Multimed Tools Appl (2018) 77:30791–30814
https://doi.org/10.1007/s11042-018-6107-1

An error-based video quality assessment method with temporal information

Woei-Tan Loh[1] · **David Boon Liang Bong**[1]

Received: 26 July 2017 / Revised: 27 March 2018 / Accepted: 8 May 2018 /
Published online: 1 June 2018

Abstract Videos are amongst the most popular online media for Internet users nowadays. Thus, it is of utmost importance that the videos transmitted through the internet or other transmission media to have a minimal data loss and acceptable visual quality. Video quality assessment (VQA) is a useful tool to determine the quality of a video without human intervention. A new VQA method, termed as Error and Temporal Structural Similarity (EaTSS), is proposed in this paper. EaTSS is based on a combination of error signals, weighted Structural Similarity Index (SSIM) and difference of temporal information. The error signals are used to weight the computed SSIM map and subsequently to compute the quality score. This is a better alternative to the usual SSIM index, in which the quality score is computed as the average of the SSIM map. For the temporal information, the second-order time-differential information are used for quality score computation. From the experiments, EaTSS is found to have competitive performance and faster computational speed compared to other existing VQA algorithms.

Keywords Video quality · Temporal effects · Temporal distortions · Multimedia content

1 Introduction

Many multimedia applications deal with visual assets nowadays. This is more evident in mobile devices such as smartphones, tablets, smart watches, and smart glasses. These devices use microprocessors or microcontrollers with high processing ability. This processing ability enables them to process data on their own without relying on external processing devices. According to the research done by Ooyala [18], mobile video views increased from 6.3% of the overall mobile traffic data in 2012 to 45.1% in Q3 2015. This shows the importance of videos in multimedia equipment. Assessment of the video qualities of these products is crucial as a quality feedback tool for device manufacturers and content service providers.

✉ David Boon Liang Bong
davidblbong@yahoo.com; bbldavid@unimas.my

[1] Faculty of Engineering, Universiti Malaysia Sarawak, Kota Samarahan, Malaysia

Since humans are ultimate users of these visual related devices, their quality ratings of images and videos are the most accurate. The quality assessment which involves humans as evaluators is known as subjective quality assessment. In spite of its accuracy, subjective quality assessment cannot be performed to all assessment tasks as it is costly and time-consuming. Thus, automatic quality assessment methods without human involvement are highly desirable. This requirement motivates the development of objective quality assessment method. Instead of obtaining quality scores from humans, algorithms are used to rate images and videos automatically. An objective method is more cost and time effective, especially for real-time tasks. However, the problem of inaccuracy in terms of human perception is still a concern for objective methods. The accuracy or the performance of objective methods is measured through the correlation of objective scores and subjective scores (ground truth results). In recent decades, hard works are devoted by researchers to improve the accuracy of objective quality assessment methods. The proposed method in this paper also aims to predict video quality with high correlation to subjective scores.

A video composed of a series of images. The changes of images or frames over time create an additional temporal dimension. Temporal information is extremely useful for interpreting motion on which many applications are based on. One typical example of these applications is activity recognition [15, 16]. In terms of quality, temporal information can affect the perceived quality of a video. Different types of temporal distortion are found on videos, such as compensation mismatch, jitter, and mosquito noise. A short review of various types of temporal distortions and their causes can be found in [23]. On the other hand, the motion from temporal effects can mask distortions too. This phenomenon is known as motion silencing. Motion silencing is gaining attention and interest from researchers recently. A popular illusion of motion silencing is demonstrated in Suchow and Alvarez's work [30]. Hence, temporal information should be considered during video quality assessment for its adverse and advantageous effects.

In this paper, a new video quality assessment (VQA) algorithm is proposed. It is dubbed as Error and Temporal Structural Similarity (EaTSS). EaTSS is derived from the authors' previous work [17] in which only compressed videos are dealt with. There are several highlights of EaTSS. Firstly, EaTSS is used to assess videos of all types of distortions. This property is highly desirable. There are image quality assessment (IQA) and VQA metrics which only deal with specific distortion. The no reference IQA metric in [3] is one of them. This metric only assesses qualities of Gaussian blur distorted images. These types of metrics have limited applications. Besides that, localized error based weight is incorporated in EaTSS. Structural Similarity Index (SSIM) [40] map is weighted by this weight for spatial quality evaluations. This weighting is very different from other weightings used in the variants of SSIM such as the works in [12, 39]. For the temporal part, EaTSS involves second-order time-differential information of a video. To the best of our knowledge, no research on VQA to date employed second-order time-differential information. There are only VQA methods that make use of first order time-differential information [4, 7, 36]. Lastly, EaTSS also has low complexity and good efficiency as shown in Section 4.4.

The remainder of this paper is organized as follows. Section 2 presents a brief review of the related previous works by other researchers and the authors [40]. EaTSS is elaborated in Section 3. In Section 4, correlation results of EaTSS are presented and discussed. Computational time and complexity are also shown in this section. In Section 5, a general conclusion is presented.

2 Previous work

VQA methods can be categorized into full reference (FR), reduced reference (RR), and no reference (NR) methods [34]. FR methods conduct the assessment task with all information of pristine videos available. Instead, part or no information from ground truth data is accessible for RR and NR methods respectively. This paper concentrates on FR methods so some state-of-the-art FR VQA methods are briefly discussed in this section. They are classified into IQA based and non-IQA based methods. A previous work by the authors [17] upon which the proposed method is built is also detailed in this section.

2.1 Previous work by other researchers

Some researchers had proposed VQA metrics which are based on IQA metrics. Existing IQA metrics can assess video quality too. However, they disregard temporal distortions and effects in a video. This is the main reason for inaccuracy and inappropriateness of IQA metrics for VQA tasks. Some popular IQA metrics that are commonly utilized and modified in VQA metrics are Mean Squared Error (MSE) or Peak Signal to Noise Ratio (PSNR) [38], SSIM [40], Multiscale SSIM (MSSIM) [42], Most Apparent Differences (MAD) [10], and Visual Information Fidelity (VIF) [27]. Some of their extensions are briefly discussed here.

By extending MSE, Rimac-Drlje et al. proposed Foveated Mean Squared Error (FMSE) [22]. FMSE makes use of center bias and eccentricity of the human visual system (HVS) for spatial quality assessment. To consider temporal effects, foveation-based contrast sensitivity of the method in [11] is applied for scenes with movement. Wang et al. adapted SSIM for VQA in Video Structural Similarity (VSSIM) index [41]. VSSIM further incorporates chrominance components while the luminance component is given more weight. The spatial quality score of each frame is weighted by global motion to generate an overall video quality score. Vu and Deshpande had proposed ViMSSIM [36] which builds upon MSSIM. Spatial quality scores are obtained by a modified exponential moving average procedure for MSSIM indices of every frame. For the temporal part, MSSIM indices for the reference and distorted frame difference information are computed. Recently, Vu and Chandler derived ViS_3 [35] from MAD. Firstly, motion magnitudes from optical flow are combined with the MAD indices to obtain the first score. The second score is the resulting dissimilarities of reference and distorted spatiotemporal slices (STS) frames. Then, two scores are integrated into an overall video quality. Another method based on VIF is put forward by Sheikh and Bovik [26]. This method incorporates source model, distortion model, and information fidelity criterion. The mutual information from wavelet subband coefficients of reference and distorted videos is employed for measuring spatial quality. For temporal quality scores, temporal distortions are measured by the information loss due to motion. The loss is computed by deviations in spatiotemporal derivatives of a video.

Nevertheless, there are VQA metrics that are not based on any IQA approach. VQM [20], proposed by Pinson and Wolf, is perhaps the first non-IQA based VQA metric. Reference and distorted videos are sampled and calibrated first. This is followed by the extraction of perception based features [44], computation of video quality parameters, and calculation of VQM models. Watson et al. had proposed Digital Video Quality (DVQ) [43] which embodies just noticeable difference (JND). In DVQ, discrete cosine transform (DCT) coefficients of a video are first filtered by contrast sensitivity function (CSF). Later, a JND threshold is applied to the filtered coefficients to generate a quality score. Another well-known VQA method that is

non-IQA based is motion-based video integrity evaluation (MOVIE) [23]. MOVIE is based on statistics of natural videos. Spatial impairments are calculated from the combination of contrast masking and 3D Gabor filtered videos. 3D Gabor filtered videos also interact with motion information from the optical flow to account for temporal distortions. The overall video quality is the geometric mean of spatial and temporal scores. Pinson et al. [21] had extended VQM to VQM variable frame delay (VQM-VFD) in 2014. VQM-VFD further embodies two additional perception based parameters for measuring frame delay. A neural network is trained to integrate all parameters into a quality score. More recently, Choi and Bovik [5] had improved the MOVIE framework by injecting the flicker sensitive index. They prove that the temporal masking of flicker impairments improves VQA performances.

From the works mentioned above, it is obvious that temporal effect plays an important role in VQA. This is evident from VQM-VFD and the method by Choi and Bovik (dubbed C-B method hereafter). They supplement their previous methods with additional temporal effect consideration. Their performance improvements motivate us to improve the temporal part of our previous work. Furthermore, the spatial part of the previous work is also improved in this paper. This is motivated by the high performances of the VQA methods that implicate Gabor filters. However, for the sake of complexity, only the localization nature of Gabor filters [22] is incorporated in the proposed method. Thus, weighting based on the local errors is implemented. Before introducing the proposed method, our previous work is elaborated concisely in the next section.

2.2 Previous work by the authors

EaTSS extends the previous method by the authors. This previous method is known as MSE Difference SSIM (MD-SSIM) [17]. There are two main parts in MD-SSIM, i.e. local and global. The overall video quality is the arithmetic mean of local and global quality scores. Local quality scores are derived from spatial and temporal quality scores. For spatial quality scores, they are computed by weighting SSIM map with the MSE map. This is shown as [17]:

$$MDSSIM(spatial) = \frac{\sum_x \sum_y E(x,y) S(x,y)}{\sum_x \sum_y E(x,y)} \tag{1}$$

where $MDSSIM(spatial)$ refers to spatial quality scores of the local part of MD-SSIM, $E(x, y)$ is the error map from MSE, and $S(x, y)$ is the SSIM map from SSIM.

Temporal quality scores are the differences between the spatial quality score of every successive frame. Then, local quality scores are defined as the weighting of spatial quality scores by temporal scores. This is shown as [17]:

$$MDSSIM(local) = \frac{\sum_f^F W_f MDSSIM(spatial)}{\sum_2^F W_f} \tag{2}$$

In (2), $MDSSIM(local)$ is the local quality score of MD-SSIM, W_f is the temporal quality score at frame f, and F is the total number of frames. For the global part, the quality score is the average of SSIM indices on a frame-by-frame basis. It is defined as [17]:

$$MDSSIM(global) = \frac{1}{F} \sum_f^F SSIM(f) \tag{3}$$

where *SSIM*(*f*) is the SSIM quality score of frame f. The overall MD-SSIM score of a video is the arithmetic mean of local and global quality scores [17]:

$$MDSSIM = \frac{1}{2}[MDSSIM(local) + MDSSIM(global)] \tag{4}$$

Based on the results in [17], the average of the SSIM map in usual SSIM metric is insufficient to capture distortions perceived by HVS in videos. MSE weighting of the SSIM map is an alternative to compute the SSIM index from SSIM map. This method is proven [17] to correlate better with subjective scores for compressed videos. Based on MD-SSIM, EaTSS extends to evaluate qualities of videos suffered from various distortions, other than compressed videos only. Next, EaTSS is deliberated.

3 Error and temporal structural similarity - EaTSS

In this section, EaTSS is discussed in detail. Similar to most of the existing metrics, EaTSS composes of spatial and temporal components. The spatial component is extended from MD-SSIM [17] while the temporal component is inspired by the works of Vu and Deshpande [36], Cardoso et al. [4], and Ítalio et al. [7]. The overall video quality score is the arithmetic mean of spatial and temporal scores. Spatial part, temporal part, and their combination are discussed sequentially in the following subsections. The overall workflow of EaTSS is shown in Fig. 1.

3.1 Spatial part of EaTSS

The weighting method from [17] is extended for the spatial part of EaTSS. In [17], MSE of reference and distorted frames are used as the weight. According to Akramullah [1], MSE is inaccurate because a particular pixel in a frame is visually affected by its surrounding pixels.

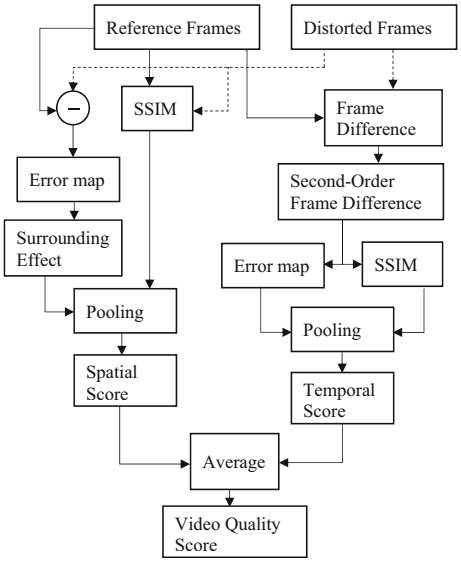


Fig. 1 Workflow of EaTSS

The work by Akramullah is supported by Limb's experiment [14]. This experiment shows that image quality ratings by humans are the average of errors in local areas with the highest error values. In other words, the image quality is proportional to the level of distortion in distorted regions. This result is extended to videos for EaTSS. Each frame in a video is exposed to HVS for a very short instant of time. Thus, we assume that HVS is not able to focus on every local part of the whole frame. Instead, HVS concentrates on local salient components in a frame. In EaTSS, distortions are deemed as the salient components. Thus, we compute the weighting in accordance with the degree of local errors to model localization. There is a distinction between this method and the result in [14]. Global effect is considered in our weighting method. Yet, more weight is given to distorted local regions. This corresponds to the characteristic of HVS that acts as a bandpass filter [6] when it searches for salient regions. This characteristic is modeled as MSE of the error of a pixel to errors of its surrounding pixels. Also, as stated in Section 2, the incorporation of localization characteristic is motivated by high performances of VQA methods which involve Gabor filter. Since high localization is the main property of Gabor filter, it is modeled and incorporated in EaTSS.

Dissimilar to MD-SSIM [17], error maps from the MSE of reference and distorted frames are not used for weighting. Instead, the differences from the subtraction of the two frames are used. This aims to include direction factors in the weighting. Direction factors indicate whether errors are causing the original luminance values to become brighter (positive direction) or darker (negative direction). Errors in different directions are considered the same in MSE maps due to the squaring of errors. By direct subtraction of frames, directions of the errors are used for the later computation of localization model. The computation of the new error map is shown as:

$$E(x,y) = R(x,y) - D(x,y) \tag{5}$$

In (5), $E(x,y)$ is the error map while $R(x,y)$ and $D(x,y)$ are the frames of reference and distorted videos respectively.

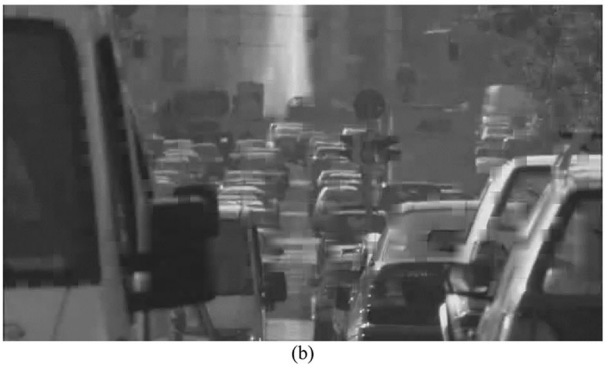
Next, the localization model is applied to the error map. MSE of the error in a pixel with the errors of its surrounding pixels is calculated. This is shown as:

$$W(x,y) = \frac{\sum_{i=x-1}^{x+1}\sum_{j=y-1}^{y+1}[E(x,y)-E(i,j)]^2}{s} \tag{6}$$

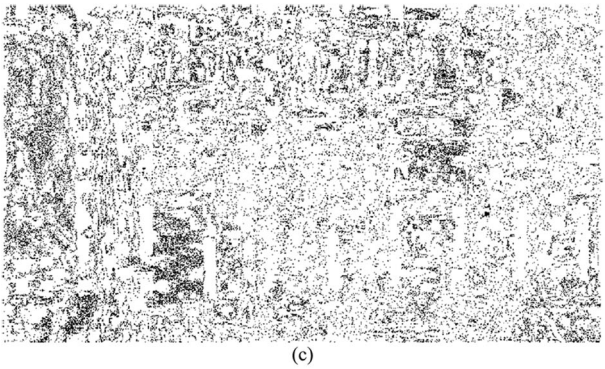
In (6), $W(x,y)$ is the weighting function, parameters i and j are the pixel positions around the target pixel, and s is the number of pixels surrounding the target pixel. The squaring of the numerator in (6) is to prevent negative values. This process does not conflict with the aforementioned statement of direction factors. This is because the subtraction in the numerator in (6) has already taken into consideration of the direction factor. The goal of squaring is to prevent inaccurate normalization in the computation of spatial score later. The clarification of localization is shown in Fig. 2. Figure 2a and b show the first frame of the reference and distorted videos from the LIVE video database [24, 25]. Figure 2c shows the error map of both frames. The illustrated error map is the result of inverting and modulus of results from (5). Figure 2d is the weight

Fig. 2 Effects of localization: **a** first frame from the original video **b** first frame from distorted video "Reproduced with permission, courtesy of Seshadrinathan et al. [6, 24]" (c) error map of **a** and **b** after modulus and inversion and **d** inverted and normalized weight function from **c** ▶

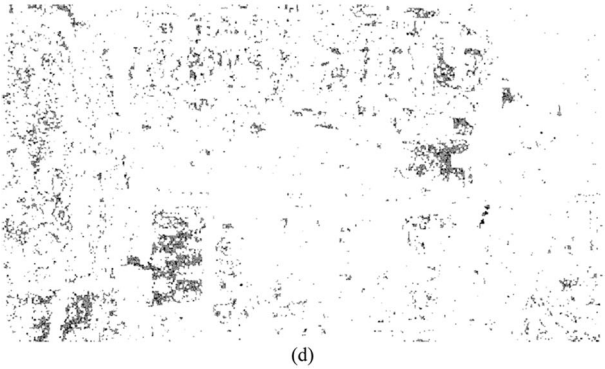
(a)



(b)



(c)



(d)

function calculated using (6). The illustrated weight function is inverted and normalized for illustration purpose. The regions where distortions are more severe have been emphasized in Fig. 2d. Moreover, regions with less significant errors have less weight according to the weight function.

After that, SSIM maps can be computed. It is shown in the equation below [40]:

$$SSIM(x,y) = \frac{\left(2\mu_x\mu_y + C_1\right)\left(2\sigma_{xy} + C_2\right)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)} \tag{7}$$

where $SSIM(x,y)$ is the SSIM map, x and y refer to a particular frame from reference and distorted videos respectively, μ is the mean intensity, σ is the standard deviations of the intensities, and σ_{xy} is the cross correlation of intensities from the reference and distorted videos. Parameters C_1 and C_2 are constants added into (7) to prevent instability. The main reason for using SSIM instead of other IQA metrics is that it has good performance with low complexity. Although MSE and PSNR are simple and fast to compute, their performances are unsatisfactory. On the other hand, other good performing IQA metrics like MSSIM and MAD have a much higher complexity than SSIM. In order to strike a balance between performance and complexity, SSIM is chosen in our method.

Then, $W(x,y)$ from (6) is used to weight SSIM maps from (7) to obtain a spatial quality score. This is done by the weighted summation of $W(x,y)$ and $SSIM(x,y)$. The weighted summation is also referred as local distortion-based pooling in [39]. Although the pooling method proposed in [39] can definitely improve the performance of SSIM maps, the aspects that we are considering here is different from the work in [39]. EaTSS focuses more on distortions while the work in [39] focuses more on the content in relation to HVS. The authors in [39] also compared their works with local distortion-based pooling. However, the weight used is totally different. Moreover, EaTSS focuses more on local distortions. The weighted summation is shown as:

$$EaTSS_{spatial} = \frac{\sum_{x=1}^{M}\sum_{Y=1}^{N}[W(x,y) \times SSIM(x,y)]}{\sum_{x=1}^{M}\sum_{Y=1}^{N}W(x,y)} \tag{8}$$

where M and N are the width and height of the frames of the video. This weighting causes the resulting spatial score to focus more on severely distorted regions in a frame.

3.2 Temporal part of EaTSS

Temporal information is very useful in order to consider temporal impairments and effects in a video. This is evident from the previous works by other researchers [5, 11, 20–23, 26, 35, 36, 41, 43, 44]. The most direct method to obtain temporal information of particular frames is through frame subtraction. This method is the generalization of the equation in [8] which measures temporal information of a video:

$$TI = max_f\left\{std_{x,y}\left[A_{f-1}(x,y) - A_f(x,y)\right]\right\} \tag{9}$$

where TI is the temporal information, max_f is the maximum over time dimension and $std_{x,y}$ is the standard deviation over space dimension. Parameter $A_f(x,y)$ is the current frame of a video

while $A_{f-1}(x,y)$ is the previous frame of the same video. Thus, the temporal information in terms of pixels can be measured by frame subtraction disregarding max_f and $std_{x,y}$. To capture temporal distortions and effects better, Vu and Deshpande had proposed an alternative in [36]. Two types of temporal information are computed for comparison; reference and distorted temporal information. Reference temporal information, TI_R, is computed by generalizing (9). This is shown as:

$$TI_R(x,y) = \left[R_f(x,y) - R_{f-1}(x,y)\right] \tag{11}$$

Distorted temporal information is obtained from two frames. One of them is the current time frame from the distorted video, $D_f(x,y)$. The other frame is the previous time frame from reference video, $R_{f-1}(x,y)$. It is shown as:

$$TI_D(x,y) = \left[D_f(x,y) - R_{f-1}(x,y)\right] \tag{10}$$

where TI_D is distorted temporal information. Both TI_R and TI_D are actually time-differential or frame differenced information. By utilizing frames from reference and distorted videos, TI_D further incorporates spatial information that is affected by temporal transitions. This corresponds to HVS functions whereby temporal and spatial sensitivities affect each other [37]. According to these functions, HVS's temporal sensitivity depends on spatial information while temporal information changes human spatial contrast sensitivity functions [37].

The alternative by Vu and Deshpande has high correlations to subjective scores. This is shown in the results in [36]. A recent RR VQA method, spatiotemporal reduced reference entropic differences (STRRED) [29], also utilized frame differences. Wavelet coefficients from frame differences are used to capture spatial and temporal information distinction of reference and distorted videos. Authors in [29] called the frame differences as time-differential information. The excellent performances of these two VQA metrics motivate us to further extend time-differential information in a different fashion to [29, 36]. We choose to extend the method by Vu and Deshpande in this paper. This method does not involve domain transformation and has lower computational cost. Most VQA methods obtain temporal information in terms of motion from optical flow or other types of motion estimation techniques. These estimation techniques are known to be complex and need long computational time. For instance, the work in [19] states that MOVIE spends more than 55% of its computation time in computing optical flow. As shown in (10) and (11), there is no multiplication or division needed for time-differential information. Thus, this method has a very low complexity that is desirable for real-time applications.

To extend time-differential information, second-order time-differential information or the differences of temporal information are computed. To date, research in the second-order time-differential information for quality assessment is still lacking. This choice is enlightened by the good performances in object recognition [13] and categorization [2] by utilizing second-order features. Thus, an assumption that second-order information is useful in predicting the video quality is made. Second-order time-differential information is defined as:

$$VTI_D(x,y) = \left[D_f(x,y) - R_{f-1}(x,y)\right] - \left[D_{f-1}(x,y) - R_{f-2}(x,y)\right] \tag{12}$$

$$VTI_R(x,y) = \left[R_f(x,y) - R_{f-1}(x,y)\right] - \left[R_{f-1}(x,y) - R_{f-2}(x,y)\right] \tag{13}$$

where $VTI_D(x,y)$ and $VTI_R(x,y)$ represent second-order of distorted and reference temporal information respectively. The first part of the right hand side of (12) and (13) correspond to

temporal information at the current time frame. Meanwhile, the second part of both equations refers to temporal information of the previous time frame.

In [36], MSSIM index is adopted for temporal information comparison. Meanwhile, STRRED incorporated Gaussian scale mixture model of wavelet coefficients of frame differences [29]. In this paper, a new variant of SSIM is used. The rationale for this choice is to reduce computational complexity. MSSIM is well known for its high computational complexity, although with good performance. Similarly, wavelet transformations also involve complex computations. Direct difference and MSE are not utilized to prevent dissimilar range to that of the spatial quality scores. A simpler variant of SSIM is utilized to maintain the performance while reducing computational complexity. SSIM index is simplified instead of utilizing it directly as in the spatial part. This is because resulting maps after computing second-order temporal information are similar to whitened images. No luminance component left in VTI_D and VTI_D after subtractions in (12) and (13). They are similar to whitened images in which luminance information is absent. In the original SSIM index, there are luminance, contrast, and structure comparisons of two frames. Consequently, the luminance comparison function can be discarded. The mean is set to zero in parameters σ_{xy}, σ_x, and σ_y. Standard deviation is defined as the correlation of pixels to their mean value. Instead of mean, the correlation of pixels second-order time-differential information to a static scene is more desirable in this case. This is to reflect characteristics of temporal information better. For a static scene, there is no temporal information. So, every pixel in a static scene will have zero values. In consequence, the mean value is replaced by zero. In the case of the cross correlation, similar reason holds. Cross correlation is defined as the correlation of standard deviations of reference and distorted temporal information. In each standard deviation, temporal information is compared to a static scene. Therefore, the mean value in the original equation can be replaced by zero. In overall, the new variant is simplified from (7) as:

$$SSIM\_no\_lc(x,y) = \frac{(2xy + C_2)}{(x^2 + y^2 + C_2)} \tag{14}$$

where *SSIM_no_lc* is the variant of SSIM map with no luminance comparison. Parameters x and y in (14) correspond to distorted and reference second-order time-differential information at a particular pixel location. The resulting map function, $map(x,y)$, that compares $VTI_D(x,y)$ and $VTI_R(x,y)$, is shown by the equation below:

$$map(x,y) = SSIM\_no\_lc(VTI_D(x,y), VTI_R(x,y)) \tag{15}$$

Similar to the spatial part of EaTSS, the localized weighting function is utilized to weight $map(x,y)$. The reason for this weighting is similar to the spatial part. Due to the short instant of frames changes, HVS can only focus on certain regions of the video. The more distorted regions in terms of temporal information are considered as the salient regions that HVS will concentrate on. The weighting function is shown as:

$$e(x,y) = VTI_D(x,y) - VTI_R(x,y) \tag{16}$$

$$w(x,y) = \frac{\sum_{i=x-1}^{x+1}\sum_{j=y-1}^{y+1}[e(x,y) - e(i,j)]^2}{s} \tag{17}$$

where $e(x, y)$ is the error map of $VTI_D(x, y)$ and $VTI_R(x, y)$, $w(x, y)$ is the weighting function for $map(x, y)$, and the parameters i, j and s are the same as in (6). The temporal quality score of EaTSS is defined as:

$$EaTSS_{temporal} = \frac{\sum_{x=1}^{M}\sum_{y=1}^{N}[w(x,y) \times map(x,y)]}{\sum_{x=1}^{M}\sum_{y=1}^{N} w(x,y)} \tag{18}$$

where the definition of M and N are the same as in (8). The weighting function $w(x, y)$ has the similar function to $W(x, y)$. It enables the temporal quality score to focus more on the distorted region. It also weights local salient temporal information more than of global temporal information.

3.3 Overall video quality

The overall video quality score of EaTSS is the combination of spatial and temporal quality scores. The same importance is given to both quality scores. As stated in Section 3.2, they can affect each other. Since both spatial and temporal quality scores are normalized to the same range, the geometric mean is unnecessary. Thus, the arithmetic mean of the spatial and temporal quality scores is used. The overall video quality score is defined as:

$$EaTSS = \frac{1}{2} \times (Spatial + Temporal) \tag{19}$$

4 Results and discussion

The performance of EaTSS is evaluated by comparing its correlation results with existing VQA metrics. The results are based on two benchmark video databases, the LIVE video database [24, 25] and the CSIQ video database [35].

4.1 Details of the databases

The LIVE video database was released by the University of Texas, Austin with all video files having planar YUV 4:2:0 format and do not contain any headers. The spatial resolution of all videos is 768 × 432 pixels. The types of distortions and their respective numbers of videos and frames are listed in Table 1. There are four levels of severity for each type of distortion except for the IP distortion with only three levels of severity. There are 10 reference videos in the

Table 1 Details of the LIVE video database

Distortion	No. of Videos	No. of Frames
Wireless	40	12,868
Internet Protocol (IP)	30	9651
H.264 compression	40	12,868
Moving Picture Experts Group type 2 (MPEG-2) compression	40	12,868
Wireless	40	12,868
Total	150	48,255

database with 15 distorted videos for each reference video. Overall, there are 150 videos that consist of 48,255 frames that need to be tested.

Second benchmark database, the CSIQ video database, was released by the 318 Advanced Technology Research Center from Oklahoma State University. All videos are in the YUV 4:2:0 format with the resolution of 832×480. All videos have 10 s duration. There are 12 reference videos and 216 distorted videos. Each reference video is distorted by six types of distortions as listed in Table 2. For each type of distortion, there are three levels of severity. Altogether, there are 82,260 frames to be tested. All distortions in both benchmark video databases are common types of errors found in video transmission and display.

4.2 Evaluation metrics

In order to measure the performance of EaTSS and compare to existing VQA methods, correlations of VQA methods quality scores and human-rated subjective scores from the benchmark databases are computed. As recommended in [33], two types of correlation coefficients are utilized. They are Spearman rank-order correlation coefficient (SC) and Pearson linear correlation coefficient (PC). SC computes prediction monotonicity. Its values reflect the degree of objective and subjective scores can fit a monotonic function. On the other hand, PC measures prediction accuracy to indicate linearity between objective and subjective scores. So as to avoid inaccuracies due to the nonlinearity of objective scores and subjective scores, objective scores need to be transformed. This is the standard procedure recommended by [33]. This procedure is also being used by other VQA methods for performance comparison such as in [5, 21, 23, 29, 35]. Before computing PC, objective quality scores are fitted to subjective scores by using a four-parameter logistic function which is defined as [33]:

$$f(x) = \tau_2 + \frac{\tau_1 - \tau_2}{1 + e^{\left[-\frac{x-\tau_3}{|\tau_4|}\right]}} \tag{20}$$

where τ_1, τ_2, τ_3, and τ_4 are regression parameters to be fitted. In this paper, these parameters are fitted using nonlinear least squares optimization based on the subjective scores from the two databases utilized in this paper. This fitting complies with the ITU guidelines [33] and is also implemented by most objective assessments [5, 21–23, 26, 35, 41, 43]. Parameter x is the objective scores and $f(x)$ are fitted scores.

F-test is utilized in this paper to test the statistical significance of VQA methods. It tests the ratio of variances of two sets of scores at 95% significant level. The null hypothesis considers two set of scores are indistinguishable. During the testing, larger variances are put in the

Table 2 Details of the CSIQ database

Distortion	No. of Videos	No. of Frames
H.264/ Advanced Video Coding (AVC) compression	36	13,710
Packet Loss Rate (plr)	36	13,710
Motion Joint Photographic Experts Group (MJPEG) compression	36	13,710
Wavelet compression	36	13,710
White Noise	36	13,710
High Efficiency Video Coding (HEVC) compression	36	13,710
Total	216	82,260

numerator. The procedures follow descriptions in [24, 34]. Interested readers can refer to [24, 34] for detailed explanations. First of all, differences of objective scores after the nonlinear transformation and subjective score, D, are computed. This is shown as [24, 34]:

$$D = \{f(x)_k - Sub_k, k = 1, 2 \ldots K\} \tag{21}$$

where $f(x)$ is the transformed objective scores from (20), *Sub* is the subjective score from databases, and K is the total number of videos to be tested. These differences are assumed to follow normal distributions. As utilized by [28], the kurtosis-based criterion is used to test Gaussianity. If the kurtosis lies between 2 and 4, the dataset is Gaussian.

EaTSS is compared with some popular IQA (applied in frame-by-frame basis) and VQA metrics. The IQA metrics to be compared are MSE/PSNR, SSIM, and MSSIM. VQA metrics involved include MOVIE [23] in addition to recently proposed VQA metrics. They are ViS_3 [35], STRRED [29], VQM-VFD [21] and C-B method [5]. The RR VQA method is included as it can be applied in FR manner. Moreover, it involves time-differential information which is similar to second-order time-differential information used in EaTSS. The previously proposed metric, MD-SSIM [17], is also being compared to show improvements of EaTSS. It is implemented to videos of all distortions types. All VQA metrics are applied in their default implementations. For VQM-VFD, only results for the LIVE video database are computed. This is because the platform for testing is of insufficient memory while computing results from the CSIQ database. Since C-B method coding is not publicly available, its results for the LIVE video database are directly quoted from [5]. Therefore, VQM-VFD and C-B method are not being compared in the CSIQ database. Spatial and temporal quality scores of EaTSS are also compared in both databases. This is to show relative contributions of spatial and temporal parts of EaTSS to EaTSS. Spatial and temporal parts of EaTSS are denoted as EaTSS (Spatial) and EaTSS (Temporal) respectively.

4.3 Performance evaluation

Tables 3, 4, 5, 6 and 7 show SC and PC of all metrics for the LIVE and CSIQ video databases. Figure 3 shows scatter plots of EaTSS against human-rated subjective scores for the LIVE and CSIQ video database.

Table 3 Spearman rank order correlation for the LIVE database

Metric	Wireless	IP	H.264	MPEG-2
MSE/PSNR	0.6574	0.4167	0.4585	0.3317
SSIM	0.6516	0.6160	0.7109	0.5933
MSSIM	0.7280	0.6543	0.7336	0.5898
MOVIE [23]	**0.8109**	0.7157	0.7664	**0.7733**
ViS_3 [35]	**0.8394**	0.7918	0.7685	0.7362
MD-SSIM [17]	0.7364	0.6743	0.7812	0.7686
STRRED [29]	0.7857	0.7722	0.8193	0.7193
VQM-VFD [21]	0.7510	0.7922	0.6525	0.6361
C-B Method [5]	0.7949	0.7513	**0.8265**	0.7671
EaTSS (Spatial)	0.7704	**0.8243**	0.7345	0.6338
EaTSS (Temperal)	0.7113	0.7838	**0.8255**	0.5137
EaTSS	0.7728	**0.8309**	0.7992	**0.7716**

Bolded values indicate the best two correlation values

Table 4 Pearson linear correlation for the LIVE database

Metric	Wireless	IP	H.264	MPEG-2
MSE/PSNR	0.5940	0.3836	0.4112	0.3520
SSIM	0.6653	0.6855	0.7377	0.5980
MSSIM	0.7122	0.7282	0.7341	0.6766
MOVIE [23]	0.8386	0.7622	0.7902	0.7596
ViS_3 [35]	**0.8574**	0.8349	0.7993	0.7574
MD-SSIM [17]	0.7438	0.7260	0.7903	**0.7901**
STRRED [29]	0.8039	0.8020	0.8228	0.7467
VQM-VFD [21]	0.8144	**0.8616**	0.7403	0.7172
C-B Method [5]	**0.8533**	0.8193	**0.8624**	0.7973
EaTSS (Spatial)	0.7935	0.7686	0.7389	0.6677
EaTSS (Temperal)	0.7275	0.7566	**0.8249**	0.5397
EaTSS	0.7962	**0.8499**	0.8181	**0.8076**

Bolded values indicate the best two correlation values

The best two VQA metrics for each distortion are highlighted in Tables 3, 4, 5, 6 and 7. The values that are in italic form are the best performing metrics for the comparison of EaTSS to IQA metrics only. In terms of SC in the LIVE video database, EaTSS has better performance than all IQA metrics (MSE/PSNR, SSIM, and MSSIM) as shown in Table 3. Compare with existing VQA metrics (MOVIE, ViS_3, MD-SSIM, ST-MAD, STRRED, VQM-VFD, and C-B method), EaTSS achieves the best results for IP distorted videos. For MPEG-2 distorted videos, EaTSS has a very close performance to the best performing metric, i.e. MOVIE. EaTSS performs better than VQM-VFD and MD-SSIM for wireless distorted videos. Yet, it falls behinds MOVIE and ViS_3. Meanwhile, it achieves the third best results in H.264 compressed videos. Overall, in terms of SC, EaTSS and MOVIE are the best performing metrics. They both attain the best two results in half of the total distortion types. While comparing spatial and temporal parts of EaTSS, it is obvious that EaTSS outperforms the implementation of solely spatial or temporal part. The improvement is particularly significant for MPEG-2 compressed videos. Both spatial and temporal parts of EaTSS perform poorly, but EaTSS attain the second best result. The possible reason is that EaTSS (Spatial) and EaTSS (Temporal) consider complementary aspects of the distortions. Thus, their combination is more effective.

Table 4 tabulates PC for all distortions in the LIVE video database. Similar to SC, EaTSS outperforms all IQA metrics. Compared to existing VQA metrics, EaTSS is still among the best two for IP and MEPG-2 distortions. However, there are some differences in relation to

Table 5 Spearman rank order correlation for the CSIQ database

Metric	H.264/ AVC	plr	MJPEG	Wavelet	White Noise	HEVC
MSE/PSNR	0.8340	0.7920	0.5089	0.7691	0.9035	0.4983
SSIM	0.9485	0.8654	0.8255	0.8607	0.9238	0.8748
MSSIM	0.9495	0.8566	0.9040	0.8561	0.9210	0.8939
MOVIE [23]	0.8970	**0.8860**	**0.8870**	0.9000	0.8430	0.9330
STRRED [29]	**0.9768**	0.8476	0.7290	**0.9459**	0.9305	0.8930
ViS_3 [35]	0.9200	0.8560	0.7890	0.9080	0.9280	0.9170
MD-SSIM [17]	**0.9483**	0.7601	0.8517	0.9001	0.9143	**0.9449**
EaTSS (Spatial)	0.9349	0.8535	**0.9367**	**0.9225**	**0.9310**	0.9235
EaTSS (Temporal)	0.7472	**0.8680**	0.6739	0.7568	0.9055	0.6043
EaTSS	0.8803	0.8646	0.8587	0.8690	**0.9336**	**0.9465**

Bolded values indicate the best two correlation values

Table 6 Pearson linear correlation for the CSIQ database

Metric	H.264/AVC	plr	MJPEG	Wavelet	White Noise	HEVC
MSE/PSNR	0.8502	0.6305	0.4557	0.7921	0.7931	0.8174
SSIM	0.9494	0.8476	0.8563	0.8820	0.9410	0.9502
MSSIM	0.9495	0.8179	0.9152	0.8880	0.9485	0.9602
MOVIE [23]	0.9040	**0.8820**	**0.8820**	0.8989	0.8550	0.9370
STRRED [29]	**0.9759**	0.8691	0.7517	**0.9530**	**0.9508**	0.9070
ViS_3 [35]	0.9180	0.8500	0.8000	0.9080	0.9160	0.9330
MD-SSIM [17]	0.9390	0.7323	0.8570	0.9188	0.9372	**0.9645**
EaTSS (Spatial)	**0.9525**	0.8673	**0.9417**	**0.9341**	0.9490	0.9403
EaTSS (Temperal)	0.7513	0.8776	0.6989	0.7926	0.9308	0.6924
EaTSS	0.9038	**0.8813**	0.8603	0.9015	**0.9656**	**0.9539**

Bolded values indicate the best two correlation values

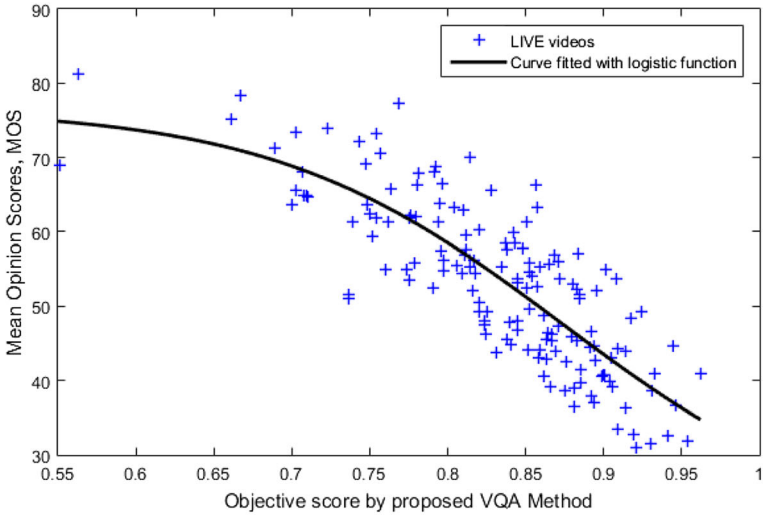
results of SC. EaTSS has better performance than MOVIE for MPEG-2 distorted videos. MOVIE's performance is rather insignificant vis-à-vis SC. On the contrary, C-B method performs better. Similar to EaTSS, it has good correlations in two types of distortions. Again in terms of PC, EaTSS has the competitive performance. For spatial and temporal parts of EaTSS, the overall results are similar to SC. The improvement of EaTSS for IP and MPEG-2 distorted videos are more significant. In terms of the LIVE video database, EaTSS performs the best. It is amongst the best two methods for both SC and PC. Comparing to MOVIE and C-B Method, they attain good results for either SC or PC only. This shows that EaTSS is more consistent and correlates better to HVS.

EaTSS also demonstrates good performances in the CSIQ database. We reclaim that VQM-VFD and C-B method are excluded in this comparison. This is due to the incompetence of hardware and unavailability of codes as well as results. The results in terms of SC and PC are shown in Tables 5 and 6 respectively. EaTSS outperforms all IQA metrics for all distortions except for H.264 compressed videos. Similar condition holds for other VQA metrics. MSSIM has the best SC and PC for H.264/AVC. This attributes to the variable block-size motion compensation (multiscale operation) for segmenting movement regions [9] in H.264/AVC. MSSIM also performs the best for MJPEG compressed videos. The underlying rationale is that MJPEG only involves intra-frame compressions [32]. Thus, IQA metrics perform better than VQA metrics for MJPEG compressed videos. When only the spatial part of EaTSS is tested, it achieves 0.9367 and 0.9417 for SC and PC respectively. The results are significantly better

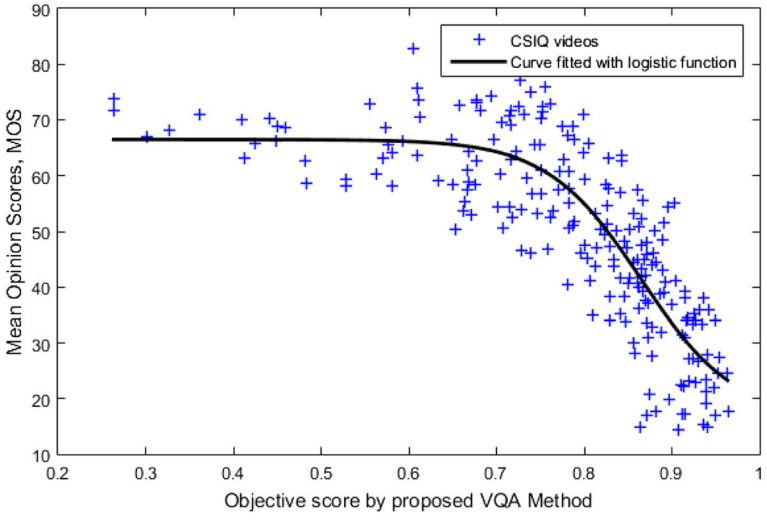
Table 7 Overall performance of VQA metrics

Metric	LIVE		CSIQ		Weighted	
	SC	PC	SC	PC	SC	PC
MOVIE [23]	0.7890	0.8116	0.8060	0.7880	0.7990	07976
ViS_3 [35]	**0.8160**	**0.8300**	**0.8410**	0.8300	**0.8148**	**0.8300**
MD-SSIM [17]	0.7787	0.7862	0.7934	0.8000	0.7874	0.7943
STRRED [29]	0.8007	0.8062	0.8129	0.7894	0.8079	0.7963
VQM-VFD [21]	0.7354	0.7763	–	–	–	–
C-B Method [5]	0.8061	**0.8278**	–	–	–	–
EaTSS (Spatial)	0.7608	0.7797	0.8243	**0.8410**	0.7983	0.8159
EaTSS (Temporal)	0.7542	0.7592	0.7353	0.7503	0.7430	0.7539
EaTSS	**0.8127**	0.8201	**0.8327**	**0.8426**	**0.8327**	**0.8417**

Bolded values indicate the best two correlation values



(a)



(b)

Fig. 3 Scatter plots: **a** EaTSS scores against subjective scores for the LIVE video database and **b** EaTSS scores against subjective scores for the CSIQ video database

than MSSIM and other IQA metrics. This proves that metrics without temporal consideration can perform better in MJPEG compressed videos. MSSIM also achieves the best PC for wavelet compressed videos. Since HEVC utilizes multiscale transform units in the inverse transform [31], MSSIM can perform better in HEVC compressed videos.

For VQA metrics excluding spatial and temporal parts of EaTSS, EaTSS achieves top two performances in three out of six distortion categories for both SC and PC. In overall, EaTSS only achieve lower correlations for two distortion types, i.e. H.264/AVC and wavelet. Yet, EaTSS still perform as good as MOVIE for these two distortions. Apparently, EaTSS (Spatial) performs better than the EaTSS (Temporal) and the EaTSS. The probable cause of this

Table 8 Kurtosis for the LIVE database

Metric	Wireless	IP	H.264	MPEG-2	Overall
MOVIE [23]	2.2984	2.4279	2.4481	2.1993	2.5630
ViS_3 [35]	3.2403	2.4673	2.2415	2.5675	2.4863
MD-SSIM [17]	2.6688	2.5333	3.3545	2.1559	2.6834
STRRED [29]	2.0530	2.6452	2.5865	2.5522	2.6692
VQM-VFD	2.7327	2.9508	3.0336	2.4012	2.8811
EaTSS (Spatial)	2.4965	2.5260	3.3166	2.0396	2.4117
EaTSS (Temperal)	2.0416	1.9005	2.7990	2.2219	2.3130
EaTSS	2.4260	2.4241	2.6088	3.4250	2.3971

condition is the imbalance of spatial and temporal distortions in videos of the CSIQ database. This is evidenced by results that in most distortion types, IQA metrics perform better than VQA metrics. This intimates that spatial distortions are dominant for most videos. Therefore, EaTSS (Spatial) performs better. To conclude, EaTSS has the best performance in the CSIQ database. It has the most top two rankings among all VQA metrics for all distortion types. Similar to the LIVE video database, the performance of EaTSS is more consistent than existing methods in the CSIQ video database. EaTSS achieves the best two results for the same distortion types for SC and PC. This is not the case for existing methods. Most of them perform well in either case only. The best performing existing method is STRRED which attains good results for two same distortion types for PC and SC.

Compare to existing VQA metrics, EaTSS has good performance for each distortion. The overall performances of all VQA metrics for each and all databases are tabulated in Table 7. For the LIVE video database with four distortion categories, ViS_3 exceeds EaTSS in terms of PC. At the same time, they have similar performances in terms of SC. Conversely, EaTSS defeats C-B method in SC but performs slightly poorer than C-B method in PC. ViS_3, C-B method, and EaTSS has very close PC. For EaTSS (Spatial) and EaTSS (Temporal), both perform equally with MD-SSIM. Yet, their combination gives a competitive result. Thus, spatial and temporal parts of EaTSS capture spatial and temporal distortions complementarily in the LIVE video database.

For CSIQ video database with six distortion types, EaTSS has the best performance in terms of PC. For SC, its result is only 0.01 lower than ViS_3. Consequently, EaTSS has similar performance to ViS_3 in the CSIQ database. The good performances of EaTSS in the CSIQ database are mainly due to the spatial part of EaTSS. This is evident as EaTSS (Spatial)

Table 9 Kurtosis for the CSIQ database

Metric	H.264/AVC	plr	MJPEG	Wavelet	White Noise	HEVC	Overall
MOVIE [23]	3.0823	3.1974	2.6003	2.6593	3.9499	3.7589	3.4389
STRRED [29]	3.1030	3.1191	3.3328	2.1521	2.1719	3.1265	3.2596
ViS_3 [35]	3.2993	2.6310	2.2721	2.3675	9.4670	2.4497	2.7634
MD-SSIM [17]	2.0690	3.1445	2.7870	1.9140	3.3767	5.2665	2.5939
EaTSS (Spatial)	2.9519	2.1405	2.6842	2.3891	2.2480	2.9312	3.1988
EaTSS (Temperal)	2.5485	2.2269	2.5818	4.5468	4.1976	3.3072	2.9485
EaTSS	2.2978	2.0418	2.5374	4.2717	2.1121	3.5842	2.6614

Table 10 F-Test (LIVE)

	M1	M2	M3	M4	M5	M6	M7	M8
M1	- - - - -	- 0 - 0 -	- 0 - 0 -	1 0 - 0 -	- 0 - 0 -	- 0 - 0 -	- 0 - 0 -	- 0 - 0 -
M2	- 1 - 1 -	- - - - -	- - - - -	1 - - - -	- - - - -	- - - - -	- - - - 1	- - - - -
M3	- 1 - 1 -	- - - - -	- - - - -	1 - - - -	- - - - -	- - - - -	- - - - -	- - - - -
M4	0 1 - 1 -	0 - - - -	0 - - - -	- - - - -	0 - - - -	0 - - - -	0 - - - -	0 - - - -
M5	- 1 - 1 -	- - - - -	- - - - -	1 - - - -	- - - - -	–	- - - 1 -	- - - - -
M6	- 1 - 1 -	- - - - -	- - - - -	1 - - - -	- - - - -	–	- - - - -	- - - - -
M7	- 1 - 1 -	- - - - 0	- - - - -	1 - - - -	- - - 0 -	–	- - - - -	- - - - -
M8	- 1 - 1 -	- - - - -	- - - - -	1 - - - -	- - - - -	–	- - - - -	- - - - -

performs much better than EaTSS (Temporal). However, there are still improvements when combing the two parts, although it is minor.

In Table 7, the weighted average of SC and PC, as in [39, 45], over the two databases are also shown. The weight for each database is computed according to the number of videos respectively. From the weighted average, EaTSS achieves the best performance for both SC and PC. This shows that EaTSS has better generalizations to distortion types and videos. ViS_3 achieves the second best result. Although EaTSS perform as well as ViS_3 with respect to each database, weighted average indicates that ViS_3 has lower generalizations to different distortion types and videos. EaTSS (Spatial) also has better generalities than EaTSS (Spatial). Still, their combination performs much better than their sole implementations. From the results in Tables 3, 4, 5, 6 and 7, EaTSS is among the best performing metrics as compared with existing metrics independent to databases.

4.4 Statistical evaluation

Statistical significance of the correlations of all VQA methods is also verified. This is done by using F-test as demonstrated by the works in [24, 34]. The Gausiannity results through the kurtosis based criterion stated in Section 4.2 are shown in Tables 8 and 9 for the LIVE and CSIQ databases respectively. For each database, there is only one category that fails the kurtosis based criterion. Thus, F-test is appropriate to be tested on these two databases. The F-test results are shown in Tables 10 and 11. Three symbols are used to indicate the result. Symbols "-", "1″, and "0″ indicates the statistical performance of VQA method placed in the row are indistinguishable, superior, and inferior to that of the method in the column respectively. In order to make the tables more compact, we use M1 to M8 to represent the VQA

Table 11 F-Test (CSIQ)

	M1	M2	M3	M4	M6	M7	M8
M1	- - - - -	- - 1 - 0 - -	- - - - -	0 - 1 - 0 - -	- - - - 0 0 -	1 - 1 1 - 1 -	- - - - 0 - 0
M2	- - 0 - 1 - -	- - - - -	- - - - 1 - -	0 - - - - - -	- - 0 - - 0 -	1 - - 1 - 1 1	- - - - -
M3	- - - - -	- - - - 0 - -	- - - - -	0 - - - 0 - -	- - 0 - 0 - 0	1 - - 1 - 1 -	- - - - 0 - 0
M4	1 - 0 - 1 - -	1 - - - - - -	1 - - - 1 - -	- - - - -	- 0 0 - - 0 -	1 - - 1 - 1 -	1 - - - - - -
M6	- - - - 1 1 -	- - 1 - - 1 -	- - 1 - 1 - 1	- 1 1 - - 1 -	- - - - -	1 - 1 1 - 1 1	1 - 1 - - - -
M7	0 - 0 0 - 0 -	0 - - 0 - 0 0	0 - - 0 - 0 -	0 - - 0 - 0 -	0 - 0 0 - 0 0	- - - - -	0 - 0 - 0 0 0
M8	- - - - 1 - 1	- - - - -	- - - - 1 - 1	0 - - - - - -	0 - 0 - - - -	1 - 1 - 1 1 1	- - - - -

Table 12 Computational complexity in terms of processing time

Metric	Processing Time (s)
MD-SSIM	9.53
EaTSS	17.79
STRRED	39.18
VQM-VFD	131.96
ViS_3	314.32
MOVIE	2,923,261.93

methods being compared. M1 to M8 indicates MOVIE, ViS_3, MDSSIM, STRRED, VQM-VFD, EaTSS (Spatial), EaTSS (Temporal), and EaTSS respectively.

The first four symbols for every entry in Table 10 refers to the F-test of wireless, IP, H.264, and MPEG-2 distortions respectively. The fifth symbol is the results for all distortion types. For all videos in the LIVE video database, all VQA methods perform equally with an exception that ViS_3 is statically superior to EaTSS (temporal). For wireless distorted videos, all VQA methods perform better than STRRED. Similar condition holds for MOVIE in IP and MPEG-2 distorted videos. For MPEG-2 compressed videos, VQM-VFD is statically superior to EaTSS (temporal). All of the VQA methods have identical performance for H.264 compressed videos.

For the CSIQ database, there are seven symbols for each entry. The first six symbols represent the distortions list in the first column of Table 2 sequentially. Meanwhile, the last symbol is the overall performance of all videos. To summarize the results, all VQA methods are superior to EaTSS (Temporal) for H.264/AVC, wavelet, and HEVC compressed videos. STRRED is superior to all methods except EaTSS (Spatial) for H.264/AVC compressed videos. For plr distortion, EaTSS (Spatial) outruns STRRED while the others perform equally. In terms of MJPEG compressed videos, EaTSS (Spatial) is superior to all methods except MOVIE. MOVIE further superiors to ViS_3, STRRED, and EaTSS (Temporal). All methods surpass MOVIE and MDSSIM except EaTSS (Temporal) for white noise impaired videos. For HEVC compressed videos, EaTSS (Spatial) outperforms MOVIE, ViS_3, and STRRED. In terms of all distortion types, ViS_3, EaTSS (Spatial), and EATSS are superior to EaTSS (Temporal). EATSS further defeat MOVIE and MDSSIM. EaTSS (Spatial) is also superior to MDSSIM.

In terms of F-test, ViS_3, MDSSIM, VQM-VFD, EaTSS (Spatial), and EaTSS are the best performing metrics for LIVE video database. On the other hand, EaTSS (Spatial) exceeds all VQA methods for CSIQ video database. This is followed by EaTSS. Thus, EaTSS is regarded as statically superior to other VQA methods excluding EaTSS (Spatial) and EaTSS (Temporal).

Table 13 Efficiency of EaTSS

EaTSS opreation	Number of addition/subtraction	Number of multiplication/division
W	$16n$	$10n$
$SSIM$	$5nm + 4n$	$5nm + 10n$
$EaTSS_{spatial}$	$2n - 2$	$2n + 1$
VTI_D and VTI_R	$6n$	0
$SSIM\_no\_lc$	$3nm$	$3nm + 5n$
w	$16n$	$10n$
$EaTSS_{temporal}$	$2n - 2$	$2n + 1$
$EaTSS$	1	1
Total	$8nm + 46n + 1$	$8nm + 39n + 3$

Table 14 CL and PC of VQA methods

Metric	Computational level	PLCC
MD-SSIM	5.10	0.7862
EaTSS	9.51	0.8201
STRRED	20.95	0.8062
VQM-VFD	70.57	0.7763
ViS_3	168.09	0.8300
MOVIE	1,563,241.67	0.8116

4.5 Computational complexity

Other than correlations, computational complexity of VQA metrics is also measured. The computational cost of each metric is measured in terms of the processing time. Table 12 shows the processing time for VQA metrics excluding C-B method. The test is done for a video (bs2_25fps.yuv) from the LIVE video database. The video is assessed by different VQA algorithms repeatedly for ten times, and the average processing time is calculated. MD-SSIM has the shortest processing time of 9.53 s. EaTSS only requires 17.79 s. The time is much less than the processing times of ViS_3 and MOVIE which are 314.32 s and 2,923,261.93 s respectively. Thus, EaTSS has low processing time and a lower computational cost than existing VQA methods other than the previous work by the authors.

The complexity of EaTSS is also analyzed theoretically. The total number of additions or subtractions and multiplications or divisions of each step in EaTSS is shown in Table 13. Parameter n refers to the total number of pixels in a frame and m is the total number of pixels in the patches used for *SSIM* and *SSIM_no_lc* in (7) and (14) respectively. In total, there are $8nm + 46n + 1$ additions and subtractions as well as $8nm + 39n + 3$ multiplications and divisions. So, EaTSS is a $O(nm)$ operation. If it is expressed in terms of frame height, M, frame width, N, and patch size, $p \times p$, then EaTSS is a $O(MNp^2)$ operation. This shows that EaTSS has low computational complexity and high efficiency.

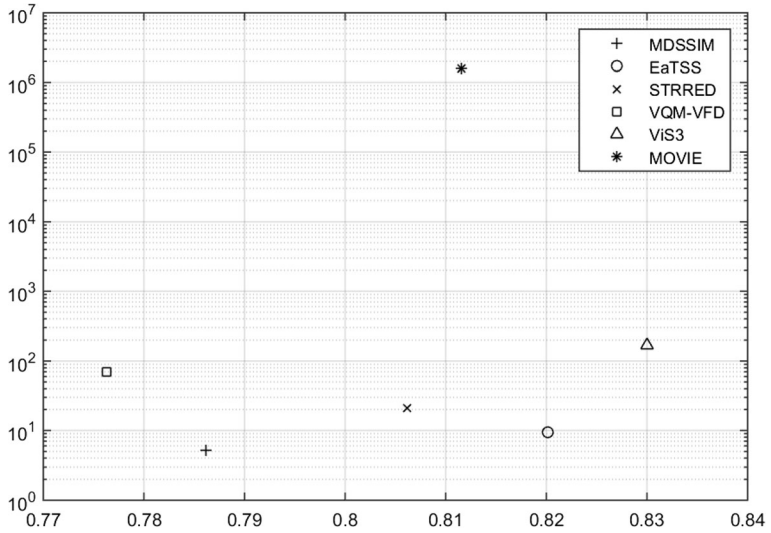


Fig. 4 CL versus PC plot

Other than computational time and theoretical analysis, the complexity of VQA methods is also compared in terms of efficiency. We follow procedures in [34] for this test. Computational Level (CL) is first computed for each method. CL is defined as the ratio of the computational time of VQA methods to the computational time of PSNR. Then, the graph of CL and PC are plotted. The results of CL and PC of every VQA methods are shown in Table 14. The graph of CL versus PC is illustrated in Fig. 4. For VQA methods with good efficiency, their points should be located as near as possible to the lower right of the plot. Meanwhile, the points of methods with poor efficiencies are located near to the upper left of the plot. From the figure, EaTSS and ViS_3 have very similar efficiency. Since the y-axis is in log scale, EaTSS has better efficiency as compared to ViS_3. MOVIE, on the other hand, has the worst efficiency due to the long computational time. In overall, EaTSS has the best efficiency among all VQA methods being tested.

5 Conclusion

A VQA method, EaTSS, is proposed in this paper based on error signals, locally weighted SSIM, and second-order time-differential information. The experiment results show that it performs very well in both benchmark databases. For the LIVE video database, it has similar performance to ViS_3 and the C-B method. It outperforms most of the recently proposed VQA metrics, i.e. STRRED, VQM-VFD, and MOVIE. EaTSS also has very good performance in the CSIQ video database where it achieves competitive performance with ViS_3. In overall, EaTSS assess the distortions in these two databases well. Weighted PC and SC show that EaTSS has good generalization to videos suffered from different types of distortions that are database independent. Furthermore, EaTSS has low computational time and cost. Besides that, it has the highest efficiency compared with existing VQA metrics.

Acknowledgments This work was supported by Ministry of Higher Education Malaysia through the provision of research grant: F02/FRGS/1492/2016.

References

1. Akramullah S (2014) Digital video concepts, methods, and metrics. New York, USA
2. Banno H, Saiki J (2015) The use of higher-order statistics in rapid object categorization in natural scenes. J Vis 15(2):1–20. https://doi.org/10.1167/15.2.4
3. Bong DBL, Khoo BE (2015) Objective blur assessment based on contraction errors of local contrast maps. Multimed Tools Appl 74(17):7355–7378. https://doi.org/10.1007/s11042-014-1983-5
4. Cardoso JVM, Alencar MS, Regis CDM, Oliveira IP (2014) Temporal analysis and perceptual weighting for objective video quality measurement. IEEE southwest symposium on image analysis and interpretation (SSIAI). IEEE, 2014, pp 57–60. https://doi.org/10.1109/SSIAI.2014.6806028
5. Choi LK, Bovik AC (2016) Flicker sensitive motion tuned video quality assessment. 2016 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI). IEEE, 2016, pp 29–32. https://doi.org/10.1109/SSIAI.2016.7459167
6. Hall CF, Hall EL (1977) A nonlinear model for the spatial characteristics of the human visual system. IEEE Trans Syst Man Cybern 7(3):274–283. https://doi.org/10.1109/TSMC.1977.4309680
7. Ítalio PO, José VMC, Carlos DMR, Marcelo SA (2013) Spatial and temporal analysis considering relevant regions applied to video quality assessment. XXXI Brazilian Telecommunications Symposium (SBrT), 2013, pp 1–5

8. ITU-T (1999) ITU-T recommendation P. 910: subjective video quality assessment methods for multimedia applications. ITU-T
9. ITU-T (2016) ITU-T recommendation H.264: advanced video coding for generic audiovisual services. ITU-T
10. Larson EC, Chandler DM (2010) Most apparent distortion: full-reference image quality assessment and the role of strategy. J Electron Imaging 19(1):011006–011006. https://doi.org/10.1117/1.3267105
11. Lee S, Pattichis MS, Bovik AC (2002) Foveated video quality assessment. IEEE Trans Multimedia 4(1): 129–132. https://doi.org/10.1109/6046.985561
12. Li C, Bovik AC (2009) Three-component weighted structural similarity index. Proceding SPIE 7242, image quality and system performance VI, 72420Q. SPIE, 2009, pp 1–9
13. Li P, Xie J, Wang Q, Zuo W (2017) Is second-order information helpful for large-scale visual recognition?. IEEE international conference on computer vision (ICCV). IEEE, 2017, pp 2070–2078
14. Limb JO (1979) Distortion criteria of the human viewer. IEEE Trans Syst Man Cybern 9(12):778–793. https://doi.org/10.1109/TSMC.1979.4310129
15. Liu Y, Nie L, Han L, Zhang L, Rosenblum DS (2015) Action2Activity: recognizing complex activities from sensor data. Proceedings of the 24th international conference on artificial intelligence. IJCAI, 2015, pp 1617–1623
16. Liu L, Cheng L, Liu Y, Jia Y, Rosenblum DS (2016) Recognizing complex activities by a probabilistic interval-based model. Proceedings of the thirtieth AAAI conference on artificial intelligence. AAAI, 2016, pp 1266–1272
17. Loh WT, Bong DBL (2015) Video quality assessment method: MD-SSIM. IEEE international conference on consumer electronics - Taiwan (ICCE-TW). IEEE, 2015, pp 290–291. https://doi.org/10.1109/ICCE-TW.2015.7216904
18. Ooyala (2016) Ooyala global video index q3 2015. Ooyala. http://go.ooyala.com/rs/OOYALA/images/Ooyala-Global-Video-Index-Q3-2015.pdf. Accessed 14 Apr 2016
19. Peng P, Cannons K, Li ZN (2013) Efficient video quality assessment based on spacetime texture representation. In: Proceedings of the 21st ACM international conference on multimedia. ACM, 2013, pp 641–644. https://doi.org/10.1145/2502081.2502168
20. Pinson MH, Wolf S (2004) A new standardized method for objectively measuring video quality. IEEE Trans Broadcast 50(3):312–322. https://doi.org/10.1109/TBC.2004.834028
21. Pinson MH, Choi LK, Bovik AC (2014) Temporal video quality model accounting for variable frame delay distortions. IEEE Trans Broadcast 60(4):637–649. https://doi.org/10.1109/TBC.2014.2365260
22. Rimac-Drlje S, Vranješ M, Žagar D (2010) Foveated mean squared error - a novel video quality metric. Multimed Tools Appl 49(3):425–445. https://doi.org/10.1007/s11042-009-0442-1
23. Seshadrinathan K, Bovik AC (2010) Motion tuned spatio-temporal quality assessment of natural videos. IEEE Trans Image Process 19(2):335–350. https://doi.org/10.1109/TIP.2009.2034992
24. Seshadrinathan K, Soundararajan R, Bovik AC, Cormack LK (2010) Study of subjective and objective quality assessment of video. IEEE Trans Image Process 19(6):1427–1441. https://doi.org/10.1109/TIP.2010.2042111
25. Seshadrinathan K, Soundararajan R, Bovik AC, Cormack LK (2010) A subjective study to evaluate video quality assessment algorithms. IS&T/SPIE Electronic Imaging. International Society for Optics and Photonics, 2010, pp 75270H-75270H-10. https://doi.org/10.1117/12.845382
26. Sheikh HR, Bovik AC (2005) A visual information fidelity approach to video quality assessment. First international workshop on video processing and quality metrics for consumer electronics. Springer, 2005, pp 23–25
27. Sheikh HR, Bovik AC, de Veciana G (2005) An information fidelity criterion for image quality assessment using natural scene statistics. IEEE Trans Image Process 14(12):2117–2128. https://doi.org/10.1109/TIP.2005.859389
28. Sheikh HR, Sabir MF, Bovik AC (2006) A statistical evaluation of recent full reference image quality assessment algorithms. IEEE Trans Image Process 15(11):3440–3451. https://doi.org/10.1109/TIP.2006.881959
29. Soundararajan R, Bovik AC (2013) Video quality assessment by reduced reference Spatio-temporal entropic differencing. IEEE Trans Circuits Syst Video Technol 23(4):684–694. https://doi.org/10.1109/TCSVT.2012.2214933
30. Suchow JW, Alvarez GA (2011) Motion silences awareness of visual change. Curr Biol 21(2):140–143. https://doi.org/10.1016/j.cub.2010.12.019
31. Sullivan GJ, Ohm JR, Han WJ, Wiegand T (2012) Overview of the high efficiency video coding (HEVC) standard. IEEE Trans Circuits Syst Video Technol 22(12):1649–1668. https://doi.org/10.1109/TCSVT.2012.2221191

32. Vo DT, Nguyen TQ (2008) Quality enhancement for motion JPEG using temporal redundancies. IEEE Trans Circuits Syst Video Technol 18(5):609–619. https://doi.org/10.1109/TCSVT.2008.918807
33. VQEG (2003) Final report from the video quality experts group on the validation of objective models of video quality assessment. VQEG
34. Vranješ M, Rimac-Drlje S, Grgić K (2013) Review of objective video quality metrics and performance comparison using different databases. Signal Process Image Commun 28(1):1–19. https://doi.org/10.1016/j.image.2012.10.003
35. Vu PV, Chandler DM (2014) ViS_3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices. J Electron Imaging 23(1):013016–013016. https://doi.org/10.1117/1.JEI.23.1.013016
36. Vu C, Deshpande S (2012) ViMSSIM: from image to video quality assessment. Proc of the 4th workshop on mobile video. ACM, 2012, pp 1–6. https://doi.org/10.1145/2151677.2151679
37. Wandell BA (1995) Foundations of vision. Sinauer Associates, Sunderland
38. Wang Z, Bovik AC (2009) Mean squared error: love it or leave it? A new look at signal fidelity measures. IEEE Signal Process Mag 26(1):98–117. https://doi.org/10.1109/MSP.2008.930649
39. Wang Z, Li Q (2011) Information content weighting for perceptual image quality assessment. IEEE Trans Image Process 20(5):1185–1198. https://doi.org/10.1109/TIP.2010.2092435
40. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 3(4):600–612. https://doi.org/10.1109/TIP.2003.819861
41. Wang Z, Lu L, Bovik AC (2004) Video quality assessment based on structural distortion measurement. Signal Process Image Commun 19(2):121–132. https://doi.org/10.1016/S0923-5965(03)00076-6
42. Wang Z, Simoncelli EP, Bovik AC (2003) Multiscale structural similarity for image quality assessment. Conference record of the thirty-seventh Asilomar Conf on signals, systems and computers. IEEE, 2003, pp 1398–1402. https://doi.org/10.1109/ACSSC.2003.1292216
43. Watson AB, Hu J, McGowan JF (2001) Digital video quality metric based on human vision. J Electron Imaging 10(1):20–29. https://doi.org/10.1117/1.1329896
44. Wolf S, Pinson M (1998) In-service performance metrics for MPEG-2 video systems. Proc made to measure 98-measurement techniques of the digital age technical seminar. IAB, 1998, pp 12–13
45. Xue W, Mou X, Zhang L, Bovik AC, Feng X (2014) Blind image quality assessment using joint statistics of gradient magnitude and Laplacian features. IEEE Trans Image Process 23(11):4850–4862

Woei-Tan Loh received his B.Eng. degree in Electronics and Telecommunication Engineering from Universiti Malaysia Sarawak in 2014. He had worked as a research assistant from the year 2014 until 2016. He is currently pursuing a postgraduate degree in Universiti Malaysia Sarawak. His research interests include video quality assessment and computer vision.

David B.L. Bong received his PhD from Universiti Sains Malaysia, Master of Science degree from Nanyang Technological University, Singapore, and Bachelor of Electrical Engineering degree from Universiti Teknologi Malaysia. He is currently a Senior Lecturer with Faculty of Engineering, Universiti Malaysia Sarawak. His research interests include image quality assessment, feature extraction, computer vision, medical imaging and intelligent systems.