Multimedia Tools and Applications
https://doi.org/10.1007/s11042-022-13416-8

Improving precision of objective image/video quality meters

Majid Behzadpour [1] · Mohammad Ghanbari [1,2]

Received: 9 December 2020 / Revised: 3 April 2022 / Accepted: 2 July 2022

Abstract
Although subjective test is the most accurate image/video quality assessment tool, it is extremely time demanding. In the past two decades, a variety of objective quality measuring tools, such as SSIM, IW-SSIM, SPSIM, FSIM, etc., have been devised, that well correlate with the subjective tests results. However, the main problem with these methods is that they do not discriminate the measured quality well enough, especially at high quality range. In this article we show how the accuracy/precision of these Image Quality Assessment (IQA) meters can be increased by mapping them into a Logistic Function (LF). The precisions are tested over a variety of image/video databases. Our experimental tests indicate while the used high-quality images can be discriminated by 23% resolution on the MOS subjective scores, discrimination resolution by the widely used IQAs are only 2%, but their mapped IQAs to Logistic Function at this quality range can be improved to $9-17\%$, depending on the characteristics of the LF function. Moreover, their precision at low to mid quality range can also be improved. At this quality range, while the discrimination resolution of MOS of the tested images is 23.2%, those of raw IQAs is nearly 8.9%, but discrimination of their adapted logistic functions can be very close to that of MOS. Moreover, with the used image databases the Pearson Linear Correlation Coefficient (PLCC) of MOS with the logistic function can be improved by $2-20\%$ as well.

Keywords Image quality assessment · Objective quality meters · Structural distortion · Structural similarity index

✉ Mohammad Ghanbari
ghan@ut.ac.ir; ghan@essex.ac.uk

Majid Behzadpour
majid.behzadpour@ut.ac.ir

[1] Department of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran

[2] School of Computer Science and Electronic Engineering, University of Essex, CO4 3SQ Colchester, UK

Published online: 28 July 2022

1 Introduction

In the past two decades there have been a lot of interests in both image and video processing. This is mainly due to the explosive growth of multimedia over the internet. Currently Cisco predicts by year 2022, more than 82% of the internet traffic will be video related material [3], let alone applications in social networks, which try to retrieve mages of various kinds in the net. Considering that raw image/video demand a large volume of data to be represented properly, their compression to achieve a manageable storage and transmission rate is inevitable. This is only possible at the cost of induced distortions in the processed image/video. It is highly desired to measure such distortions, by any objective measuring tool.

Considering that the ultimate receptor of visual content is the human visual system (HVS), the best and most accurate measuring device for assessing processed image/video distortions is again based on HVS. This is normally carried out by subjective tests, where a group of viewers watch a set of distorted image/video contents, and viewers' mean opinion score (MOS) is taken as the best representative of visual quality. However, this process apart from being time consuming, it requires certain laboratory set ups, which may not be feasible for all users.

To resolve MOS limitations, historically image/video quality is measured based on the difference between their unprocessed and processed versions and presented in terms of Peak-Signal-to-Noise Ratio, PSNR. However, it can be argued that PSNR may not be a valid quality measure in certain scenarios. For instance, if the original non-distorted image is shifted even by one pixel, the difference between the original signal and its shifted version can show a significant drop in PSNR, whereas the shifted image quality is subjectively perfect. Moreover, PSNR value is not an indication of absolute acceptable video quality, nor it can be used to compare two different visual contents. Despite this, PSNR is a valid criterion in comparing image/video of the same content, provided their dimensions are not altered. In [5], it is shown that if the image content remains unaltered, improving PSNR can definitely improve MOS. This is the reason all video codecs, through rate-distortion optimization try to minimize coding distortion (maximize PSNR) for the best subjective quality.

Over the past two decades a number of image quality assessment (IQA) tools have been devised that can alleviate PSNR limitations. One group of these IQAs are based on No-Reference (NR) concept that can measure the image quality without referring it to the non-degraded reference image. This can isolate the influence of the reference picture on the accuracy of measuring tool. For instance, we had shown that by extracting the quantizer parameter and the number of DCT blocks which have only one non-zero coefficient from the bitstream, one can gauge the video quality [4]. More sophisticated NR model can be made by measuring coding distortions, such as blurriness [1], or a mixture of blurriness and blockiness [16]. Mittal et al. have devised a NR meter, called BRISQUE which does not need to measure blurry or blocking artifacts, but instead uses scene statistics of locally normalized luminance coefficients to quantify possible losses of "naturalness" in the image due to the presence of distortions [9]. Despite the fact that normally NR has inferior accuracy compared to Full-Reference (FR), they claim this meter is even more accurate than the FR of PSNR and SSIM, without having their limitations on picture size alterations or orientations [9].

Another group of IQAs known as perceptual meters are based on Structural Similarity Index (SSIM), which like PSNR they are based on FR. A variety of these perceptual meters have been developed but all have a common problem that they lose precision and accuracy at high image quality range. The main contribution of this paper is to show how Logistic Functions (LF) can improve the performance of these quality metrics. Through the

experiments we show how LF can be easily added to all of these measuring tools, not only to improve their precision but also to increase their correlations to MOS.

The rest of the paper is organized in the following order. Section 2 looks at some of the most common IQA measuring tools and their common limitations. Section 3 introduces the proposed Logistic Function (LF), and through experiments show LF can increase the Pearson Linear Correlation Coefficient (PLCC) of all IQAs with the MOS. Section 4 extends the proposed method to enhance the widely used Video Multimethod Assessment Fusion (VMAF) of measuring video quality. Finally, Section 5 draws some concluding remarks.

2 Popular Image Quality Assessment (IQA) tools

PSNR is a kind of full reference model, but it is sensitive to picture content and cannot evaluate relative quality of two different contents. A family of full reference meters that do not have such sensitivity, are based on structural similarity, the so-called structural similarity index [23]. In this method as long as any added distortion does not alter the structure of the neighboring pixels, the human visual system is not sensitive to it. In the past two decades, numerous methods based on structural similarity have been devised.

For instance, in multi scale structural similarity [22] it is assumed that the human visual system adapts itself to extract structural information of the scene, and hence structural similarity can provide a good measure of perceived image quality. Weighting structural similarity for better adaptation is presented in IWSSIM [20]. In [27] Local weight is calculated based on the symmetry model of the reference image and more weights are given to certain areas. Using such criterion, [26] introduces VSI, a visual saliency-induced index for perceptual image quality assessment, where more weight is given in the pooling strategy. FSIM: a Feature Similarity Index for image quality assessment is described in [25]. Since human visual system is more sensitive to image edges, FSIM is mainly an edge-sensitive image quality assessor. The super-pixel method, known as SPSSIM, is another well-known and new model that divides images into meaningful areas and the evaluation model is based on the local quality of these areas [14]. Finally, an image quality assessment method based on edge-feature image segmentation (EFS) is proposed in [13].

Although these variants of Image Quality Assessment (IQA) methods have some gains or deficiencies over each other, they all suffer from a common deficiency that, at high image quality range their scores tend to saturate. This makes their measured values at high image quality to lose accuracy and make them almost unreliable quality meters. Figure 1 shows relationship between MOS and the objective quality scores by some of these methods for a TID2013 [11] image database. They include: SSIM [23], MS-SSIM [22], IW-SSIM [20], VSI [26], FSIM [25], FSIMc [25], SPSIM [14], EFS [13] and GMSD [24]. As seen at high image quality, all the measured values are very close to each other and lose precision. At lower image quality range, although some behave better than SSIM, but still image scores at this range are scattered. This paper aims to alleviate these shortfalls and hopefully to improve their correlations with the MOS.

Before explaining how the precision of measurement and its correlation with MOS can be improved, let us briefly explain how each of these measures, define structure in their definitions:

(a) SSIM: The structural similarity index measure (SSIM). which is used for measuring the structural similarity between two blocks of pixels.

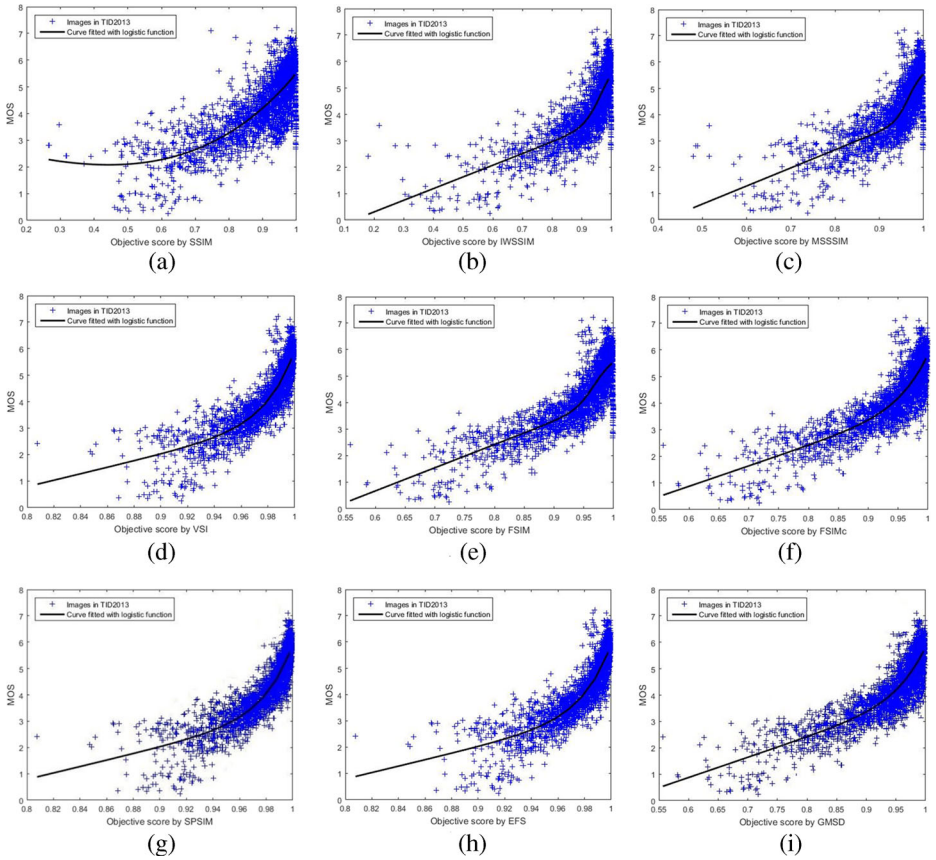

Fig. 1 Scatter plots of subjective MOS against scores obtained by model prediction on the TID2013 database. **a** SSIM, **b** IW-SSIM, **c** MS-SSIM, **d** VSI, **e** FSIM, **f** FSIMc, **g** SPSIM, **h** EFS, **i** GMSD

(b) IW-SSIM: Information Content Weighted Structural Similarity Index for IQA, which gives extra weight to the content during pooling.

(c) MS-SSIM: Calculates the multi-scale structural similarity (MS-SSIM). This function calculates the SSIM index of several versions of the image at various scales.

(d) VSI: A Visual Saliency-Induced Index for IQA. Visual saliency (VS) puts emphasis on areas of an image which will attract the most attention of the human visual system.

(e) FSIM: A Feature Similarity Index for IQA. It is based on the fact that human visual system (HVS) understands an image mainly according to its low-level features. Specifically, the phase congruency (PC), which is a dimensionless measure of the significance of a local structure.

(f) FSIMc: is a FSIM which also uses color components information in its calculation.

(g) SPSIM: A Super-pixel-Based Similarity Index for IQA. It is based on the fact that a super-pixel is a set of image pixels that share similar visual characteristics and is thus perceptually meaningful.

(h) EFS: An image quality assessment method by edge-feature-based image segmentation (EFS).

(i) GMSD: Gradient Magnitude Similarity Deviation for IQA. The reason for using such a measure in structural similarity is that: image gradients are sensitive to image distortions, while different local structures in a distorted image suffer different degrees of degradations.

On the significance of the above measures, it is worth noting that Wang et al. in a highly cited article [21] have answered the question of "why image quality assessment is so difficult", and they have concluded that a correct way in measurement is to model the image degradation as a structural distortion instead of error (which is used in PSNR). This is a good indication of why structurally based distortion measure is so popular.

We would like to emphasize that the above notes should not give the wrong impression that PSNR is a useless measuring tool, and all the credits should go to SSIM family. The fact is that usage of PSNR or SSIM family depends on the type of application. As we have pointed out in [5], PSNR is a valid measuring criterion if the image content remains the same, but it cannot be used to measure subjective quality of two different images. For instance, Fig. 2 shows the subjective quality of two different pictures with the same PSNR value of 25.1 dB. As seen, their subjective quality is very different and their IQA values measured by simple SSIM is more accurate. However, in a recent survey paper [6], we examined the suitability of 13 different SSIM-based IQA methods, as well as the PSNR in measuring the quality of error-concealed video clips. The aim was to find out which of these methods best measures the quality improvement of error-concealed packetized video clips. In these tests, the quality of error-concealed video frames alone as well as the whole video clips were evaluated by these 13 well-known IQA meters. Interesting conclusion was that none of these meters could indicate for sure, it was the best measuring tool, but among all the tests, PSNR was either first or second best. We believe the reason for success of PSNR is that, in loss concealment, the content remains the same and any improvement in PSNR under a loss concealing method, according to [5], should also improve the subjective quality.

However, all the SSIM family methods have the common weakness that, their discrimination of quality, particularly at high image quality range is poor and. We hope, by alleviating this deficiency, structural similarity based image quality meters can even be more widely used.

Fig. 2 Subjective quality of two different contents with the same PSNR

3 Proposed Logistic Function (LF)

Although as Fig. 1 shows, the variants of SSIM family improve the shortfall of SSIM at mid to low image quality range, but at high image quality, all of them similar to SSIM suffer from precision and accuracy. In [2] problem of SSIM has been mathematically studied and some improvements on SSIM-IQA has been reported. On the validation of subjective models of video quality assessment, VQEG has introduced logistic functions to approximate each objective parameter to a subjective impairment level. The functional form is a 3rd order polynomial with four- and five- parameter logistic curves which are optimized for the best performance [18, 19].

In this paper we mainly look at on the loss of precision of these metrics at high image quality range. We aim to show how a simple Logistic Function (LF) without any parameter optimization can be defined to improve the shortcoming of SSIM-family meters and in particular it is extended to video quality measurement.

According to Fig. 1, at high values of IQA, MOS grows exponentially with IQA. If it is dampened by any means, then relationship between them becomes closer to a linear function. For example, if we a define Logistic Function (LF), as given in Eq. (1):

$$\mathrm{LF} = 1 - \sqrt{1 - \mathrm{IQA}} \tag{1}$$

since IQA is a number between 0 and 1, for larger values of IQA, 1- IQA becomes much smaller. Taking square root of smaller values, leads to a larger difference with themselves. That is; the smaller is the value, the larger becomes its square root, and hence separating these points wider apart from each other. On the other hand, at lower values of IQA, 1-IQA gets larger, and its difference with its square root does not increase that much. Thus, such definition of LF makes larger values of IQA to be separated from each other more than their smaller values. This means that, LF functions like Eq. (1), can separate larger values of IQA (higher quality) more than separation between lower quality values. However, SSIM measured values at low values are sufficiently separated from each other, and they do not need further separation.

It is worth mentioning that the logistic function (LF) of Eq. (1) can be incorporated within each measuring device, to directly calculate LF value, rather than the IQA value. Please note that in quality assessment tools, normally for each block of pixels, or segment of images an IQA is calculated and the aggregate of IQAs of blocks/segments represent the final IQA score of an image. Alternatively, one may just use the overall output of IQA and map it to LF value as the final score. However, since Eq. (1) is not a linear function of IQA, the two methods are not exactly equivalent, and the former has a better accuracy over the latter. However, for simplicity and being more conservative, we have taken the worst case of mapping the final IQA to LF value. This is executed throughout all experiments and had we used per block/segment LF values, and their sum were amalgamated into a final LF, the results would have been better.

To see how the defined Logistic Function (LF) can improve the saturated high image quality, the IQA values measured by the used SSIM [23], MS-SSIM [22], IW-SSIM [20], VSI [26], FSIM [25], FSIMc [25], SPSIM [14], EFS [13] and GMSD [24] methods are mapped to LF and shown in Fig. 3. In the graphs of Fig. 3, after measuring the IQA by each measuring device, the derived IQA is mapped to its equivalent LF using Eq. (1), and hence the MOS is drawn versus their LF values.

As seen in Fig. 3, this time MOS has a better linear relation to the LF versions of the used IQA methods. Moreover, the resultant quality measure can have a higher correlation to the MOS too. Table 1 shows the Pearson Linear Correlation Coefficient (PLCC) of 9 structural similarity based IQAs for 4 sets of image databases of: CSIQ [7], LIVE [12], tid2008 [10] and tid2013 [11]. In this Table correlation between the MOS and measured IQA value of each method with and without their Logistic Functions are tabulated. The Table shows that for every measuring device, its LF version has much better correlation to MOS than IQA of that measure itself.

Please note that at the very high values of quality of Fig. 1, there are drops in quality and these are also present in their LF-versions in Fig. 3. This is because some of the images are in color and are subjectively evaluated with color fidelity, but in the objective measure, only luminance components are calculated (or vice versa, if they are in black and white, color components are also included in the objective measure). For instance, by comparing the scatter diagram of FSIM (without color) with that of FSIMc (with color), where the objective measure also includes color components, the difference in dropping the quality at the high end can be verified.

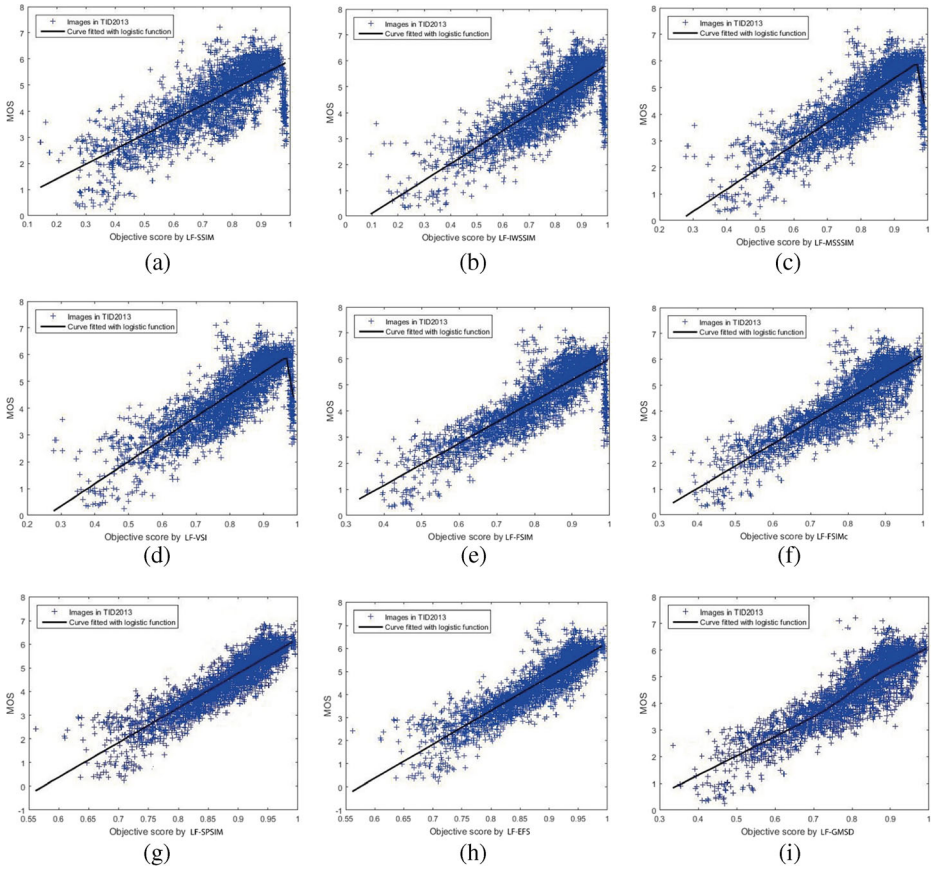

Fig. 3 Scatter plots of subjective MOS against scores obtained by model prediction with Logistic Function (LF) on the TID2013 database. **a LF**-SSIM, **b LF**-IW-SSIM, **c LF**-MS-SSIM, **d LF**-VSI, **e LF**-FSIM, **f LF**-FSIMc, **g LF**-SPSIM, **h LF**-EFS, **i LF**-GMSD

Table 1 Comparison between PLCC of MOS of various IQAs without and with their Logistic Functions (LF) for 4 image datasets

	CSIQ		LIVE		tid2008		tid2013	
	IQA	LF	IQA	LF	IQA	LF	IQA	LF
Ssim	0.7916	**0.8526**	0.829	**0.922**	0.7401	**0.7732**	0.7596	**0.7746**
iw-ssim	0.7947	**0.8891**	0.8029	**0.9207**	0.8086	**0.8555**	0.7638	**0.8172**
FSIM	0.8048	**0.893**	0.8586	**0.9407**	0.8301	**0.8724**	0.8195	**0.8408**
FSIMc	0.8208	**0.9032**	0.8595	**0.9472**	0.8341	**0.8747**	0.8322	**0.8749**
Ms-ssim	0.772	**0.87**	0.767	**0.9035**	0.7897	**0.8406**	0.7773	**0.8145**
VSI	0.8392	**0.9154**	0.7647	**0.9035**	0.8107	**0.8681**	0.8373	**0.8957**
SPSIM	0.8583	**0.9344**	0.7985	**0.9599**	0.8312	**0.8929**	0.8468	**0.9092**
GMSD	0.8812	**0.9542**	0.8629	**0.9602**	0.812	**0.8717**	0.8031	**0.8542**
EFS	0.8412	**0.9215**	0.8591	**0.9445**	0.8276	**0.8801**	0.8431	**0.9011**

In every dataset the quality of each measuring tool in IQA is represented normally, but its LF adapted version is bolded.

Another important point to note is that, if among the SSIM family, a method performs better than the other, its LF mapped version will also perform better. The reason is that, according to Eq. (1), each LF version of measured IQA is directly related to its IQA value. For instance, in Table 1, among the 9 tested IQAs, if GMSD is the best measuring device for CSIQ image database, its LF version also has the highest performance among all LF versions of this sequence. One may inspect all image databases of Table 1, for such property. This implies that, the IQA value of any measuring device can be mapped to LF, to improve its precision without damaging its correlation accuracy to MOS. The more significant point is that, since with the defined LF, higher quality points are better separated from each other, by bringing the measured quality values to their correct positions, the correlation between LF adapted SSIM with MOS increases. For instance, in all images and databases listed in Table 1, the Pearson Linear Correlation Coefficient (PLCC) between MOS and LF can be 2-20.2% better than such measure between MOS and measured IQA itself.

Apart from higher PLCC of LF measured meters, they have a better precision, not only at high quality, but also at medium and low quality as well.

To investigate the precision of the Logistic Function across all image quality ranges, as well as those in the structurally based similarity measures, we have borrowed the idea of image quality analysis method in large databases from Ponomarenko et al. [11]. In image analysis defined in [11], the MOSs of about 3000 images in the tid2013 image database is classified into three groups based on their quality range, each one of nearly 1000 images. First of all, the whole images are rated into 8 ranges (0–8). The first group called “bad quality” have an MOS in the range of 0.242–3.94. The second class called “middle quality” group contains images with MOS in the range of 3.94–5.25. Finally, the third group contains “good quality” images with MOS higher than 5.25.

Since our goal is to measure precision, we group the images in a known value of fixed qualities of 2, 4 and 6 for bad, middle and good quality respectively. It is interesting to note that, we do not need to take 1000 images, instead we have taken only 10 images from each group. We will show that even such a small sample, can prove our concept of precision measure.

Table 2 shows SSIM values and their LF versions along with MOS for 10 images, selected from the tid2013 database at almost high MOS score of 6 (good quality). Similarly, Tables 3 and 4 show these values for 10 images of middle and bad quality, respectively in the MOS scores of 4 and 2. The Tables also include the averages of MOS for the three

Table 2 Values of each IQA metric and their LF-SSIM along with MOS for good quality images from database tid2013 [11]

Image	MOS	SSIM	LF-SSIM	LF2-SSIM	LF3-SSIM
'i01_02_2.bmp'	6.1081	0.9948	**0.9282**	**0.8986**	**0.7826**
'i03_08_1.bmp'	6.3421	0.9970	**0.9456**	**0.9231**	**0.8192**
'i04_02_1.bmp'	6.2750	0.9927	**0.9148**	**0.8797**	**0.7563**
'i05_16_1.bmp'	6.1500	0.9939	**0.9219**	**0.8898**	**0.7701**
'i07_09_1.bmp'	6.4222	0.9904	**0.9022**	**0.8620**	**0.7329**
'i23_16_1.bmp'	6.3333	0.9980	**0.9552**	**0.9367**	**0.8411**
'i03_16_1.bmp'	6.8205	0.9969	**0.9443**	**0.9213**	**0.8164**
'i12_16_1.bmp'	6.5263	0.9988	**0.9656**	**0.9514**	**0.8668**
'i24_02_1.bmp'	6.4444	0.9962	**0.9386**	**0.9132**	**0.8039**
'i08_04_1.bmp'	6.0645	0.9992	**0.9708**	**0.9588**	**0.8807**
Avg	6.3487	0.9958	**0.9387**	**0.9135**	**0.8070**

LF-adapted versions of IQA values are shown in bold.

different quality ranges, as well as averages of their SSIM and LF versions of SSIM (LF-SSIM). Inspection of these data reveal the following interesting outcomes:

1. The difference between the average MOS of good quality video from the average of middle quality is 6.3487–4.5120 = 1.8367. Considering that the MOS range is 0 to 8, then this difference indicates a precision of 0.2296, equivalent to almost 23% difference in quality. Note that theoretically, the difference between good quality of 6 and middle quality of 4 (if all had MOS of exactly 6 and 4), is 0.25, corresponding to 25%, not much different from 23%. Thus, even a small sample of 10 images, show such a high precision on MOS. However, such a difference on the average of SSIM is only 0.9958–0.9757 = 0.0201. Meaning the SSIM precision in discriminating a good image quality from a middle quality is only 2%. This is significantly less than 23% of MOS and shows its weakness in assessing video/image quality at high IQA range. On the other hand, the difference between LF-SSIM from good to middle quality is 0.9387 − 0.8450 = 0.0937. This is equivalent to nearly 9.4%, almost 4.7 times better precision than SSIM at this quality range.
2. The proposed method not only improves precision at high quality range, as shown above, it also has a better performance at middle and bad quality ranges as well. This can be investigated by

Table 3 Values of each IQA metric and their LF-SSIM along with MOS for middle quality images from database tid2013 [11]

image	MOS	SSIM	LF-SSIM	LF2-SSIM	LF3-SSIM
'i01_02_4.bmp'	4.7143	0.9805	**0.8602**	**0.8033**	**0.6618**
'i01_17_3.bmp'	4.0811	0.9632	**0.8082**	**0.7312**	**0.5835**
'i03_04_4.bmp'	4.6000	0.9782	**0.8523**	**0.7923**	**0.6492**
'i03_08_3.bmp'	4.4737	0.9706	**0.8287**	**0.7595**	**0.6133**
'i04_17_3.bmp'	4.7805	0.9741	**0.8391**	**0.7740**	**0.6290**
'i05_01_3.bmp'	4.8462	0.9791	**0.8553**	**0.7964**	**0.6539**
'i05_02_5.bmp'	4.2564	0.9785	**0.8533**	**0.7937**	**0.6508**
'i05_19_3.bmp'	4.2250	0.9837	**0.8724**	**0.8203**	**0.6815**
'i07_06_2.bmp'	4.6136	0.9731	**0.8361**	**0.7697**	**0.6243**
'i23_04_5.bmp'	4.5294	0.9758	**0.8444**	**0.7813**	**0.6370**
Avg	4.5120	0.9757	**0.8450**	**0.7822**	**0.6384**

LF-adapted versions of IQA values are shown in bold.

Table 4 Values of each IQA metric and their LF-SSIM along with MOS for bad quality images from database tid2013 [11]

image	MOS	SSIM	LF-SSIM	LF2-SSIM	LF3-SSIM
'i03_09_5.bmp'	2.4737	0.8091	**0.5631**	**0.4123**	**0.2984**
'i20_10_5.bmp'	2.7436	0.8710	**0.6409**	**0.5088**	**0.3774**
'i09_07_5.bmp'	2.7097	0.8489	**0.6113**	**0.4715**	**0.3463**
'i06_07_5.bmp'	2.6857	0.8874	**0.6644**	**0.5389**	**0.4032**
'i25_10_4.bmp'	2.1471	0.9280	**0.7317**	**0.6274**	**0.4822**
'i08_07_5.bmp'	2.3871	0.8837	**0.6590**	**0.5320**	**0.3972**
'i13_15_1.bmp'	2.7317	0.9451	**0.7657**	**0.6732**	**0.5256**
'i18_22_4.bmp'	2.9524	0.9127	**0.7045**	**0.5913**	**0.4493**
'i08_03_4.bmp'	2.9688	0.8774	**0.6498**	**0.5202**	**0.3871**
'i25_22_5.bmp'	2.7941	0.9015	**0.6861**	**0.5672**	**0.4278**
Avg	2.6594	0.8865	**0.6676**	**0.5443**	**0.4095**

LF-adapted versions of IQA values are bolded.

looking at the average values of MOS, IQA of SSIM and its LF version (LF-SSIM), in going from middle to bad quality. In this case for MOS, the difference between them is: 4.5120–2.6594 = 1.8526. This on 0–8 MOS scale is equivalent to 1.8526/8 = 0.2316, which is 23.16% precision (again very close to the theoretical value of 25%). Such a precision on the SSIM discrimination between middle and bad quality is: 0.9757 − 0.8865 = 0.0894, equivalent to 8.9% precision. However, this precision for LF-SSIM is: 0.8450 − 0.6676 = 0.1774. This means the precision of LF version is 17.74%, which is much closer to MOS discrimination value of 23.16% than the SSIM alone of 8.9%. It is almost twice the precision of SSIM at bad-middle quality range.

We have tested the above scenarios with all the SSIM family measuring tools. They showed almost the same behavior as was explained above on SSIM.

The above analysis, indicates that, discrimination of the Logistic Function version of SSIM not only is more than 4 times better than SSIM itself at good-to-middle quality range, but its precision at middle-to-bad quality is still twice better.

It is worth noting that the used Logistic Function (LF) in Eq. (1), can be defined in a variety of ways. For instance, if we define the Logistic Functions according to Eqs. (2) and (3) as:

$$LF_2 = 1 - \sqrt[2]{1 - IQA^2} \tag{2}$$

$$LF_3 = 1 - \sqrt[3]{1 - IQA^2} \tag{3}$$

then, both of them similar to LF of Eq. (1) have the kind of non-linearity to discriminate the SSIM values. We have added their values to the last two columns of Table 2. Similar to the procedure under items 1 and 2 above, we can calculate the good-to-middle and middle-to-bad discrimination of

Table 5 Precision of discrimination between Good, Middle and Bad quality

Discrimination precision between quality bands [%]	MOS	SSIM	LF-SSIM	LF_2-SSIM	LF_3-SSIM
Good quality from medium quality	22.96	2.01	9.37	13.13	16.86
Medium quality from bad quality	23.16	8.94	17.74	23.79	22.90

these two new functions. Table 5 shows precision of discrimination of these two new functions along with those of MOS, SSIM and LF-SSIM of Eq. (1).

The Table shows that according to the rule defined in [11], discrimination of good quality from middle quality under SSIM is only 2%, significantly lower than almost 23% of MOS. However, in this quality range, all the logistical functions of LF, LF_2 and LF_3, have significantly improved the precision, by a factor 5–8 times better. In the middle-to-bad quality range, although the precision of discrimination under SSIM is not too bad, but again precisions under all three logistic functions have come very close to that of MOS.

It should be noted that, when the precision of discrimination between two quality bands (e.g., good-to-middle quality) is improved, then it can be concluded that the precision of discrimination between the scores within each band (e.g., good quality) is also improved. To prove this, we can either use the rule of [11] and group each quality band into three regions of upper, middle and lower sections, and process their averages. Or we may simply calculate the standard deviations of each quality band of Table 2, to indicate the spread of scores in each band. Table 6 shows the standard deviation within each quality band of data in Table 2, 3 and 4, for good, middle and bad quality, respectively. For MOS, since it is defined in the range of 0–8, the values are normalized to unity.

Although standard deviations based on only 10 samples cannot be very reliable, nevertheless they show the trend of the spread of measured scores in each quality band. The Table shows that SSIM scores within the good and middle quality bands are very dense, but within the bad quality range is close to that of MOS. On the other hand, spreads of scores within the good and middle quality ranges of all three logistic functions are very close to MOS. For these functions, at the bad quality range, spread of scores compared to MOS is over exaggerated.

4 Logistic Function (LF) for VMAF, in video quality assessment

Image quality assessment IQA parameters can also be used to measure video quality, since video is made up of a series of video frames/pictures. For instance, in [15], it is reported that for a 10 s video, average of 20% of worst IQAs measured on frame-by-frame bases has a very high correlation with the subjective scores. However, since the advent of Video Multimethod Assessment Fusion (VMAF), that predicts subjective video quality by diffusing multiple quality metrics into a single score through machine learning optimization, this method has become the most popular video quality meter. It was originally collaboratively devised by researchers in Netflix and colleagues of Professor CC Kuo, at the university of Southern California [8]. However, over the years, through more works and tests like deep learning and better training, it has become a de facto method for video quality assessment, and almost everyone uses it [17].

It would be interesting to see if the used logistic function (LF) for image quality meters, can also improve VMAF's performance for video. We have tested Netflix database, comprising of

Table 6 Percentages of standard deviations within each quality band

Discrimination Precision within each quality band [%] STD	MOS	SSIM	LF-SSIM	LF_2-SSIM	LF_3-SSIM
Good quality	2.80	0.28	2.21	3.12	4.77
Medium quality	3.17	0.58	1.81	2.51	2.76
Bad quality	3.19	3.92	5.83	7.53	6.56

75 video sequences and the LIVE video of 150 video sequences. The scatter diagram of these sequences, for VMAF and its LF version (LF-VMAF, which uses VMAF metric in Eq. (1) are shown in Fig. 4. In these tests, while PLCC of VMAF for Netflix sequence was 0.965, this value with LF-VMAF was 0.946 (0.02 point worse). On the Live video, while PLCC of VMAF to MOS was 0.7549, this value with LF-VAMF was 0.7704 (almost 0.02 better).

It is important to note that, these scatter diagrams, compared with those of images, shown in Figs. 1 and 3, that contain more than 1000 + images (e.g., tid2013 has 3000 images), are very sparse. These sparsely scattered points do not have enough data to show the saturation limitation of video quality meters. Had we had in the order of 1000 video clips we could have better results with LF-VMAF. This can be verified by comparing LIVE and Netflix datasets, where they have respectively 150 and 75 video sequences, where the dataset with larger video sequences has a better performance with LF-VMAF. It should be noted that testing with larger database of video sequences is labor intensive, as 150 video sequences of LIVE video required 40GB storage.

5 Conclusions

Image quality assessment (IQA) tools are widely used in evaluating quality of processed images. They belong to a family of structural similarity index (SSIM) method, that correlate

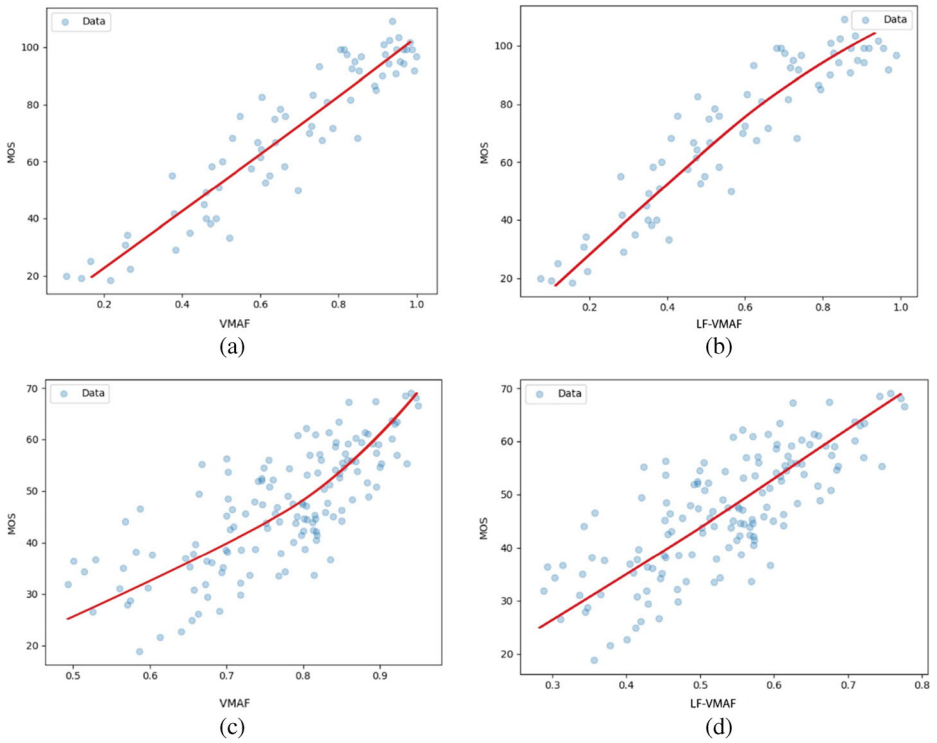

Fig. 4 Scatter plots of subjective MOS against scores, Netflix database (**a**) VMAF, (**b**) **LF**-VMAF, and Live database (**c**)VMAF, (**d**) **LF**-VMAF

well with the human visual systems behavior. Through more than two decades, numerous versions of SSIM-based image quality assessment meters have been devised. Their test results show some improvements of one method over the other. However, they all suffer from loss of precision, especially at high image quality range.

In this paper we have shown that a simple logistic function can be added to the outcome of these measuring devises, to improve their precisions. Throughout the experiments we have shown that, the added logistic function not only improves precision at high quality images, those of low-quality ones can also be improved. This improvement in precision can also increase the Pearson Linear Correlation Coefficient (PLCC) of the objective measures with the mean opinion scores (MOSs). For all images of databases listed in Table 1, while Pearson correlation of MOS to LF of any measured device has a minimum improvement of 2%, its maximum improvement is as high as 20%.

For analysis of a large database of images, if we divide them into three groups of bad, middle and good quality images, while the MOS of good-quality images, had almost 23% precision, this value for IQAs at this quality was only 2%, but their adapted logistic function at this quality was 9.4 − 17%. Such modification could also improve measured quality precision at middle to bad quality. In this case, while precision of MOS at these quality ranges was 23.2%, that of raw IQA was 8.9% and their logistic function version was increased to 17.7 − 23%, very close to the MOS range.

Finally, we have tested the impact of defined logistic function on video quality meter, especially the widely used VMAF. Although a limited number of video sequences were tested, but their outcomes indicate that a logistic function can also improv the precision of video quality meters too.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Baig MA, Moinuddin AA, Khan E, Ghanbari M (2022) DFT-based no-reference quality assessment of blurred images. Multimed Tools Appl. https://doi.org/10.1007/s11042-022-11992-3
2. Brunet D, Vrscay ER, Wang Z (2012) On the mathematical properties of the structural similarity index. IEEE Trans Image Process 21(4):1488–1499
3. Cisco (2019) Cisco visual networking index: Forecast and methodology, 2017{2022 (White Paper)
4. Ghanbari M (2011) Video quality measurement. US Patent 7,869,517
5. Huynh-Thu Q, Ghanbari M (2008) Scope of validity of PSNR in image/video quality assessment. Electron Lett 44(13):800–801

6. Kazemi M, Ghanbari M, Shirmohammadi S (2020) The performance of quality metrics in assessing error-concealed video quality. IEEE Trans Image Process 29:5937–5952. https://doi.org/10.1109/TIP.2020.2984356
7. Larson EC, Chandler DM (2009) Categorical Image Quality (CSIQ) Database, [Online]. Available: http://vision.okstate.edu/csiq. Access dates 3 April 2022
8. Liu TJ, Lin W, Kuo CCJ (2012) Image quality assessment using multi-method fusion. IEEE Trans Image Process 22(5):1793–1807
9. Mittal A, Moorthy AK, Bovik AC (2012) No-reference image quality assessment in the spatial domain. IEEE Trans Image Process 21(12):4695–4708
10. Ponomarenko N, Lukin V, Zelensky A, Egiazarian K, Carli M, Battisti F (2009) TID2008—A database for evaluation of full-reference visual quality assessment metrics. Adv Mod Radioelectron 10:30–45
11. Ponomarenko N, Jin L, Ieremeiev O, Lukin V, Egiazarian K, Astola J, Vozel B, Chehdi K, Carli M, Battisti F, Kuo CCJ (2015) Image database tid2013: Peculiarities, results and perspectives. Sig Process Image Commun 30:57–77
12. 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
13. ShiZaifeng ZhangJiaping, CaoQingjie, PangKe LuoTao (2018) Full-reference image quality assessment based on image segmentation with edge feature. Signal Process vol 145:99–105
14. Sun W, Liao Q, Xue J-H, Zhou F (2018) SPSIM: A superpixel-based similarity index for full-reference image quality assessment. IEEE Trans Image Process 27(9):4232–4244
15. Tan KT, Ghanbari M (2000) A multi-metric objective picture-quality measurement model for MPEG video. IEEE Trans Circuits Syst Video Technol 10(7):1208–1213
16. Viqar M, Moinuddin AA, Khan E, Ghanbari M (2022) Frequency-domain blind quality assessment of blurred and blocking-artefact images using Gaussian Process Regression model. Sig Process Image Commun 116651. https://doi.org/10.1016/j.image.2022.116651
17. VMAF (2017) Perceptual video quality assessment based on multi-method fusion. Netflix, Inc., 2017-07-14, retrieved 2017-07-15
18. VQEG final report (2000) the validation of objective models of video quality assessment. https://www.academia.edu/es/2102517/FINAL_REPORT_FROM_THE_VIDEO...
19. VQEG final report (2003) FR-TV phase 2 validation test. https://vqeg.org/.../2003_03_Intel_USA/VQEGIIDraftReportv2a.pdf
20. Wang Z, Li Q (2011) Information content weighting for perceptual image quality assessment. IEEE Trans Image Process 20(5):1185–1198
21. Wang Z, Bovik AC, Lu L (2002) Why is image quality assessment so difficult? 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing, Orlando, FL, pp IV-3313-IV-3316. https://doi.org/10.1109/ICASSP.2002.5745362
22. Wang Z, Simoncelli EP, Bovik AC (2003) Multiscale structural similarity for image quality assessment. IEEE Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers, vol 2
23. Wang Z, Bovik AC, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13:600–612
24. Xue W, Zhang L, Mou X, Bovik A (2014) Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. IEEE Trans Image Process 23(2):684–695
25. Zhang L, Zhang L, Mou X, Zhang D (2011) FSIM: A feature similarity index for image quality assessment. IEEE Trans Image Process 20(8):2378–2386
26. Zhang L, Shen Y, Li H (2014) VSI: A visual saliency-induced index for perceptual image quality assessment. IEEE Trans Image Process 23(10):4270–4281
27. Zhang W, Borji A, Wang Z, Callet PL, Liu H (2016) The application of visual saliency models in objective image quality assessment: A statistical evaluation. IEEE Trans Neural Netw Learn Syst 27(6):1266–1278