
Crew	SD (704 × 576)	0.01, 0.1, 0.5, 1, 1.5, 2, and 3.5 Mb/s	30	300	1, 2, 5, 10 and 20
City	SD (704 × 576)	0.01, 0.1, 0.5, 1, 1.5, 2, and 3.5 Mb/s	30	300	1, 2, 5, 10 and 20
Rugby	SD (720 × 576)	0.01, 0.1, 0.5, 1, 1.5, 2, and 3.5 Mb/s	30	220	1, 2, 5, 10 and 20
Fries	SD (720 × 576)	0.01, 0.1, 0.5, 1, 1.5, 2, and 3.5 Mb/s	30	220	1, 2, 5, 10 and 20
Race	SD (720 × 576)	0.01, 0.1, 0.5, 1, 1.5, 2, and 3.5 Mb/s	30	220	1, 2, 5, 10 and 20
Foreman	CIF (352 × 288)	0.1, 1, 5, 50, 150, 264, and 360 kb/s	30	300	1, 2, 5, 10 and 20
CoastGuard	CIF (352 × 288)	0.1, 1, 5, 50, 150, 264, and 360 kb/s	30	300	1, 2, 5, 10 and 20
Akiyo	CIF (352 × 288)	0.1, 1, 5, 50, 150, 264, and 360 kb/s	30	300	1, 2, 5, 10 and 20
Hall Way	CIF (352 × 288)	0.1, 1, 5, 50, 150, 264, and 360 kb/s	30	300	1, 2, 5, 10 and 20
Mother	CIF (176 × 144)	0.1, 0.2, 5, 10, 15, 36, 64, and 150 kb/s	25	300	1, 2, 5, 10 and 20
Mobile	CIF (176 × 144)	0.1, 0.2, 5, 10, 15, 36, 64, and 150 kb/s	25	300	1, 2, 5, 10 and 20
Carphone	CIF (176 × 144)	0.1, 0.2, 5, 10, 15, 36, 64, and 150 Kb/s	25	300	1, 2, 5, 10 and 20

5.2. Test methodology and Processing of the score

The subjective study for the database is conducted with single stimulus continuous quality evaluation (SSCQE) approach[45, 46]. In this study, only one video is shown to the viewer. The single stimulus approach needs less time by the reviewer, and also reduces the memory effects on the perceived quality. In this approach, the original reference videos are also shown to the observer, although observer is not told about its presence on the set of videos. This is used to equalize the scores. The score given to the reference videos is also a representative of the supposed bias which observer carries, and a compensation is acquired by subtracting the scores of the distorted videos with this bias, providing a difference score for that particular distorted video sequence. This measure is known as the differential mean opinion score (DMOS) [45, 46]. All of the created distorted videos are first loaded into the memory to avoid latencies. The videos are shown to the subject at a distance of four times the video height. Thirty observers participated in this study, and the subjects were also briefed about the study. The study was done in two sessions to minimize the subject fatigue. In each session, 66 videos (60 distorted + 6 references) were shown to the subjects randomly. The order was changed for each session, and also for each subject. Moreover, two consecutive sequences of the same reference were not shown to the viewer in order to minimize memory effects. Training videos were also shown to the viewers, and the subjects also scored training videos for training purpose. At the end of the presentation of the video, each viewer provided an opinion score with quality grading as: Bad (0–20), Poor (20–40), Fair (40–60), Good (60–80), and Excellent (80–100). The score that each subject assigned to a distorted sequence in a session was subtracted from the score that the subject assigned to the reference sequence in the same session, thus providing a difference score. The scores from the remaining subjects were then averaged to form the DMOS for each sequence.

5.3. Overall quality evaluator

A global quality evaluator combines the quantitative inputs from each of the functional blocks to predict the final quality score. This score represents the quality of the experience as perceived by the user of the given

application. Multivariate regression is a typical way to combine the individual metric scores into an overall quality score. It is a simple, and computationally effective method. The combination model is defined by the means of a multivariate linear regression as:

$$TotalDistortion = \alpha_1\, m_1 + \alpha_2\, m_2 + \alpha_3\, m_3 + \alpha_4\, m_4 + \alpha_5\, m_5 + \alpha_6\, m_6$$

The $m_1, m_2, ..., and\ m_6$ are the variables for the blocking, blur, temporal, JND, network, and reduced reference functional blocks. The coefficients $\alpha_1, \alpha_2, ..., and\ \alpha_6$ are the corresponding coefficients of these variables. These coefficients are computed by using the multivariate linear regression, and represent the best match with the test database to obtain the optimum solution.

5.4. Evaluation of the proposed framework on tested database

All of the individual functional blocks (network, blur, blocking, temporal, JND, and RR metrics) are processed individually on each of the tested videos. They are combined using the multivariate regression to estimate the total distortion in the received video sequence. Different values of the correlations are obtained by using, i.e. combining, different function modules in the proposed framework. Table 3 shows the correlation results of the proposed framework for different combinations and their comparison with other metrics. The Pearson correlation coefficient (PCC), root mean square error (RMSE), and the Spearman rank order correlation coefficient (SROCC) measurements are used to test the performance of the proposed framework.

Table 3. Correlation coefficient comparison with other metrics tested on developed video database.

Metrics/algorithms	PCC	SROCC	RMSE
NR+RR measurement	0.7840	0.7183	6.72
NR measurement	0.7453	0.6929	7.37
Without blockiness measurement	0.7941	0.7153	6.46
Without RR and transmission measurement	0.6644	0.6356	7.5612
STRRED	0.8253	0.8109	5.89
MSSIM	0.8076	0.8006	6.33
RRIQA	0.6332	0.6103	7.63
PSNR	0.3653	0.4389	9.03

These results show that the framework shows a good correlation with the subjective data (even operated in the NR mode only), and the effectiveness increases as we use it with the RR mode, i.e. NR+RR. The framework also shows good correlation without blockiness measurement. Therefore, the framework can ignore the blockiness functional block if RR data are available, as RR measurement estimates the blockiness distortion itself. The framework is also able enough to estimate the video quality when it is used with compression distortion measuring functional blocks only. Similarly, the correlation coefficients have also been measured for the STRRED, MSSIM, RRIQA, and PSNR algorithms. The STRRED algorithm [47] works with less original information to full original information from the reference video, i.e., effectively making it a full reference VQA algorithms. The PSNR and MSSIM [44] are also FR algorithms, while reduced reference image quality assessment (RRIQA) [48] is an RR algorithm.

Compared with other algorithms, our proposed framework (used in an NR mode only) shows better correlation than PSNR and RRIQA. The framework used in compression mode also shows marginally better

correlation than RRIQA. With the addition of the RR mode with NR, the framework shows almost the same correlation as MSSIM. The STRRED algorithm shows good correlation for SROCC measurement. However, PCC and RMSE results are not as good as SROCC when it is compared with our proposed framework used in RR+NR mode. In [47], STRRED algorithm was tested with LIVE H264 video quality database [45], and the packet losses in LIVE H264 video quality database are up to 10% only. However, packet losses in our developed database are up to 20%, which are considered as severe losses. The STRRED showed good results in [47] when compared with MSSIM, i.e. when it was tested with a packet loss ratio of maximum 10%. However, STRRED algorithm does not show same good results as in [47] when compared with MSSIM, and our proposed framework (NR+RR mode) for a packet loss ratio of up to 20%. Moreover, our framework measures transmission distortion by working on the received bit stream only instead of decoded pixels. Our proposed framework still shows good correlation with subjective data with this severe packet loss. Overall, the video quality estimation provided by the framework is very satisfying, i.e. even with NR mode.

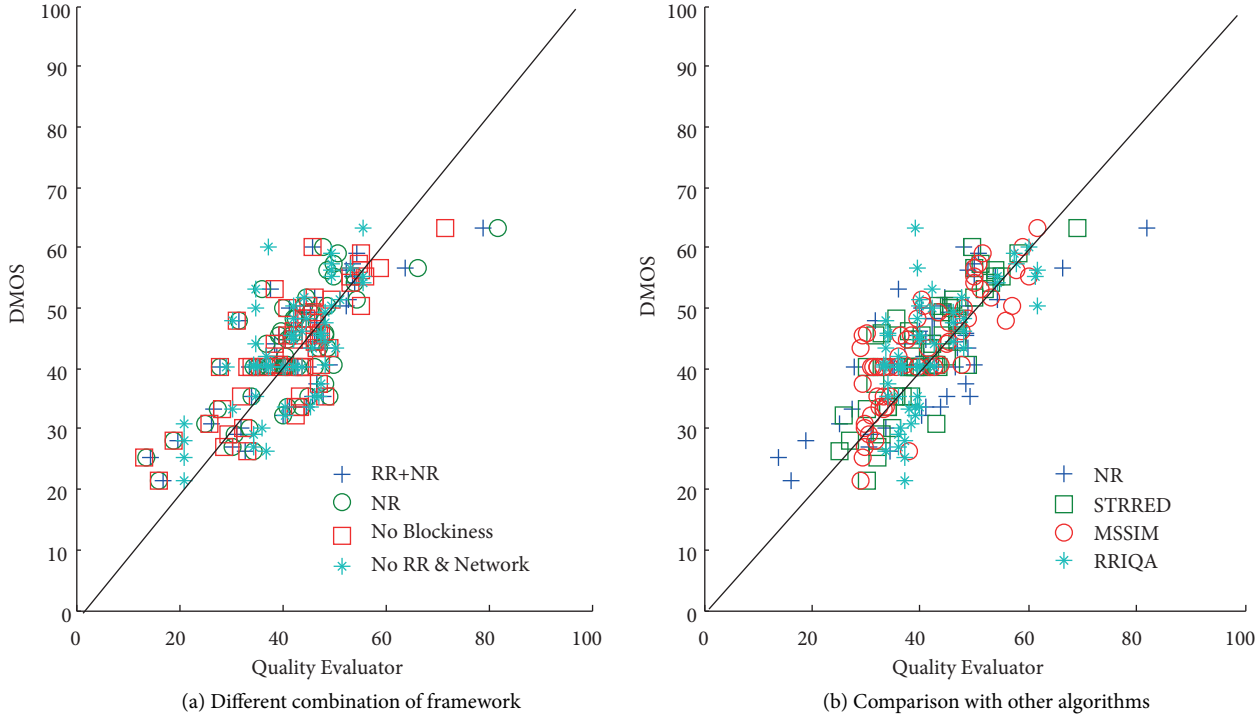


Figure 2. Scatter plots (using developed database).

Figure 2a shows the scatter plot comparisons, i.e. visual comparison of the framework when it is tested by combining different function modules. The figure shows that the number of outliers is not large when used in different modes. Figures 2b and 3a show the scatter plot comparison with other algorithms. Both figures show that the number of outliers is less as compared with RRIQA. However, the STRRED algorithm shows the best linear relation to the subjective evaluation.

5.5. Evaluation of the the proposed framework with LIVE Video Quality database

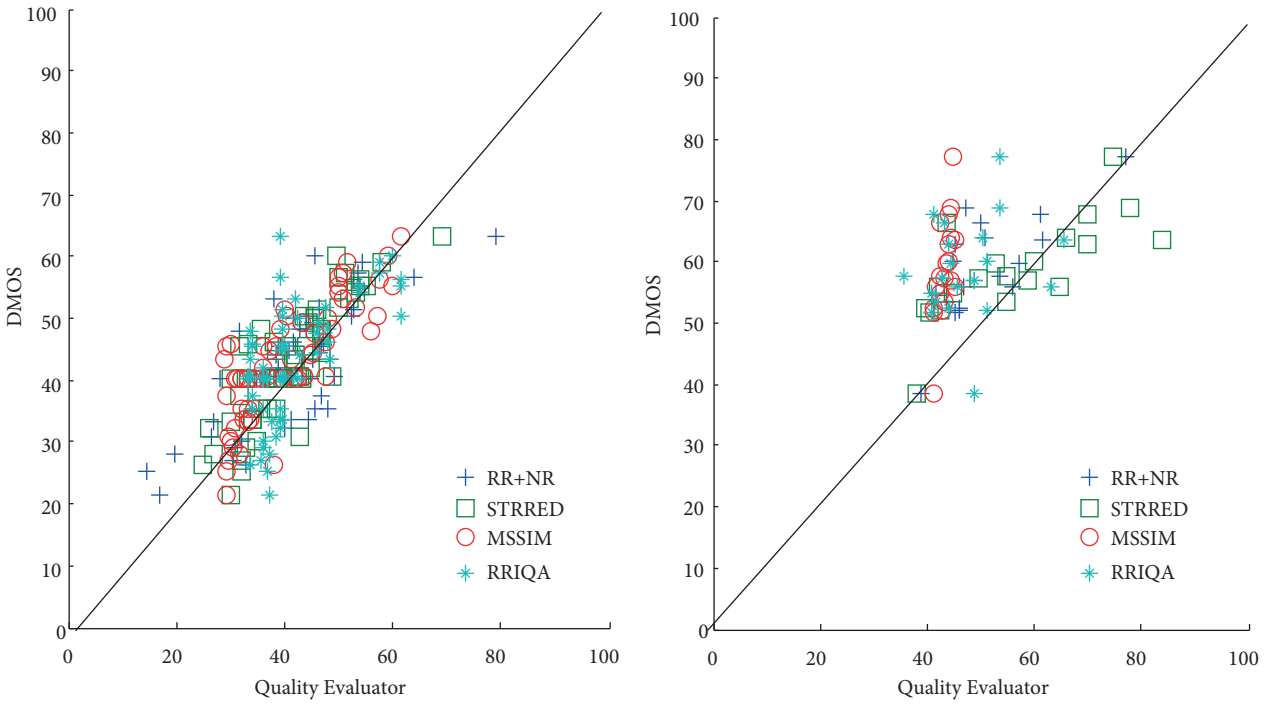
The framework was also tested on the LIVE Video Quality (H.264 compressed videos.) Database [46]. Table 4 shows the correlation results for different combinations. These results again show the same trend as we have seen

in the previous section. The results again validate our assertion that the framework produces good results in the NR mode and its effectiveness increases as we combine it with RR mode. Compared with other algorithms, our framework almost shows the same performance when compared with STRRED, i.e. NR+RR modes. The results also show that our framework used in NR+RR mode has a better correlation than MSSIM.

Table 4. Correlation coefficient comparison with other metrics tested on LIVE H.264 compressed database.

Metrics/algorithms	PCC	SROCC	RMSE
NR (without transmission module)+RR measurement	0.7221	0.6391	10.66
NR measurement (without transmission module)	0.536	0.4487	16.92
NR+RR (without blockiness and transmission distortion measurement)	0.7389	0.6722	13.47
RR measurement	0.5529	0.3814	14.01
STRRED	0.7261	0.7934	9.91
MSSIM	0.6539	0.6453	17.12
RRIQA	0.2085	0.2663	14.88
PSNR	0.4858	0.4314	16.39

Figure 3b shows the scatter plot comparison with other algorithms (NR+RR mode without blockiness and transmission distortion measurement). The figure shows that the number of outliers is less as compared with MSSIM and RRIQA. However, STRRED shows almost the same number of outliers as our proposed framework.



(a) Comparison with other metrics with developed database

(b) Comparison with other metrics with LIVE database

Figure 3. Scatter plots

A framework has been tested on two different databases, and shows the same trend. Overall test results show that the framework achieves good performance, scalability, and feasibility to assess the video quality. The framework enables us to work on two different kinds of data and work in three different modes. The ability of the framework to work on either RR or NR mode make it very feasible in any kind of available bandwidth environment. Moreover, the scalability is greatly enhanced to perform by using different function modules. The framework also shows better performance, and more flexibility as compared with other existing algorithms.

6. Conclusions and future work

This work deals with the issue of video quality evaluation, which is increasingly becoming an important issue in audiovisual communications. To improve the current state of the art, factors of NR estimation (most suitable for bandwidth limited wireless environments), and the way human observers evaluate the content of video were considered. We first presented a detailed survey, and comparison of the existing FR, RR, and NR algorithms in terms of their feasibility in a wireless environment. Based on the survey, a framework is proposed in order to overcome the limitations of the existing algorithms, i.e. to predict the video quality for different content type, data rate, and error rate combinations. The proposed framework is a hybrid of the integral of RR, and NR modes. The framework can work in three different modes depending on the availability of the RR data. The RR and NR modes can also evaluate the quality individually. Moreover, each individual function block in the NR mode can also work independently, showing the flexibility/capability of the framework. The framework is able to work with two different kinds of received data. For transmission distortion, the framework works only with the received bit stream, while the framework works on the decoded video to measure the blocking, blurring, temporal, RR, and JND distortions. All of these metrics (function modules) had already been individually tested on the publicly available database in our previous works. A new video database containing compressed and distorted videos due to channel-induced distortions was also developed to test the performance of the proposed framework. The proposed framework was also further verified with the H.264 compressed database provided by the LIVE Video Quality Database. The results show that our framework demonstrates a strong correlation with the subjective evaluation for the two separate video databases as compared with existing algorithms. The framework also shows good performance when it is used in NR mode only. The framework is highly flexible due to its ability to add new function modules to enhance its performance and the capacity of working in any bandwidth environment. Depending on the application, hardware capacity, and the availability of the bandwidth, a user can pick a function module/mode for the prediction of video quality. Furthermore, a user can pick a specific function module to find out degradation in quality due to a specific distortion, i.e. blocking, etc.

The development of new function modules which will estimate distortion using quantization step size, and based on the number of encoded DCT coefficients is part of our future work. Moreover, the process of fine tuning the proposed framework with subjective scores measured over a broad range of video contents, and processing schemes is a continuing topic of interest. Due to its great flexibility, the framework can also be used to estimate video quality for video encoded in any advanced video codec, i.e. H.265/HEVC.

References

[1] Lajos H, Peter J, Jurgen S. Wireless Video Communication: Second to Third Generation Systems and Beyond. New York, NY, USA: Wiley, 2001.

[2] Theodore S. Wireless Communication: Principle and Practice. NJ, USA: Prentice Hall, 2002.

[3] Haohong W, Lisimachos P. 4G Wireless Video Communications. London, UK: Wiley, 2009.

[4] Otto R, Schd S. Optical and photoelectric analog of eye. Journal of Optical Society of America 1956; 46 (9): 721-738. doi: org/10.1364/JOSA.46.000721

[5] Mannos J, Sakrison D. The effects of a visual fidelity criterion of the encoding of images. IEEE Transactions on Information Theory 1974; 20 (4): 525-536. doi: 10.1109/TIT.1974.1055250

[6] Hekstra P, Beerends J, Ledermann D, Caluwe F, Kohler S. PVQM – a perceptual video quality measure. ELSEVIER Signal Processing: Image Communication 2002; 17 (10): 781-798. doi: org/10.1016/S0923-5965(02)00056-5

[7] Wang Z, Simoncelli P, Bovik A. Multiscale structural similarity for image quality assessment. In: IEEE 2003 Conference on Signals, Systems and Computers; Asilomar, California, USA; 2003. pp. 1398-1402.

[8] Shnayderman A, Gusev A, Eskicioglu M. An SVD-based grayscale image quality measure for local and global assessment. IEEE Transactions on Image Processing 2006; 15 (2): 422-429. doi: 10.1109/TIP.2005.860605

[9] Jianxin P, Rong Z, Lu L, Jinhui T, Zhengkai L. A projection-based image quality measure. International Journal of Imaging Systems and Technology 2008; 18 (2): 94-100. doi: org/10.1002/ima.20156

[10] Jianxin P, Rong Z, Zhenkai L. Image quality assessment metrics with radon transform. In: IEEE 2008 International Conference on Systems, Man and Cybernetics; Singapore, Singapore; 2008. pp. 1-6.

[11] Arthur A, Coleen T, Stephen W. Objective video quality assessment system based on human perception. Proceedings of SPIE 1993; 1913: 15-26. doi: 10.1117/12.152700

[12] Ulrich E, Tubagus K, Jurgen Z. Perceptual quality assessment of wireless video applications. In: IEEE 2006 4th International Symposium on Turbo Codes and Related Topics; Munich, Germany; 2006. pp. 1-6.

[13] Stephen W, Margaret H. Spatial-temporal distortion metric for in-service quality monitoring of any digital video system. Proceedings of SPIE 1999; 3845: 266-277. doi: org/10.1117/12.371210

[14] Uzair M, Fayek D. Reduced reference image quality assessment using principal component analysis. In: IEEE 2011 International Symposium on Broadband Multimedia Systems and Broadcasting; Munich, Germany; 2011. pp. 1-6.

[15] Miyahara M, Kotani K, Algazi R. Objective picture quality scale (PQS) for image coding. IEEE Transactions on Communications 1998; 46 (9): 1215-1226. doi: 10.1109/26.718563

[16] Chan W, Goldsmith P. A psychovisually-based image quality evaluator for JPEG images. In: IEEE 2000 International Conference on Systems, Man, and Cybernetics; Nashville, TN, USA; 2000. pp. 1541-1546.

[17] Wu H. A generalized block-edge impairment metric for video coding. IEEE Signal Processing 1997; 4 (11): 317-320.

[18] Uzair M, Fayek D. An efficient no-reference blockiness metric for intra-coded video frames. In: IEEE 2011 14th International Symposium on Wireless Personal Multimedia Communications; Brest, France; 2011. pp. 1-6.

[19] Erasmus S, Smithi K. An automatic focusing and astigmatism correction system for the SEM and CTEM. Journal of Microscopy 1982; 127 (2): 185-199. doi: 10.1111/j.1365-2818.1982.tb00412.x

[20] Wang Z, Sheikh H, Bovik A. Handbook of Image and Video Processing. New York, NY, USA: Academic Press, 2000.

[21] Uzair M, Dony R. An efficient no-reference blurriness metric for images and video frames. In: IEEE 2016 Canadian Conference on Electrical and Computer Engineering; Vancouver, Canada; 2016. pp. 1-4.

[22] Amir M. A video streaming quality assessment scheme based on packet level measurement. In: IEEE 2015 International Conference on Communications and Signal Processing; Sharjah, UAE; 2015. pp. 1556-1562.

[23] Matteo N, Marco T, Stefano T. No-reference video quality monitoring for H.264/AVC coded video. IEEE Transactions on Multimedia 2015; 11 (5): 932-946. doi: 10.1109/TMM.2009.2021785

[24] Mario V, Denis V. Video transmission artifacts detection using no-reference approach. In: IEEE 2018 Zooming Innovation in Consumer Technologies Conference; Novi Sad, Serbia; 2018. pp. 1932-1937.

[25] Zhiguo H, Qiqiang Z. A new approach for packet loss measurement of video streaming and its application. Springer Multimedia Tools and Applications 2018; 77 (10): 1158–1168. doi: 10.1007/s11042-016-3566-0

[26] Uzair M, Dony R. No-reference transmission distortion modelling for H.264/AVC coded video. IEEE Transactions on Signal and Information Processing over Networks 2015; 1 (3): 209-221. doi: 10.1109/TSIPN.2015.2476695

[27] Chou C, Chen C. A perceptually optimized 3-D subband codec for video communication over wireless channels. IEEE Transactions on Circuits and Systems for Video Technology 1996; 6 (2): 143–156. doi: 10.1109/76.488822

[28] Anmin L, Fan Z. Just noticeable difference for images with decomposition model for separating edge and textured regions. IEEE Transactions on Circuits and Systems for Video Technology 2010; 20 (11): 1648-1652. doi: 10.1109/TCSVT.2010.2087432

[29] Ahumada A, Peterson H. Luminance-model-based DCT quantization for color image compression. Proceedings of SPIE 1992; 1666: 365-374. doi: org/10.1117/12.135982

[30] Chou C, Chen W. A perceptually optimized 3-D subband codec for video communication over wireless channels. IEEE Transactions on Circuits and Systems for Video Technology 1996; 6 (2): 143-156. doi: 10.1109/76.488822

[31] Kelly D. Motion and vision. II. stabilized spatio-temporal threshold surface. Journal of the Optical Society of America 1979; 69 (10): 1340-1349. doi: 10.1364/JOSA.69.001340

[32] Daly S. Engineering observations from spatiovelocity and spatiotemporal visual models. Proceedings of SPIE 1998; 3299: 180-191. doi: doi.org/10.1117/12.320110

[33] Uzair M, Dony R. Estimating just-noticeable distortion for images/videos in pixel domain. IET Journal in Image Processing 2017; 11 (8): 559-567. doi: 10.1049/iet-ipr.2016.1120

[34] Mannos J, Sakrison D. Video quality assessment accounting for temporal visual masking of local flicker. ELSEVIER Signal processing: Image communication 2018; 67: 182-198. doi: org/10.1016/j.image.2018.06.009

[35] Peng P. An efficient temporal distortion measure of videos based on space time texture. ELSEVIER Pattern Recognition 2017; 70: 1-11. doi: 10.1016/j.patcog.2017.04.031

[36] Bampis G. Enhancing temporal quality measurements in a globally deployed streaming video quality predictor. In: IEEE 2018 25th IEEE International Conference on Image Processing; Athens, Greece; 2018. pp. 614-618.

[37] Yuming L, Lai P, Fang Y. No-reference image quality assessment with deep convolutional neural networks. In: IEEE 2016 International Conference on Digital Signal Processing; Beijing, China; 2016. pp. 685-689.

[38] Maria T, Antonio L. Predictive no-reference assessment of video quality. ELSEVIER Signal Processing: Image Communication 2017; 52: 20-32. doi: 10.1016/j.image.2016.12.001

[39] Sewong A. No-reference video quality assessment based on convolutional neural network and human temporal behavior. In: Asia-Pacific Signal and Information Processing; Hawaii, USA; 2018. pp. 318-325.

[40] Sebastian B, Wojciech S. Deep neural networks for no-reference and full-reference image quality assessment. IEEE Transactions on Image Processing 2018; 27 (1): 206-219. doi: 10.1109/TIP.2017.2760518

[41] Yuming L, Chun C, Xuyuan Xu. No-reference video quality assessment with 3D shearlet transform and convolutional neural networks. IEEE Transactions on Circuits and Systems for Video Technology 2016; 26 (6): 575-589. doi: 10.1109/TCSVT.2015.2430711

[42] Varga A, Hornig R. An overview of the OMNeT++ simulation environment. In: SimuTools 2008 International Conference on Simulation Tools for Communications and Networks; Marseile, France; 2008. pp. 1-11.

[43] Alexis M, Karsten S, Gary S. H.264/14496-10 AVC Reference Software Manual. London, UK: Dolby Laboratories Inc., Fraunhofer Institute HHI, Microsoft Corporation, 2009.

[44] Wang Z, Simoncelli P, Bovik A. Multiscale structural similarity for image quality assessment. In: IEEE 2003 Conference on Signals, Systems and Computers; Asilomar, California, USA; 2003. pp. 1398-1402.

[45] Seshadrinathan K, Soundararajan R, Bovik A, Cormack L. Study of subjective and objective quality assessment of video. IEEE Transactions on Image Processing 2010; 19 (6): 1427-1441. doi: 10.1109/TIP.2010.2042111

[46] Moorthy A, Seshad K, Soundar R, Bovik A. Wireless video quality assessment: a study of subjective scores and objective algorithms. IEEE Transactions on Circuits and Systems for Video Technology 2010; 20 (4): 587-599. doi: 10.1109/TCSVT.2010.2041829

[47] Rajiv S, Alan C. Video quality assessment by reduced reference spatio-temporal entropic differencing. IEEE Transactions on Circuits and Systems for Video Technology 2013; 23 (4): 684-694. doi: 10.1109/TCSVT.2012.2214933

[48] Wang Z, Simoncelli P. Reduced-reference image quality assessment using a wavelet-domain natural image statistic model. Proceeding of SPIE 2005; 5666: 1578-1583. doi: org/10.1117/12.597306