International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

Spatiotemporal Video Quality Assessment Method via Multiple Feature Mappings

Daniel Oppong Bediako[a*], Yi Zhang[b], Xuanqin Mou[c]

[a,b,c]*Institute of image processing and Pattern Recognition, Xi'an Jiaotong University, Xi'an China*
[a]*Email: danieloppongbediako@yahoo.com* , [b]*Email: yi.zhang.osu@xjtu.edu.cn* , [c]*Email: xqmou@mail.xjtu.edu*

Abstract

Progressed video quality assessment (VQA) methods aim to evaluate the perceptual quality of videos in many applications but often prompt to increase computational complexity. Problems derive from the complexity of the distorted videos that are of significant concern in the communication industry, as well as the spatial-temporal content of the two-fold (spatial and temporal) distortion. Therefore, the findings of the study indicate that the information in the spatiotemporal slice (STS) images are useful in measuring video distortion. This paper mainly focuses on developing on a full reference video quality assessment algorithm estimator that integrates several features of spatiotemporal slices ($\mathrm{STS_S}$) of frames to form a high-performance video quality. This research work aims to evaluate video quality by utilizing several VQA databases by the following steps: (1) we first arrange the reference and test video sequences into a spatiotemporal slice representation. A collection of spatiotemporal feature maps were computed on each reference-test video. These response features are then processed by using a Structural Similarity (SSIM) to form a local frame quality. (2) To further enhance the quality assessment, we combine the spatial feature maps with the spatiotemporal feature maps and propose the VQA model, named multiple map similarity feature deviation (MMSFD-STS). (3) We apply a sequential pooling strategy to assemble the quality indices of frames in the video quality scoring. (4) Extensive evaluations on video quality databases show that the proposed VQA algorithm achieves better/competitive performance as compared with other state- of- the- art methods.

Keywords: Full reference video quality assessment; MMSFD-STS; spatiotemporal slice images.

----

* Corresponding author.

1. Introduction

The rapid distribution of advanced visual information being transmitted globally through communication devices and social networks has gain advantages due to its capabilities. Consequently, millions of videos are uploaded to these devices each day. However, with a large number of the advanced information transmitted to different end-users, the nature of the video quality must be the fundamental concern. Moreover, during the purchase, transmission, compression, and reproduction processes, videos experience various distortions, and the problem of video quality turns into a focal concern. Different from many signal-handling applications, the recipient of video signals is assumed to be human. Thus video quality algorithms must endeavor to assess the perceptual quality of videos in a manner that is consistent with human judgment. Powerful and efficient objective video quality assessment (VQA) methods are profoundly attractive in current visual communication frameworks for the execution of quality control assessment and classification purposes. Straightforward VQA algorithms might produce direct augments of image quality assessment (IQA) approaches on frame-by-frame bases but often lead to poor performance. Some of the approaches that go beyond frame-level processing will be reviewed further in related work. There are two methods to measure video quality. The first method is through subjective evaluation, which takes many subjects to give mean opinion scores (MOS) or difference mean opinion scores (DMOS) for each test video. Because of human involvement, the subjective test becomes impractical and cumbersome for many VQA applications and also consumes time, costs, and labor. The other method is to use the objective VQA algorithms which is more effective and efficient. The objective method has three categories. The full reference (FR) algorithms operate by measuring the difference between the distorted and the reference video in a way to imitate the human visual system (HVS). Usually, such approaches require reference video that for comparison. Reduce reference (RR) algorithm is suitable for some limited situations on the bandwidth that predict video quality by using partial reference video information. No reference (NR) algorithm assesses the quality without referring to the original high-quality image/video. In this paper, we address the task of FR VQA in which a reference video is available to the VQA algorithm. The simplest way for VQA is to measure the mean square error (MSE) or peak to signal noise ratio (PSNR) between the reference and distorted videos in a frame-by-frame manner. Notwithstanding the simplicity, the PSNR has a weak correlation with subjective visual quality. [1-3] The MSE cannot satisfactorily reflect perceptual quality since it does not consider the structural information hidden in neighboring pixels. Recent IQA algorithm was successfully used in VQA. Although most of the recent VAQ algorithms that depend on a frame-by-frame measurement is effective, there is still room to further improve their performances by considering the temporal information of videos in a more effective way. Wang and his colleagues [4] were the first to extend the use of structural similarity (SSIM) index to VQA through the application of perceptual weights to local regions and whole frames. Subsequently, the three-SSIM algorithm [5] further considered the impact of image content. Following the different sensitivities of perceptions in human vision, the algorithm divides the contents of video frames into edges, textures, and smooth areas. Recently, video parsing applications employ features extracted from a spatiotemporal slice (STS). Spatiotemporal MAD (ST-MAD) [6] employs a combination of image frame distortion and STS image similarity measures between the reference and distorted videos in the VQA model design. Although ST-MAD has excellent performance, the computation complexity is high due to the complexity of the MAD model and the HVS-based model. Wang and his colleagues [7] extracted spatial edge features and characteristics of temporal motion from localized space-time regions and used them to represent video structures. Notwithstanding some improvement, performance of the VQA model is generally lagging behind that of the best IQA models. To compute video STS image similarity, Peng Yan and his colleagues [8] proposed SSTS-GMSD/ STS-GMSD, a simple and efficient VQA framework. As the framework's starting point, the motion structural similarity of original and test video STS images is detected using an excellent structural similarity algorithm. Because the GMSD algorithm performs well in structural distortion detection, the GMSD similarity of STS images is used for video quality evaluation to measure these distortions. Given the high performance of STS-based VQA algorithms, they discovered that STS images contain useful information for a variety of features. Following a series of experiments to determine the effect of different object motion tracks, the VQA model was designed. Based on this studies, we propose an FR-VQA metric called MMSFD-STS that produces promising quality prediction accuracy with a simple computational framework. Our method, in particular, employs space-time differential change information in the form of spatial gradients and a temporal frame map, both of which are stored in the same STS data structure [9, 10]. We create a spatial STS gradient magnitude map as well as a temporal STS imaginary map. The gradient is widely used for measuring brightness and color variance [11]. It was also used as a successful descriptor in IQA models [12, 13]. Li and his colleagues [14] conducted a multi-distorted image quality assessment by combining the image gradient with the statistics of local binary pattern (LBP). The gradient of the STS offers comprehensive details on local time-space degradation of the video [15,16]. Based on the good performance of the gradient, we calculate the relative gradient (RM) map and the gradient orientation (GO) map on the STS along three dimension. In order to perform video quality assessment, we calculate the three-dimensional coefficients for each of the reference-distorted STS map pairs. Such pairs of reference-distorted STS feature maps are then analyzed using SSIM to form a quality of the local frame. Since multiple maps are produced containing values that may be distributed over different ranges, we therefore apply a sequential pooling strategy to assemble the quality indices of frames into the overall video quality. We make the following contributions as follows. First, we demonstrate that STS images alone can predict the perceptual quality of video sequences, which leads to an efficient VQA metric by using the MMSFD-STS metric to quantify the discrepancy between the reference and distortion STS images. We propose a general framework that transforms video sequence into STS representations of high- performance video quality predictor. A process of subsampling is applied in the STS domain to reduce computational complexity, without loss of quality prediction performance. Multiple quality-aware maps are analyzed for the STS representations, including gradient magnitude map, relative gradient map, imaginary enhance map, gradient orientation maps, and luminance maps in each pixel. All the adjacent frames are combined to form a group of frames (GOF). Finally, we use a sequential pooling approach to assemble the quality indices of the frame into one video quality score. The remaining of this paper is structured as follows: Section 2 presents related work; Section 3 specifies the method of the proposed VQA; Section 4 presents the experiment setup and result. Finally, Section 5 presents the discussion and conclusion.

2. Related Work

We introduce some well-known performed (VQA) as follows. First, the National Telecommunications and Information Administration (NTIA) [17] developed the Video Quality Metric (VQM) algorithm. The algorithm depends on the loss in computing the spatial gradients of the luminance features components and the color impairment of the VQM. The American National Standard Institute (ANSI) and International Telecommunications Union Recommendation [18, 19] have adopted VQM as a national standard in the VQEG Phase II validation tests.

Seshadrinathan and his colleagues [20] proposed the Motion-based Video Integrity Evaluation (MOVIE) index. In their experiment, the authors propose three MOVIE version: The Spatial MOVIE index, the Temporal MOVIE index, and the MOVIE index. Seshadrinathan and his colleagues presented a framework for video quality through spatio-temporal quality, temporal quality, and spatial quality measurement and integrated the groups of three-dimensional video frames into a Gabor filter bank. They estimate the overall quality by combining the spatial and temporal components. In the study [21], the authors suggested that SSIM be tailored to VQA ' true motion ' for SS-SSIM, assessing spatial quality and creating an easy temporal metric that would lead to a VQA algorithm that could be applied practically in conjunction with SSIM. They combined the algorithm of block-based motion evaluation with the SS-SSIM for the evaluation of temporal quality.

In comparison to an optical flow computation algorithm, these two operations require low computational complexity. Besides, they only use neighboring frames on the video rather than using a filter bank that needs considerable temporal support. Although the estimated time for temporal vision is about 200–300 milliseconds, the scheme will contribute to evaluating instant quality in a real-time scenario without waiting for sufficient frames to fill the buffer. The new algorithm, which evaluates structural retention between motion-compensated areas in a frame, is known as the motion-compensated structural similarity index (MC-SSIM).

Reference [22] investigated the nature of spatially localized flickers in pristine digital videos and the potential modeling of temporal visual masking of local flicker to improve VQA performance; the paper uses an enhanced VQA model by exploiting the psychophysical model of a temporal flicker mask. Specifically, the temporal flicker mask model used to expand the well-known MOVIE Index and developed a quantitative model of local flicker perception for the use of human subjective studies to more precisely predict Svideo quality. Flicker's influence on VQA also examined in terms of the compression bitrate, object motion, and temporal subsampling. These steps significantly improve the VQA models by taking into account the effects of temporary visual masking on flickering distortions in a perceptual manner by developing a MOVIE in the temporal dimensions associated with flickering. However, the method uses a spatiotemporal Gabor filter bank to compute bandpass filter responses on reference and distorted videos. Yan and his colleagues [8] decompose the video into several units and extract features from the spatiotemporal slice. Inspired by SSIM, the authors compute the gradient magnitude similarity (GMS) from the spatial-temporal slice and make use of GMSD to achieve the quality index in all directions. They make use of the temporal response to propose FR-VQA metric, name as STS-GMSD. To determine the influence of spatial information, the authors added the spatial part of GMSD and lead to another VQA metric, named SSTS-GMS. Although it does not improve performance, instead it reduces the linearity between the predicted score and the subjective score. Besides, the best performance on the LIVE database was V-Slice images that contain more information of motion than the H-Slice images; here, we show that the combination of all the slices can be employed to develop the high performance of VQA metric. In [23], the authors introduce a Model approach to VQA and decompose the video into spatiotemporal features using a 3D gradient that integrate both spatial and temporal slice information to measure video quality. Then for each group of frames (GOF), a three-dimensional gradient masking is employed to obtain each pixel in all directions. Based on the index, the authors combined the spatiotemporal gradient differencing GSDST between the reference and the distorted video block using the gradient in all directions. The authors believe that using a machine learning strategy can improve performance quality.

FR-VQA matrix is achieved by efficiently using the FR-IQA frame by frame method, and the results are weighted into one video index. These per-frame quality estimates can also be broken down over time to calculate an overall video quality assessment. Nevertheless, this basic line technique lacks temporal information and is not well correlated with the quality perceived by human observers. There are some well-known and useful FR-IQA metrics, including GSSIM [24], SSIM [2], MS-SSIM [25], VSNR [26], MAD [27], and VIF [28], some of which have been used in VQA model design. We will introduce briefly some state-of-the-art FR-VQA algorithms that will be used as competitors in this paper to compare with the proposed method.

Wang and his colleagues [29] proposed the video structural similarity (VSSIM) index with different SSIM indices at three levels. At the local region level, the SSIM index of each region is processed for luminance and chrominance components, with the luminance component being weighted higher. Finally, the SSIM frame weighted by global motion at the sequence level provides an estimate of the quality of the video. Reference [30] proposed an algorithm called the STS-Based Motion Structure Similarity (STS-MSPS) and found that STS images mainly have two areas of motion structure; the complex motion area, and the simple motion area. In their study, they have thus developed a motion structure partition based on the FR-VQA method for STS images and attempted to detect their influence in VQA. Information that originates from the structure's changes is used to measure the quality of the video. The GMSD algorithm is an excellent perceptual FR-IQA method, among other algorithms, to quantize the motion structure distortion. The authors tested the GMSD algorithm in the LIVE VQA database. The results show that the GMSD algorithm is the best algorithm for detecting structural similarity in STS images. Therefore, the GMSD algorithm is used to compute the motion structure partition similarity of STS images.

Once GMSD has been chosen to detect distortion of motion structure in STS images, the two different regions (simple and complex areas of motion structure) of STS images on VQA are further studied. Study results show that the complex motion areas have a significant more impact on VQA than the simple motion areas, which indicate that the HVS is more sensitive to distortion in the complex motion areas. Freita and his colleagues [31]. Use separate set features to calculate independently of each other, each set is concatenated to generate a vector feature. The feature vector is used as an input to predict the quality score in a random forest regression (RFR). Although the proposed approach outperforms state-of-the-art video quality metrics for data sets and distortion types, fails to do the best for the other metrics on the data sets. The ViS3 [32] video quality assessment algorithm uses STS images on a large timescale. Besides, the HVS model captures distortion in STS images. The ViS3 becomes more complicated as the design model is composed of the MAD and HVS models.

Based on the above discussion, we then concluded that the STS images are an accurate prediction for perceptual quality of video and then contribute to an efficient VQA algorithm by combining multiple map techniques to best represent VQA prediction model, such as gradient magnitude map, relative gradient map, imaginary enhance map, gradient orientation maps, and luminance maps in each pixel. The details will be addressed in the next section.

3. VQA Method Proposed

In this paper, we propose a general framework for video quality assessment which is based on analyzing multiple feature maps of spatiotemporal slices. The proposed model of the block diagram is shown in Fig 1.The reference and the distorted video are processed into STS representation. Further processes are performed in the STS representations to produce an imaginary Gabor enhance map, a gradient magnitude map, a relative gradient map (RM), a gradient orientation (GO) map, and a Luminance map. These feature maps are well defined and yield several outputs that are pooled using statistical based strategy to produce the final video quality score

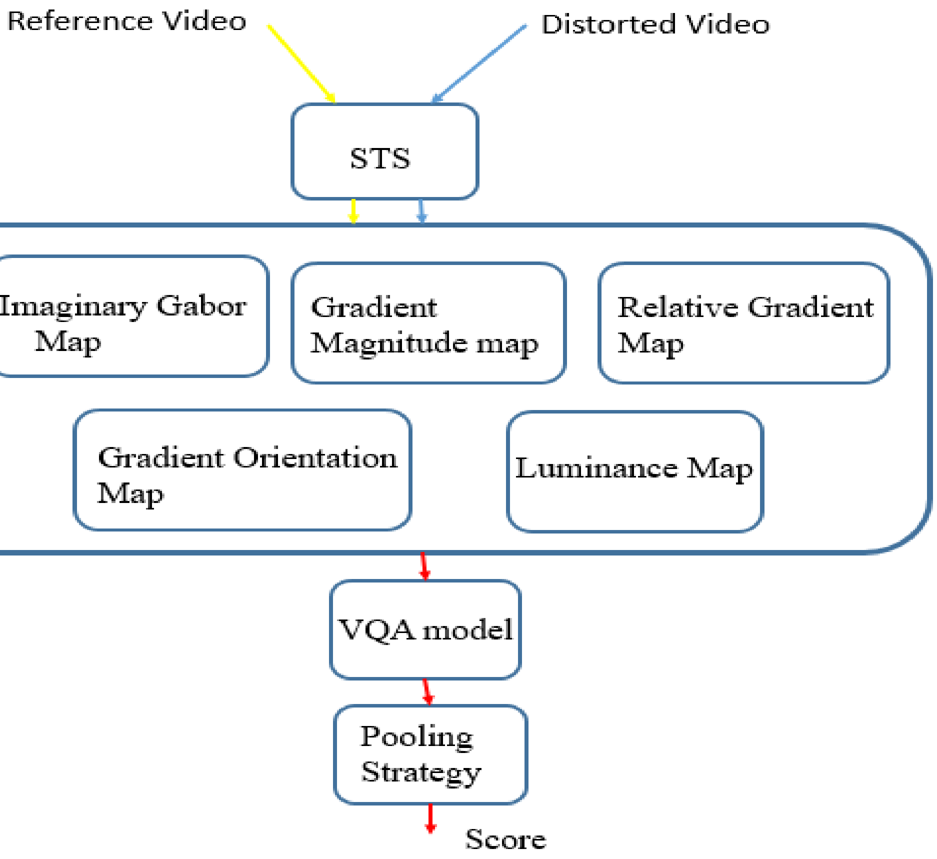


Figure 1: Block diagram of the proposed method.

3.1 Spatiotemporal slice images

We follow the method in [33] to produce images that contain spatiotemporal information. A video can display three variables F (x, y, and t). As shown in Figure.2 (a), the extracted slices are only standard video images when the video sliced perpendicularly to *t* dimensions. However, the slices of the video can also extract images containing spatiotemporal information perpendicular to the other two spatial dimensions. The new slices are called STS images. Figure.1 (b), shows the extracted slices of STS images taken from the perpendicular to the *x*-

dimensions and are referred to as the vertical STS images (V-STS).The extracted STS images were also taken perpendicular to the *y* dimension as shown in Figure.1(c) and is called horizontal STS images (H-STS).

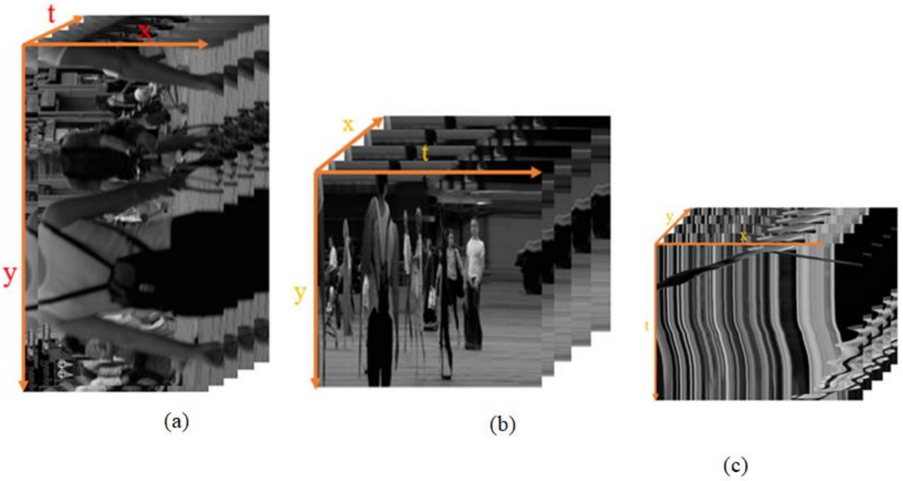


Figure 2: The slice images are seen in different dimensions of the video. (a) Standard video frames. (b) Vertical STS images. (c) Horizontal images of the STS.

4. Similarity Feature maps

Features developed from STS responses have been widely used in many applications of computer vision and image/video processing [34]. The most common distortion are structure effect with different orientations and scale, sudden changes in texture, contrast and lighting, all of which can affect the appearance of STS. To this end, we use much simpler method, which employs the SSIM index to measure various type of luminance distortions in X, Y, and T direction. The feature similarity measurement will be conducted by comparing the luminance maps L_r(x, y) and L_d (x, y) to formulate luminance similarity maps S_L(x, y).

A 2D Gabor filter can be described as a 2D Gaussian kernel function, modulated by a direction-oriented complex sinusoidal plane wave. The Gabor filter can be expressed as a set of two 2D filters—one involves the cosine term (i.e., generating the Gabor's real part, equivalently), and the other involves the sine term (i.e., the Gabor's imaginary part), respectively, as follows

$$G_{\mathrm{Re}}[i, j] = \frac{\left(i^2 + j^2\right)}{2\sigma^2} \cos\left(2\pi f \left(i \cos\theta + j \sin\theta\right)\right) \tag{1}$$

$$G_{\mathrm{Im}}[i, j] = \frac{\left(i^2 + j^2\right)}{2\sigma^2} \sin\left(2\pi f \left(i \cos\theta + j \sin\theta\right)\right) \tag{2}$$

The parameters *j* and *i* determine the directions of the Gabor kennel. The filter response is determined by the

Gaussian spread parameter and the radial frequency *f* of the sinusoid modulator. The orientation θ is chosen to create a set of Gabor filter. The Gabor filter, however, has a limitation on its bandwidth selection: when the bandwidth is set to a value that is larger than 1 octave, its DC response becomes non-negligible. To overcome the bandwidth limitation of the Gabor filter as mentioned above, we set the bandwidth to 0.5 and 1 octave respectively and chose the best result for this paper.

Since the HVS is highly sensitive to the edge information (e.g., [35, 36, 37]), this motivates us to exploit the imaginary part Gabor filter (i.e., $G_{\mathrm{Im}}[i, j]$) to convolve with the luminance component of the reference and the distorted counterpart along each direction

$$H_{\varsigma} = L_H \otimes G_{\mathrm{Im}} \tag{3}$$

$$V_{\varsigma} = L_V \otimes G_{\mathrm{Im}} \tag{4}$$

$$T_{\varsigma} = L_T \otimes G_{\mathrm{Im}} \tag{5}$$

Where the symbol $"\otimes"$ denotes the convolution operation; h_h and h_v denotes the prewitt filters along the horizontal and vertical directions. Using the representation of (1) and (2), the magnitudes of the reference image *r* and the distorted image *d*, denoted by *mr* and *md* along with the STS images, are computed as follows:

$$m_r = \sqrt{\left(G_{\mathrm{Re}}[i, j]\right)_r^2 + \left(G_{\mathrm{Im}}[i, j]\right)_r^2} \tag{6}$$

$$m_d = \sqrt{\left(G_{\mathrm{Re}}[i, j]\right)_d^2 + \left(G_{\mathrm{Im}}[i, j]\right)_d^2} \tag{7}$$

The gradient is an effective feature to capture shifts in structure, contrast changes and textural changes [38] that may arise from distortion. Here, we use the real and imaginary parts of the Gabor filter response coefficient to calculate two simple gradient-based metrics [12]: the relative gradient magnitude (RM) and the gradient orientation (GO) to capture local and global variations of the STS. Specifically, RM and GO are defined as follows:

$$t-direction: I_{RM}^{t}(i) = \sqrt{\left(I^{t}{}_{GRA}(i) - I_{\mathrm{Re}}^{t}(i)\right)^2 + \left(I_{GRA}^{t}(i) - I_{\mathrm{Im}}^{t}(i)\right)^2} \tag{8}$$

$$h-direction: I_{RM}^{h}(i) = \sqrt{\left(I_{GRA}^{h}(i) - I_{\mathrm{Re}}^{h}(i)\right)^2 + \left(I_{GRA}^{h}(i) - I_{Im}^{h}(i)\right)^2}$$

(9)

$$v-direction: I_{RM}^{v}(i) = \sqrt{\left(I_{GRA}^{v}(i) - I_{\mathrm{Re}}^{v}(i)\right)^2 + \left(I_{GRA}^{v}(i) - I_{\mathrm{Im}}^{v}(i)\right)^2}$$

(10)

$$I_{GO}(i) = \tan^{-1}\left(\frac{I_{\mathrm{Re}}(i)}{I_{\mathrm{Im}}(i)}\right)$$

(11)

Where $I_{\mathrm{Re}}^{t} / I_{m}(i)$, $I_{\mathrm{Re}}^{h} / I_{m}(i)$, and $I_{R_m}^{v} / I_{m}(i)$ indicate the real/imaginary part of the Gabor filter response along the horizontal, vertical, and temporal direction, respectively.

We utilize $\mathrm{G_{Im}}$ maps, RM maps, GO maps, GM map, and Luminance maps to enhance the STS content representation at all directions. These, together with the original STS and its complemented spatial-temporal representation, constitute the map pairs (reference and distorted) that are used to compute the multiple map similarity feature (MMSF) at all three dimensions. As indicated below, we obtain the MMSF of each pixel in the *x*, *y*, and *t* directions of the response maps mentioned above. *r* denotes the group of frames from the reference video and *d* denotes the distorted video. As in SSIM [25], we calculate the multiple map similarity features as follows:

$$MMSF = \frac{2m_r m_d + c}{m_r^2 + m_d^2 + c}$$

(12)

where m_r and m_d are the reference maps and distorted maps in $x, y,$ and t response, *c* represents a positive constant providing numerical stability. If m_r and m_d are identical, the map will reach the maximum value of 1. The brighter the gray level, the greater the similarity, and thus the higher local quality predicted by *MMSF*.

The MMSF is used to represent the distinction between STS images and to measure the video distortion severity. The outputs of T-STS, V-STS, and H-STS were combined to form the final video quality index of MMSF-STS.

4.1 Spatiotemporal multiple map similarity feature Deviation

As discussed above, there are three elements to predict the video quality (1) Spatial quality map, (2) H-STS quality map, and (3) V-STS quality map. Based on this assumption, we combine all the responses in the three elements respectively, and compute the multiple map similarity features of spatiotemporal slices MMSF-STS between the reference and distorted videos in *x, y,* and *t* directions as shown below;

$$MMSF\_STS = \left(\frac{2m_{rx}m_{dx} + C_1}{m_{rx}^2 + m_{dx}^2 + C_1}\right)^{\beta} \left(\frac{2m_{ry}m_{dy} + C_2}{m_{ry}^2 + m_{dy}^2 + C_2}\right)^{\alpha} \left(\frac{2m_{rt}m_{dt} + C_3}{m_{rt}^2 + m_{dt}^2 + C_3}\right)^{\gamma}$$

(13)

where α, β, and γ are local quality indices of weights of x, y, and t. Here we set α, β, and γ parameters to be 1.

To determine the average value of the first 10 percent, we sort the $MMSF\_STS$ values for each dimension in descending order and compute the average value of the first 10 percent. This process is called average percentile[39,40] and is usually applied to clarify the interpretation of scores on standardized tests. After obtaining the $MMSF\_STS$ score, we calculate the quality score and denote the percentile average by PAR%, along the dimension as:

$$MMSF\_STS = PAR10\%(MMSF\_STS)$$

(14)

$$\overline{MMSF\_STS} = \frac{1}{n}\sum_{i=1}^{n} GMSF\_STS(i)$$

(15)

where n is the total number of pixels in one video frame. A higher score provides better video quality. Average pooling assumes that in estimating the image quality, each pixel has the same perceptional quality value. However, as the general perception of how local quality degradation differs may differ in distinct regions, we use the standard deviation of MMSF-STS as the final quality index which is given by;

$$MMSFD\_STS = \left[\frac{1}{n}\sum_{1=1}^{n}\left(MMSF\_STS(i) - \overline{MMSF\_STS}\right)^{2}\right]^{\frac{1}{2}}$$

(17)

This outcome shows that the higher the MMSFD_STS score, the lower the perceptual image quality if distortion is negligible, the MMSFD_STS value will be 0. Fig 3 shows the framework of proposed multiple map similarity feature deviation (MMSFD-STS) for spatiotemporal slice model

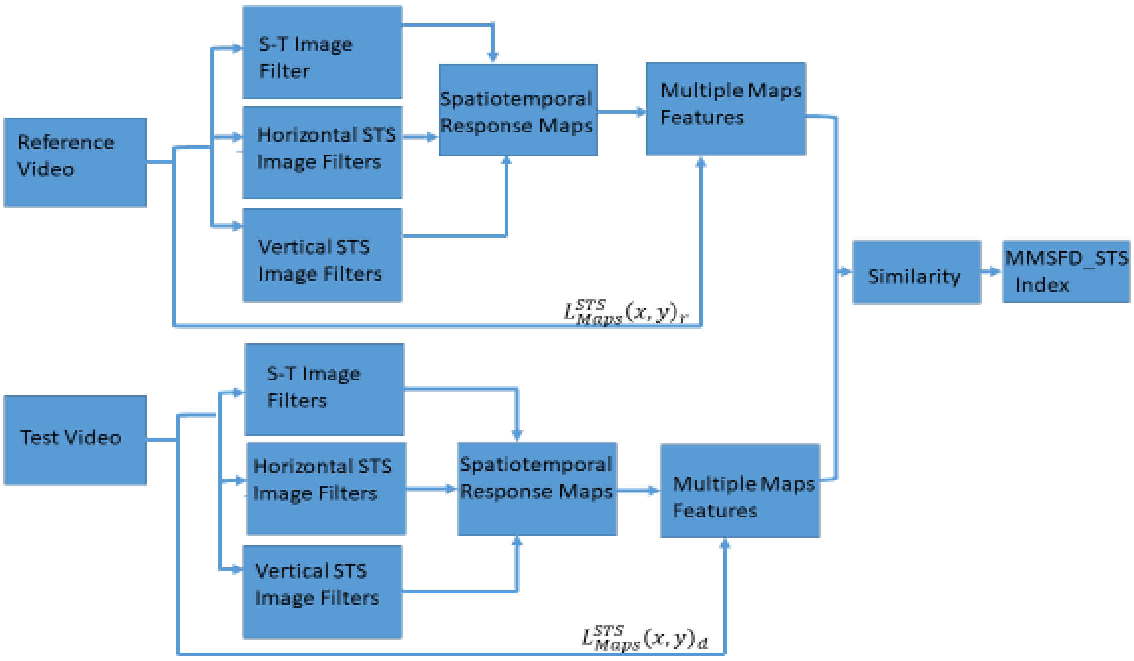


Figure 3: The framework of the proposed multiple map similarity feature deviation (MMSFD-STS) for spatiotemporal slice model.

5. Experiment and Results

In this section, we demonstrate the performance of our proposed method on two publicly available video databases. The first database is the LIVE VQA quality assessment database [41, 42] from TEXAS. The second database is Ecole Polytechnique Frale de Laussane (EPFL) and Politecneco di Milano (PoliMI) database [43, 44] over many other approaches for video quality prediction.

5.1 Performance of MMSFD_STS on the VQA Database

We use the EPFL-PoliMI VQA database and the LIVE VQA database as the test databases to evaluate the effectiveness of the proposed MMSFD_STS algorithm. The LIVE VQA database consists of a group of ten high-quality reference videos and a set of 150 distorted videos, all of which are YUV format whose spatial resolution is 768×432. The database has four types of distortion, which include compression artifacts due to H.264 and MPEG-2, the error introduced due to transmission over IP network and error imported over wireless networks. The mean DMOS values and standard deviation of the DMOS results are indexes in the LIVE video database. The EPFL VQA database consists of six original CIF spatial resolution video sequences (352×288 pixels) and six original 4CIF spatial resolution video sequences (704×576 pixels) coded with H.264/AVC and corrupted by simulating the packet loss caused by transmission over an error-prone network. Twelve corrupted bit streams were created for each original video sequence by dropping packets according to the error pattern. Therefore, 156 video sequences were finally scored by 40 subjects. Subjective scores were also made available as mean opinion scores (MOS) carried out at the premises of two academic institutions. For our experiment, however, only the Y component was used. We compared our algorithm with well-known VQA FR methods, as depicted in Tables 1 and 2. The experiment measured the objective and subjective values of the spearman rank order correlation SROCC and the Pearson linear correlation coefficient PLCC. We applied a four-parameter logistic transformation to the predicted raw results, as suggested by VQEG [45], before calculating any PLCC value. The transformed formula is as follows;

$$Q_j^{'} = \frac{(\beta_1 - \beta_2)}{1 + \exp\left[\dfrac{Q_J - \beta_3}{|\beta_4|}\right]} + \beta_2 \tag{18}$$

where Q_J indicates the original FR-VQA objective score, $Q_j^{'}$ is the expected video test subjective value for *j* and the parameters β_1, β_2, β_3 and β_4 are chosen as the free best fit subjective score to predict the overall subjective results. Our proposed method (MMSFD-STS) was compared with the conventional metrics (PSNR, SSIM, V-VIF, MS-SSIM, VSNR, MOVIE, VQM, and ST-MAD). Note that the origin of PSNR, VSNR, and MS-SSIM are derived from the IQA metrics and later extended by applying them on a frame-by-frame basis using the LIVE video database and average the scores for every frame. The MMSFD_STS algorithm comparison is presented in Table 1 in the LIVE database using SROCC. The PLCC comparison appears in Table 2. As shown in Tables 1 and 2, ST-MAD is the highest for distortion of H.262 and distortion of MPEG-2 in terms of SROCC and LCC. SSTS-GMSD is the highest for IP distortion in terms of LCC. Except for that, our MMSFD_STS method achieves the best performance for all. MMSFD-STS is the highest for MPEG-2 distortion, wireless, and IP distortion in terms of LCC and SROCC. Above all, our method emerged the highest performance in terms of ALL Data for SROCC and LCC, respectively. Besides, our method is simple compared to the traditional metric. To further probe the results of our proposed method, a scatter plot of the predicted scores versus the subjective DMOS scores by the four VQA metrics which can achieve good results is presented in Figure 4. As observed in the plots, the proposed method tested on both the LIVE database and the EPFL-PoLiMi database shows a better relationship between the metric and the subjective quality.

Table1: Comparison of LIVE database results by SROCC.

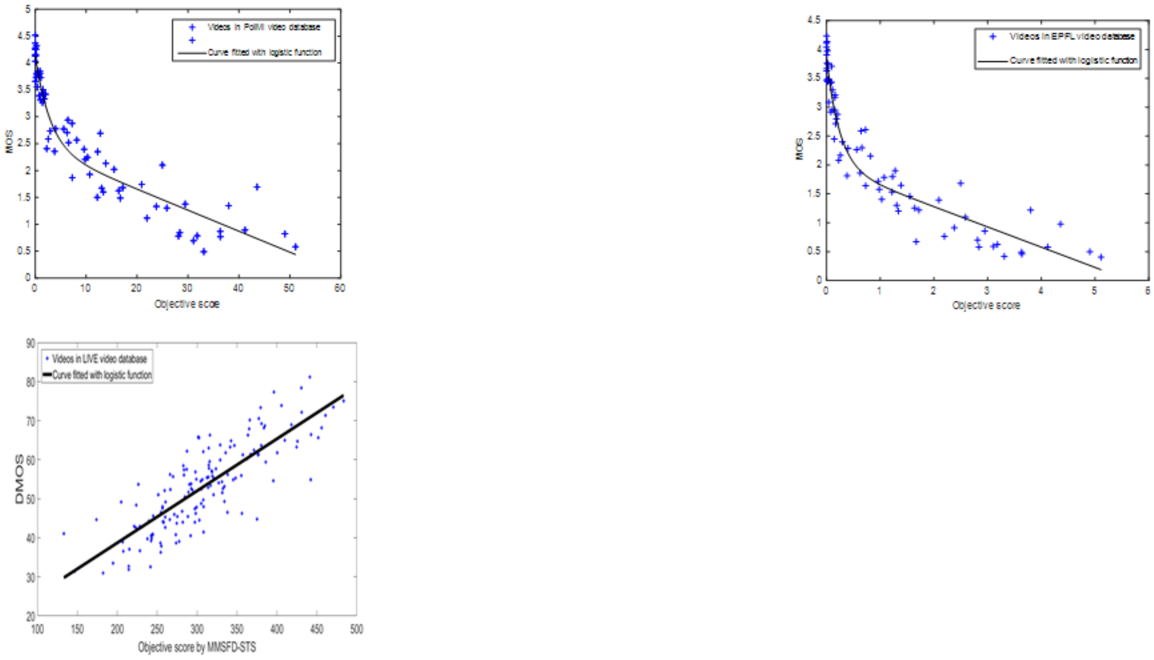
VQA	Wireless	IP	H.264	MPEG-2	ALL Data
SSIM	0.5233	0.4559	0.6514	0.5545	0.5257
MS-SSIM	0.7289	0.6534	0.7313	0.6684	0.736
V-VIF	0.5507	0.4736	0.6807	0.6116	0.571
VSNR	0.7019	0.6894	0.6460	0.5915	0.6755
PSNR	0.6574	0.4167	0.4585	0.3862	0.5397
VQM	0.7214	0.6383	0.6520	0.7810	0.7026
MOVIE	0.8109	0.7157	0.7664	0.7733	0.7890
ST-MAD	0.8099	0.7758	**0.9021**	**0.8461**	0.8251
SSTS-GMSD	0.8116	0.7855	0.7917	0.8151	0.8383
ViS3	0.8372	0.7885	0.7685	0.7362	0.8168
MMSFD_STS	**0.8476**	**0.7992**	0.7886	0.8314	**0.8432**

Table 2: Comparison of LIVE database results by LCC.

VQA	Wireless	IP	H.264	MPEG-2	ALL Data
SSIM	0.5401	0.5119	0.6656	0.5491	0.5444
MS-SSIM	0.7170	0.7219	0.6919	0.6415	0.5756
V-VIF	0.5488	0.5102	0.6911	0.6415	0.5756
VSNR	0.6992	0.7341	0.6216	0.5980	0.6896
PSNR	0.6690	0.4645	0.5492	0.3891	0.5621
VQM	0.7324	0.6480	0.6459	0.7860	0.7236
MOVIE	0.8386	0.7622	0.7902	0.7595	0.8116
ST-MAD	0.8591	0.8065	**0.9155**	**0.8560**	0.8332
SSTS-GMSD	0.8403	**0.8267**	0.8166	0.8218	0.8416
ViS3	0.8473	0.8164	0.7890	0.7510	0.8263
MMSFD_STS	**0.8596**	0.7684	0.7890	0.8273	**0.8499**

Table3: Performance comparison on EPFL-PoliMl video database.

	SROCC		PLCC	
Algorithm	EPFL	PoliMl	EPFL	PoliMl
VSi3	0.9023	0.8731	0.9214	0.8570
ST-MAD	0.9075	0.9465	0.9017	0.9489
MMSFD_STS	0.9665	0.9584	0.9588	0.9680
SSTS-GMSD	0.9419	0.9476	0.9547	0.9545



(a) (b) (c)

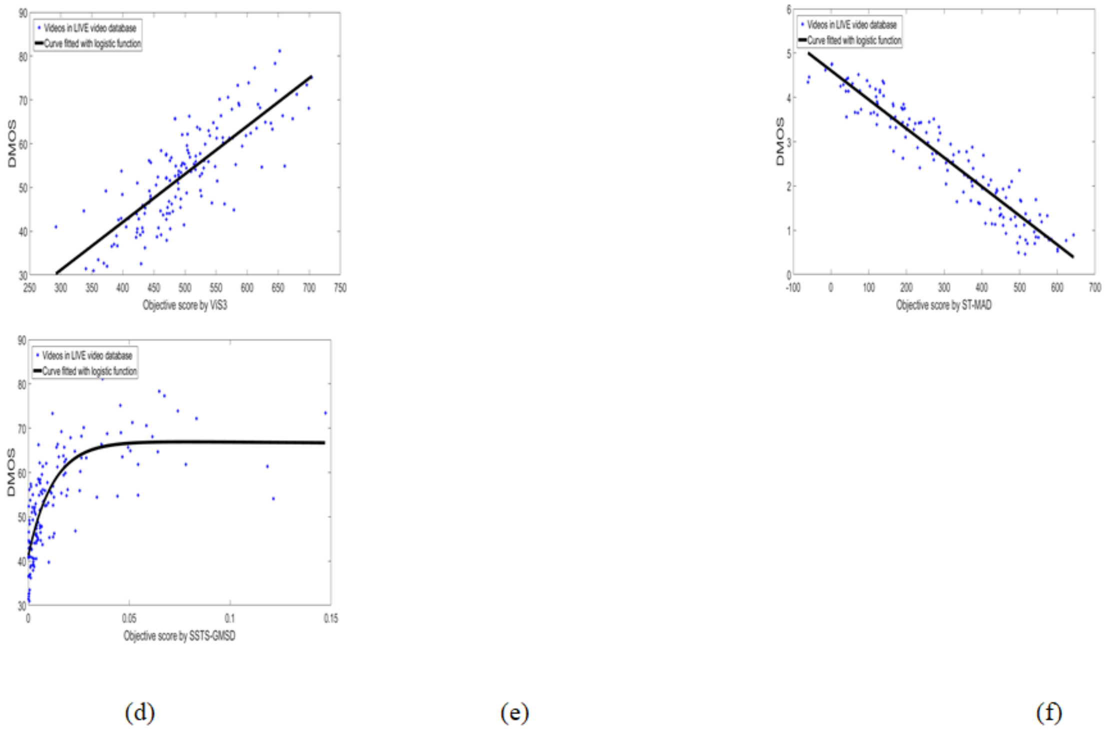


Figure 4: The scatter plots of the objective VQA scores versus DMOS for all videos in the LIVE video quality database and EPFL-PoliMi video quality database. (a) Proposed method tested on the EPFL database, (b) Proposed method tested on the PoliMl database. (c) Proposed method tested on LIVE database. (d) ViS3 method tested on LIVE database. (e) ST-MAD method tested on LIVE database. (f) SSTS-GMSD method tested on LIVE database.

5.3. Performance of MMSFD_STS on the EPFL-PoliMi video database

Tables 3 show MMSFD_STS performance in the EPFL-PoLiMi database. The EPFL-PoLiMi database provides subjective outcomes from two academic institutions; we separate our experiment using the mean opinion value (MOS) of both academic institutions. In that study, we conducted a more intensive EPFL-PoLiMi VQA database experiment and then validated the effectiveness of MMSFD_STS with SSTS-GMSD, ViS3 and ST-MAD algorithms, as these state-of- the- art FR-VQA model provide results comparable to the proposed metric in terms of the correlation scores shown in Tables 3. These study indicate that MMSFD_STS is the best-performing algorithm in terms of SROCC and LCC. Since the MOS of the two academic institutions is slightly different, the SROCC and LCC fluctuate slightly. We, therefore, claim that our multiple map response well to smaller and higher size video database resolution. The score on the SROCC and LCC demonstrates that the correlation of spatiotemporal features of our method is an effective way to predict video quality.

We demonstrated that MMSFD-STS3 performs well in predicting video quality by testing it on various video-quality databases. It not only excels at VQA for entire databases with varying types and levels of distortion, but it also performs well on videos with a specific type of distortion. Our performance evaluation shows that MMSFD-STS3 outperforms or is statistically tied with current state-of-the-art VQA algorithms. Our analysis also reveals that MMSFD-STS outperforms *ST-MAD*, ViS3, and SSTS-GMSD in predicting the quality of videos from specific databases.

5.4. Running Time

We examine the proposed algorithm running time and compare it with the existing algorithms. We selected tr16_25fps.yuv from both the reference and distorted videos in the LIVE database. All the algorithms were performed on Dell computer with an Intel(R) i5-8500 CPU@3.00GHz, 8.00 GB RAM, 64-bit Microsoft window 10, and 2019 MATLAB software. The video to be tested consist of 216 frames with 432×768 resolution. All MATLAB source code were downloaded from the author's website. As shown in Table 4, our proposed algorithm has the fastest running time than other competitive algorithms.

Table 4: Runtime comparison of the STS-based VQA algorithms.

VQA algorithms	Running time (s)	Ratio to MMSFD-STS
ST-MAD	399.45	2.22
ViS3	342.05	1.90
SSTS-GMSD	270.60	1.51
MMSFD-STS	179.84	1

5.5. *Discussion*

As shown in the LIVE VQA database, our proposed algorithm (MMSFD_STS) shows a strong relationship with human perception of video than most of the algorithms studied in other papers. The performance of the MMSFD_STS indicates that by analyzing the dissimilarities of STS images, temporal video distortion can be well captured. In both video databases, we found that the object background has more horizontal and vertical motion. The MMSFD-STS perform well in predicting video quality when combining all the directional features. It not only excels at VQA for LIVE database, but it also performs well on EPFL-PoliMI video database. Performance on the EPFL-PoliMI video database demonstrates the efficacy of the proposed algorithm. The values of the parameters α, β, and γ are set to 1, 1, and 1, respectively, and are modified to carry out this experiment. In this experiment, three different values of α, β, and γ are used. To conduct our experiment, we chose the best performance to perform the experiment. MMSFD_STS shows better predictions on EPFL-PoliMI databases than LIVE database. MMSFD showed higher SROCC and LCC than either ST-MAD or ViS3 alone. We observe different relative performances depending on the database. Nonetheless, ST-MAD has the highest SROCC and LCC for H.262 distortion and MPEG-2 distortion. SSTS-GMSD, on the other hand, exhibits high IP distortion in terms of LCC. The SSTS-GMSD extracts features from videos that are promising for FR-VQA model design using gradient magnitude; in this case, we used multiple algorithms to extract the features. The combined features shows the best results as compared with the state of the algorithm including SSTS-GMSD model.

Although our proposed approach is the best models and the best running time performance compared to other state-of-the-art models, it still fails to perform well for some of the distortions in the LIVE database. However, in ALL data, our method still performs well. MMSFD-STS, however, is not without limitations. One significant limitation is the potentially large memory requirements for long videos. The STS images of a long video may necessitate a prohibitively large width or height for the time dimension. One solution in this case would be to divide the video into small chunks across time. Another limitation of MMSFD-STS is that it only considers the luminance component of the video at the moment. Additional gains may be realized by taking chrominance degradations into account. Another potential enhancement would be to use a more accurate pooling model of the spatiotemporal responses used in the spatiotemporal dissimilarity stag. In our future work, we will extend the experiment validation on the other video databases. We would also apply our experiment to individual directions and compare it to the state of the art.

6. Conclusion

In this paper, we studied and proposed a novel computational efficient digital video quality assessment (VQA) algorithm named multiple map similarity feature deviation (MMSFD-STS). The involved procedure of the algorithm includes the framework for VQA based on spatiotemporal slices. We developed a framework that can capture multiple features in digital video and constructed a MMSFD-STS index. We combined all the features and constructed a VQA index. We also demonstrated the ability of the VQA index to predict human opinion scores on an extensive database of videos. The performance of several leading objective VQA algorithms was evaluated using the results of the study. The proposed algorithm developed as part of this paper has proven to perform very well in this study and shown to be competitive with other state-of-the-art methods.

References

[1]. B. Girod, “Psychovisual Aspects Of Image Processing: What’s Wrong With Mean Squared Error?,” in Proceedings of the Seventh Workshop on Multidimensional Signal Processing, p. P.2-P.2, IEEE, Lake Placid, NY (1991) [doi:10.1109/MDSP.1991.639240].

[2]. Z. Wang et al., “Image Quality Assessment: From Error Visibility to Structural Similarity,” IEEE Trans. on Image Process. **13**(4), 600–612 (2004) [doi:10.1109/TIP.2003.819861].

[3]. A. C. Bovik, *The essential guide to image processing*, Academic Press, London ; Boston (2009).

[4]. W. Xue and L. Zhang, “Gradient Magnitude Similarity Deviation: An Highly Efficient Perceptual Image Quality Index,” 12.

[5]. A. C. Bovik, “Content-weighted video quality assessment using a three-component image model,” J. Electron. Imaging **19**(1), 011003 (2010) [doi:10.1117/1.3267087].

[6]. 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, pp. 2505–2508, IEEE, Brussels, Belgium (2011) [doi:10.1109/ICIP.2011.6116171].

[7]. Y. Wang, T. Jiang, S. W. Ma, and W. Gao, “Novel spatio-temporal structural information based video quality metric,” IEEE Transactions on Circuits and System for Video Technology, vol. 22, no. 7, pp. 989-998, Jun. 2012

[8]. P. Yan, X. Mou, and W. Xue, “Video quality assessment via gradient magnitude similarity deviation of spatial and spatiotemporal slices,” presented at IS&T/SPIE Electronic Imaging, 11 March 2015, San Francisco, California, United States, 94110M [doi:10.1117/12.2083283].

[9]. D. A. Migliore, M. Matteucci, and M. Naccari, “A revaluation of differencing frame in fast and robust motion detection,” In Proceedings of the 4th ACM International Workshop on Video Surveillance and Sensor Networks, Oct. 2006, pp. 215-218

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

[11]. C. W. Niblack, R. Barber, W. Equitz, and M. D. Flickner, “QBIC project: querying images by content, using color, texture, and shape,” In Storage and Retrieval for Image and Video Databases, Apr. 1993, pp. 173-188.

[12]. L. Liu, Y. Hua, Q. Zhao, H. Huang, and A. C. Bovik, “Blind image quality assessment by relative gradient statistics and adaboosting neural network,” Signal Processing: Image Communication, vol. 40, no. 1, pp. 1-15, Jan. 2016.

[13]. 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, Feb. 2014.

[14]. Q. Li, W. Lin, and Y. Fang, “No-reference quality assessment for multiply-distorted images in gradient domain,” IEEE Signal Processing Letters, vol. 23, no. 4, pp. 541-545, Apr. 2016.

[15]. C. Lee, S. Cho, J. Choe, T. Jeong, W. Ahn, and E. Lee, “Objective video quality assessment,” Optical Engineering, vol. 45, no. 1, article no. 017004, Jan. 2006.

[16]. W. Xue, X. Mou, L. Zhang, A.C. Bovik, and X. Feng, “Blind image quality prediction using joint statistics of gradient magnitude and laplacian features,” IEEE Transactions on Image Processing, vol. 23, no. 11, pp. 4850–4862, Nov. 2014.

[17]. .M. H. Pinson and S. Wolf, “A New Standardized Method for Objectively Measuring Video Quality,” IEEE Trans. on Broadcast. **50**(3), 312–322 (2004) [doi:10.1109/TBC.2004.834028].

[18]. 18.“RECOMMENDATION ITU-R BT.1907 - Objective perceptual video quality measurement techniques for broadcasting applications using HDTV in the presence of a full reference signal,” 26.

[19]. .M. H. Pinson, N. Staelens, and A. Webster, “The history of video quality model validation,” in 2013 IEEE 15th International Workshop on Multimedia Signal Processing (MMSP), pp. 458–463, IEEE, Pula (CA), Italy (2013) [doi:10.1109/MMSP.2013.6659332].

[20]. K. Seshadrinathan and A. Bovik, “Motion Tuned Spatio-Temporal Quality Assessment of Natural Videos,” Image Processing, IEEE Transactions on **19**, 335–350 (2010) [doi:10.1109/TIP.2009.2034992].

[21]. A. K. Moorthy and A. C. Bovik, “Efficient Video Quality Assessment Along Temporal Trajectories,” IEEE Trans. Circuits Syst. Video Technol. **20**(11), 1653–1658 (2010)[doi:10.1109/TCSVT.2010.2087470].

[22]. L. K. Choi and A. C. Bovik, “Video quality assessment accounting for temporal visual masking of local flicker,” Signal Processing: Image Communication **67**, 182–198 (2018) [doi:10.1016/j.image.2018.06.009].

[23]. W. Lu et al., “A spatiotemporal model of video quality assessment via 3D gradient differencing,” Information Sciences **478**, 141–151 (2019) [doi:10.1016/j.ins.2018.11.003].

[24]. G. Chen, C. Yang, and S. Xie, “Gradient-Based Structural Similarity for Image Quality Assessment,” in 2006 International Conference on Image Processing, pp. 2929–2932, IEEE, Atlanta, GA (2006) [doi:10.1109/ICIP.2006.313132].

[25]. 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, pp. 1398–1402, IEEE, Pacific Grove, CA, USA (2003) [doi:10.1109/ACSSC.2003.1292216].

[26]. D. M. Chandler and S. S. Hemami, “VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images,” IEEE Trans. on Image Process. **16**(9), 2284–2298 (2007) [doi:10.1109/TIP.2007.901820].

[27]. D. M. Chandler, “Most apparent distortion: full-reference image quality assessment and the role of strategy,” J. Electron. Imaging **19**(1), 011006 (2010) [doi:10.1117/1.3267105].

[28]. H. R. Sheikh and A. C. Bovik, “IMAGE INFORMATION AND VISUAL QUALITY,” 4.

[29]. Z. Wang, L. Lu, and A. C. Bovik, “Video quality assessment based on structural distortion measurement,” Signal Processing: Image Communication **19**(2), 121–132 (2004) [doi:10.1016/S0923-5965(03)00076-6].

[30]. P. Yan and X. Mou, “Video quality assessment based on motion structure partition similarity of spatiotemporal slice images,” J. Electron. Imag. **27**(03), 1 (2018) [doi:10.1117/1.JEI.27.3.033019].

[31]. P. G. Freitas, W. Y. L. Akamine, and M. C. Q. Farias, “Using multiple spatio-temporal features to estimate video quality,” Signal Processing: Image Communication **64**, 1–10 (2018) [doi:10.1016/j.image.2018.02.010].

[32]. P. V. Vu and D. M. Chandler, “ViS3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices,” J. Electron. Imaging **23**(1), 013016 (2014) [doi:10.1117/1.JEI.23.1.013016].

[33]. “Feature Extraction and Analysis using Gabor Filter and Higher Order.pdf.”

[34]. M. J. Wainwright and E. P. Simoncelli, “Scale mixtures of Gaussians and the statistics of natural images,” Advances in Neural Information Processing Systems, Nov. 2000, pp. 855- 861

[35]. A. Liu, W. Lin, and M. Narwaria, “Image quality assessment based on gradient similarity,” IEEE Trans. Image Process. vol. 21, no. 4, pp. 1500–1512, Apr. 2012.

[36]. W. Xue, L. Zhang, X. Mou, and A. C. Bovik, “Gradient magnitude similarity deviation: A highly efficient perceptual image quality index,” IEEE Trans. Image Process., vol. 23, no. 2, pp. 684–695, Feb. 2014.

[37]. Z. Ni, L. Ma, H. Zeng, C. Cai, and K.-K. Ma, “Screen content image quality assessment using edge model,” in Proc. IEEE Int. Conf. Image Process., Aug. 2016, pp. 81–85.

[38]. B. Tao and B. W. Dickinson, “Texture recognition and image retrieval using gradient indexing,” Journal of Visual Communication and Image Representation, vol. 11, no. 3, pp. 372-342, Sep. 2000

[39]. A. K. Moorthy and A. C. Bovik, “Visual Importance Pooling for Image Quality Assessment,” IEEE J. Sel. Top. Signal Process. **3**(2), 193–201 (2009) [doi:10.1109/JSTSP.2009.2015374].

[40]. A. K. Moorthy and A. C. Bovik, “Perceptually significant spatial pooling techniques for image quality assessment,” presented at IS&T/SPIE Electronic Imaging, 5 February 2009, San Jose, CA, 724012

[doi:10.1117/12.810166].

[41]. K. Seshadrinathan et al., "Study of Subjective and Objective Quality Assessment of Video," IEEE Trans. on Image Process. **19**(6), 1427–1441 (2010) [doi:10.1109/TIP.2010.2042111].

[42]. K. Seshadrinathan et al., "A subjective study to evaluate video quality assessment algorithms," presented at IS&T/SPIE Electronic Imaging, 4 February 2010, San Jose, California, 75270H [doi:10.1117/12.845382].

[43]. F. De Simone et al., "Subjective Quality Assessment of H.264/AVC Video Streaming with Packet Losses," EURASIP Journal on Image and Video Processing **2011**, 1–12 (2011) [doi:10.1155/2011/190431].

[44]. F. De Simone et al., "A H.264/AVC video database for the evaluation of quality metrics," in 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2430–2433, IEEE, Dallas, TX, USA (2010) [doi:10.1109/ICASSP.2010.5496296].

[45]. K. Brunnstrom et al., "VQeg validation and ITU standardization of objective perceptual video quality metrics [Standards in a Nutshell]," IEEE Signal Process. Mag. **26**(3), 96–101 (2009) [doi:10.1109/MSP.2009.932162].

List of Figures

List of Tables

Daniel Oppong Bediako received his Bachelor of Engineering (BEng) in Electrical Electronic Engineering from Accra Institute of Technology, Accra, Ghana, in 2013 and his MS degree in Information and Communication Engineering from Xi'an Jiaotong University, Xi'an, Shaanxi, China, in 2017. Currently, he is pursuing a Ph.D in information and communication engineering at Xi'an Jiaotong University. His research interests include image quality assessment and video quality assessment.

Yi Zhang received the B.S. and M.S. degrees in electrical engineering from Northwestern Polytechnical University, Xi'an, China, in 2008 and 2011, respectively, and the Ph.D. degree in electrical engineering from Oklahoma State University, Stillwater, OK, USA, in 2015. From 2016 to 2018, he was a Postdoctoral Research Associate with the Department of Electrical and Electronic Engineering, Shizuoka University, Japan. He is currently a Faculty Member with the School of Electronic and Information Engineering, Xi'an Jiaotong University, China. His research interests include 2D/3D image processing, machine learning, pattern recognition, and computer vision

Xuanqin Mou has been with the Institute of Image Processing and Pattern Recognition (IPPR), Electronic and Information Engineering School, Xi’an Jiaotong University, since 1987. He has been an associate professor since 1997 and a professor since 2002. Currently, he is the director of IPPR. Prof. Mou has authored or coauthored more than 200 peer-reviewed journal or conference papers. He has been granted as the Technology Academy Award for invention by the Ministry of Education of China, and the Technology Academy Awards from the Government of Shaanxi Province, China