Journal of Universal Computer Science, vol. 23, no. 5 (2017), 505-527
submitted: 13/12/16, accepted: 14/5/17, appeared: 28/5/17

A Comparative Study of Objective Video Quality Assessment Metrics

Carlos A.B.Mello, **Marília M.Saraiva** and **Diego P.A.Menor**
(Centro de Informática, Universidade Federal de Pernambuco, Recife, Brazil
{cabm, mms5, dpam}@cin.ufpe.br)

Ricardo Nishihara
(Motorola Company, São Paulo, Brazil
wrn009@motorola.com)

Abstract: This paper presents a comparison of several video quality metrics, analysing their performance against different types of distortions. Usually, comparisons are made considering a full dataset with few different degradations. We are presenting here a comparison using three very different datasets (VQEG Phase I, LIVE VQA and ReTRIeVED) and a fourth dataset which was generated in a mobile phone network simulator. This was done to check if the video quality metrics can correctly measure the degradations created by variations in the network, very close to real scenarios. The analysis was done with 13 full reference metrics (including Opticom's PEVQ commercial tool) and two no-reference metrics. We have concluded that NTIA's VQM achieved the best results, in most of the cases. It is an open source algorithm that outperformed most of the other techniques, including the licensed PEVQ.

Key Words: video quality assessment, video distortions, PEVQ, SSIM, VQM
Category: J.0, J.6

1 Introduction

With the advances of the Internet, the broadcasting of multimedia files has become one of the most desired services for users. The possibility to access these files through different types of devices, including hand held devices, increased the popularity of audio and video broadcasting. Compression standards, as H.264 or H.265 (also known as High Efficiency Video Coding - HEVC), have reached high video compression rates also taking into account the quality of the images displayed. The compression rates jointly with the band width currently available in several countries have made it possible to access such files even from mobile phones. However, as the bandwidth increases, the user expectations for better services also increases. These services are not just related to Internet service providers, but also to the development of new compression methods. This is a very complex problem for audio and video data, as they deal not only with the reduction of the file size (in a lossy or lossless way), but also with the maintenance of the perceived quality of the signal. This is particularly difficult in videos, as the human being is naturally a visual creature[Firsby and Stone 2010]. Thus, video compression algorithms must deal with a subjective analysis of the compressed file, and not just the satisfaction of compression rate requirements.

Automatic quality analysis is usually made in one of three possible ways: as full-reference, as reduced reference or as no-reference. A full reference method needs an original, undegraded, signal to make a comparison with the target signal. It is very useful in test environments to conduct experiments on coders and decoders. A no-reference method does not require a clean signal; it makes the analysis of the target signal based only on its features and an expectation of what is considered a good quality signal to the human perceiver. A reduced reference algorithm gets the features for quality analysis from the video itself. For example, it can analyse the difference between adjacent frames.

In this paper, we explore the performance of some state-of-the-art metrics for video quality assessment under several different levels of degradations. These degradations come from three different well known datasets and from the simulation of a real scenario of broadcasting in cellular networks. Our major contributions are: (i) the analysis of video quality metrics in videos under 10 different types (and levels) of degradations; (ii) the analysis of video quality metrics in an almost real scenario in cell phone devices (provided by a mobile network simulator); (iii) the analysis of state of the art no-reference video quality metrics; (iv) a final analysis of cost *versus* benefit of the metrics, considering also Opticom's PEVQ commercial tool.

This paper is divided as follows: in the next section, the metrics under analysis are described. In Section 3, the experiments are presented and the results are discussed, while Section 4 concludes the paper.

2 Video Quality Assessment Metrics

Due to the complexity of the subject, there have been several attempts to automatically evaluate video quality. These approaches were first derived from static image evaluation in a frame-by-frame analysis. Some new approaches consider that a video frame must be related to its previous and following frames; other methods work with analysis in different domains (such as frequency). We have chosen a set of 10 algorithms (some of them have variations making a total amount of 13 metrics). The criteria used in this choice was the applicability (for example, SSIM can be found in testing devices as Rohde-Schawrz CMW500 - a communication tester), industrial use (as Opticom's PEVQ), quality of results (as NTIA's VQM) and novelty (as BLIINDS and VIIDEO). Among the many variations created from SSIM, some of them have more innovative proposals: MS-SSIM, stSSIM, GSSIM and 3SSIM. There are also some different implementations of SSIM that lead to different results: they are called "precise" and "fast". The only difference is in how they evaluate the variables of SSIM based on a Gaussian filter (for precise version) or an average filter (for fast version). The same happens to MS-SSIM. There is still another implementation of SSIM that comes with MathWorks' MatLab that presented different scores when compared to the other versions. The algorithms are divided into full-reference and no-reference. Most of them are full-reference with just BLIINDS and VIIDEO as no-reference methods, due to the difficulty in creating such algorithms. No reduced reference algorithm was implemented as there are none with high scientific impact, to the best of our knowledge.

2.1 Full-Reference Metrics

2.1.1 MSE and PSNR

The Mean Squared Error (MSE) is probably the simplest way to compare two signals. It is just the summation of the absolute difference between the two signals, being evaluated as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(x(i) - y(i))^2 \tag{1}$$

where x and y are the signals, both with length N. One of the drawbacks of MSE is that it is not limited to a specific range of values, which means that it only makes some sense when compared to other MSE values. As an error, the lower, the better; thus a MSE of 30 is better than a 40, but this is only a relative value. The other major problem is that this metric does not take into account any structural information between the signals. For example, it is not suitable to compare two texture images even if they are from the same material; it can also fail in the comparison of two speech signals even if they are the same words spoken by the same person. As its value can increase rapidly, a scaled version of it is proposed in the evaluation of Peak Signal to Noise Ratio (PSNR):

$$PSNR = log\frac{C^2}{MSE} \tag{2}$$

where C is the maximum possible value that can be found in the signal. Of course, PSNR has the same problems as MSE but in a different range of values. For more about the problems related to the use of MSE and PSNR as video quality metrics, please see [Wang and Bovik 2006].

2.1.2 Video Quality Metric - VQM

VQM stands for Video Quality Metric. As it is a common name, there are different algorithms with the same name. The most universally accepted is presented in [Pinson and Wolf 2004]. This general model and its associated calibration techniques comprise a complete automated objective video quality measurement system. The calibration of the original and the processed video streams includes spatial alignment, valid region estimation, gain-level offset calculation, and temporal alignment. VQM calculation involves extracting perception-based features, computing video quality parameters, and combining these parameters to construct the general model.

An example of other VQM metrics is a DCT-based approach that can be found in the MSU Video Quality Measurement Tool [MSU 2005]. It is based on wavelet analysis.

2.1.3 SSIM and Variations

In [Wang and Bovik 2002], Z.Wang and A.C.Bovik introduced an universal quality index for image comparison. The Q index is defined as:

$$Q = \frac{4.\sigma_{xy}.\overline{x}.\overline{y}}{(\sigma_x^2 + \sigma_y^2).(\overline{x}^2 + \overline{y}^2)} \tag{3}$$

with x and y as reference and target images, σ_x, σ_y, $\overline{x}$ and $\overline{y}$ are the standard deviation and average values of the colours for both images, respectively. σ_{xy} is the correlation between the images. The index ranges between -1 and 1 with the 1 value reached for identical images. The index considers the correlation and the similarity between luminance and contrast of the images to reach its final score.

The Q index fails in some situations; for example, again, in the comparison of texture images. Although textures (of a same pattern) could present perceived similarities, the index can not detect them. A structural similarity index (SSIM) was then proposed in [Wang et al 2004b] (it was further called SS-SSIM - for Single Scale SSIM, in opposition to MS-SSIM - Multi-Scale SSIM [Wang et al 2003]). The new metric was proposed to evaluate the structural similarity between two images, comparing local patterns that are normalized for luminance and contrast. To create the new index, it was considered that while the luminance of the surface of an object under observation is the product of the brightness and its reflexive properties, the structure of the objects in the scene does not depend of the illuminant. The structural features are then the attributes that represent the structure of the objects of the scene, independently of average luminance and contrast. This is why local luminance and contrast are used in the index definition. SSIM can be evaluated as follows:

$$SSIM = \frac{(2\overline{x}.\overline{y} + C_1).(2.\sigma_{xy} + C_2)}{(\overline{x}^2 + \overline{y}^2 + C_1).(\sigma_x^2 + \sigma_y^2 + C_2)} \tag{4}$$

where $C_1 = (K_1L)^2$ and $C_2 = (K_2L)^2$ are constants, with L as the maximum value of the pixel color (255 for a 8-bits image) and $K_1 << 1$ and $K_2 << 1$.

The Q index is a special case of this equation when $C_1 = C_2 = 0$. This index was further improved to deal with video analysis [Wang et al 2004a]. For video quality assessment, the frames are analysed in YCbCr color space. The index is applied for each component and each frame of the videos under comparison. In fact, the frames are divided into windows and SSIM is applied at each valid window (a window is considered invalid if it has a low contrast value). Weights are assigned to the score generated between windows and for each component (Y, Cb and Cr) to produce the final score.

Due to its simplicity and good results (in ordinary situations) several variations of SSIM have been developed. We summarize some of them next with other variations stated in the experiments section:
1) MS-SSIM [Wang et al 2003]: The scale-space theory [Witkin 1983] is the motivation behind the MS-SSIM, or Multi-Scale SSIM. As in the original SSIM, luminance, contrast and structural information are extracted from the image. The difference is that it is done in different scales of the original and reference images. To simulate the reduction in scale, the images are low-pass filtered and downsampled by a factor of 2. The final score is a combination of the features extracted from each scale.
2) stSSIM [Moorthy and Bovik 2010]: Spatio-Temporal video SSIM came to bring more velocity to MOVIE algorithm (MOtion-based Video Integrity Evaluation index) [Seshadrinathan and Bovik 2010], but keeping its efficiency. It starts with a version of SSIM that runs on videos, analysing temporal neighbours of a pixel (i.e., the past and future values of that pixel in a video). This temporal evaluation creates a score for each pixel. A temporal score is evaluated for each frame (considering all the pixels of the frame) and the final score comes from the product of the average value of the temporal scores and the spatial score (evaluated using common SSIM).
3) 3SSIM[Li and Bovik 2009]: The three component SSIM was proposed to improve SSIM in blurred and noisy images (or frames in a video). The original image is decomposed into edge, textures and smooth regions. Different weights are associated to each one of these regions considering that they produce different stimuli to our perceptive system. Thus, the original reference image is segmented into these three classes and the final score is given by the average of each SSIM value. In practice, three images are created for each reference image and SSIM is evaluated for each one of them with the final index found by the weighted average value of these indexes.

2.1.4 Perceptual Evaluation of Video Quality - PEVQ

OPTICOM's PEVQ [Opticom 2017] is a software for video quality analysis; it can be found in the PEXQ framework which also includes PESQ (Perceptual Evaluation of Speech Quality) software. PEVQ's algorithm is divided into four major blocks: (i) spatial and temporal alignment as a preprocessing step; (ii) evaluation of perceptual difference with emphasis on luminance and chrominance domains; (iii) the previous phase creates indexes that are now classified to detect specific distortions; and (iv) all previous scores are combined in the final PEVQ score. The parameters considered in the final score are delay, brightness, contrast, PSNR and distortion indicators. PEVQ is a commercial software which requires the acquisition of a license (its cost was 9,000.00 euros in 2015 per one single license for one unique computer - it cannot be re-installed in another computer).

2.2 No-Reference Metrics

2.3 Video BLIINDS

Video BLIINDS [Saad et al 2014] is a no-reference image quality assessment algorithm with an approach that relies on a spatio-temporal model of video scenes in the discrete cosine transform domain, and on a model that characterizes the type of motion occurring in the scenes, in order to predict video quality. The major point is the analysis of the statistical distribution of the locally evaluated discrete cosine transform (DCT) coefficients over frame-difference. The empirical probability distribution of frame difference of the DCT coefficients evaluated from the pristine video are more heavy-tailed than the DCT coefficients of the distorted video. To capture both spatial and temporal local frequencies, a 2-dimensional spatial DCT is applied to frame-difference patches in n $\times$ n blocks (to guarantee locality). A 1D generalized Gaussian function is used to model the probability distribution of these coefficient histograms. Several features are extracted for prediction as: the motion coherency measure and the global motion measure which are key characterizations of the temporal behaviour exhibited by a video sequence, five Natural Video Statistics (NVS) spectral ratios (shape parameters), absolute temporal derivative of mean DC coefficients, and the purely spatial frame naturalness measure. Each feature is computed from each frame difference (except the spatial naturalness measure), then temporally pooled over a 10 second interval. Prior to feeding the features into the SVR (Support Vector Regression), the spatio-temporal features (other than the naturalness index) are subjected to a logarithmic non-linearity. Quality prediction is then performed on the entire video segment. A simple scheme of BLIINDS can be seen in Figure 1 .

2.3.1 Video Intrinsic Integrity and Distortion Evaluation Oracle - VIIDEO

The next step on no-reference metrics is VIIDEO[Mittal et al 2016]. Although BLIINDS provides a good answer for quality prediction, it is too slow with

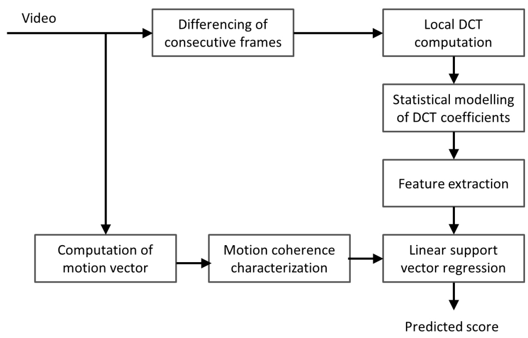


Figure 1: BLIINDS no-reference video analysis scheme.

minutes of execution time in MatLab using the implementation provided in [Mittal et al 2016]. In this sense, VIIDEO provided a faster solution, although not so accurate as BLIINDS (lower correlations). Another advantage of VIIDEO is that it does not require the use of any additional information other than the video being evaluated. The method is based on a representation of the video in coefficients generated from the frame differences, processed considering local mean and divided by contrast normalization. All the local features are evaluated on a 3×3 window. A Gaussian weighted function is used to act as a semi-saturation constant. It was observed that the histogram of the coefficients of pristine videos are Gaussian-like functions, wherein each distortion modifies the histograms in an specific way. The product of neighbouring coefficients has shown to be well-modelled as following a zero mode asymmetric generalized Gaussian distribution (AGGD) with parameters that can be efficiently estimated in four orientations, amounting 12 parameters that are used to evaluate the final score.

3 Datasets

We have used three well known datasets (VQEG Phase I, LIVE VQA and ReTRiEVED) and a dataset created in our experiments with a cellular network simulator. The major features of the first three datasets are explained next, while our dataset is detailed in the experiments section.

3.1 VQEG FR-TV Phase I

The Video Quality Expert Group (VQEG) [VQEG 2000] has 320 test sequences and it is divided into 25 FPS(50Hz)/low bitrate, 25 FPS(50Hz)/high bitrate, 30 FPS(60Hz)/low bitrate, and 30 FPS(60Hz)/high bitrate sets. About the data format, the videos are either 625/50 or 525/60 format, in YCbCr color system. The term 625/50 refers to a system of 625 scanning lines at 50 hertz (PAL/SECAM). This 50 hertz scanning rate produces a full frame every 1/25 of second. A similar description can be used for the 525/60 format (found in PAL-M). In our tests, the comparison was made in just two sets, 25 FPS (50Hz) and 30 FPS (60Hz). Figure 2 presents some samples from this dataset.

Figure 2: VQEG Phase I sample frames.

3.2 LIVE VQA

The LIVE Video Quality Assessment database [LIVE 2003] uses ten uncompressed high-quality videos with a wide variety of content as reference videos. A set of 150 distorted videos were created from these reference videos (15 distorted videos per reference) using four different distortion types - MPEG-2 compression, H.264 compression, simulated transmission of H.264 compressed bit streams through error-prone IP networks and through error-prone wireless networks. Some sample frames can be seen in Figure 3.

Distortion strengths were adjusted manually to ensure that the different distorted videos were separated by perceptual levels of distortion. Each video in the LIVE Video Quality Database was assessed by 38 human subjects in a single stimulus study with hidden reference removal, where the subjects scored the video quality on a continuous quality scale.

Figure 3: LIVE VQA sample frames.

3.3 ReTRiEVED

The ReTRiEVED dataset [ReTRiEVED 2014] was first introduced in the work of [Seshadrinathan et al 2010]. It is composed of 184 degraded videos created from eight undistorted videos with different types of content. The degradations applied to the videos are related to possible problems in video broadcasting: packet loss, jitter, delay, and throughput. About the eight original videos, six of them are from EPFLPoliMI [EPFL 2009] dataset and the other two are from the Consumer Digital Video Library [CDVL 2003]. They have different frame rate, bit rate and length. Just the dimensions of the frames are the same (704×576 pixels). 41 subjects ranked the videos in a MOS experiment. Figure 4 illustrates ReTRiEVED dataset with sample frames from videos from the Packet Loss Rate (PLR) distortion.

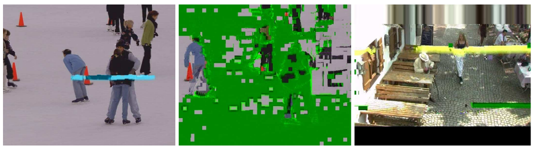


Figure 4: Sample frames with packet loss rate from ReTRiEVED dataset.

As it can be seen in [Seshadrinathan et al 2010], the dataset is comprised of clusters of videos with different temporal and spatial information. Some of them has more spatial information than temporal information (as the soccer video), while others have the opposite relation (as the ducks take off video). With this in mind, it is possible to infer the possible behaviour of some algorithms according to the results found. In fact, the ReTRiEVED database is a great challenge for every algorithm, as we present in the next section.

4 Experiments

The experiments are divided into two: first, the metrics will be tried against all the three well known datasets (LIVE VQA, VQEG Phase I and ReTRIeVED). After this, a simulation scenario is used based on a cellular network model provided by Anritsu's MD847503A.

To analyse the performance of the algorithms, Pearson Linear Correlation Coefficient (PLCC), Spearman Rank Order Correlation (SROCC) are used just as Root Mean Squared Error (RMSE). While SROCC measures how the relation between two variables (in our case, the automatic score and the human score) can be described by a monotonic function, PLCC measures the prediction accuracy. In both cases, 1 represents the best value. PLCC is evaluated after a non-linear regression on the video quality assessment algorithm scores to map them into DMOS (Differential Mean Opinion Score) scores using a five parameters logistic function. Scatter plots are usually helpful to give a visual understanding of the behaviour of the algorithms. They are presented in some cases just for a better understanding of the results. For RMSE, as a measure of error, the lower, the better.

4.1 Datasets

For the datasets reported in Section 3, the experiments run for full-reference metrics (FR) and no-reference metrics (NR) in separate. Currently, NR metrics are still under development with high advances but still not so good as full reference metrics. Because of this, we avoid a direct comparison between no-reference algorithms and full reference algorithms.

4.1.1 Full Reference Metrics

Tables 1 to 3 present the results (Pearson, Spearman and RMSE) for VQEG Phase I dataset, Tables 4 to 6 do the same for LIVE VQA dataset, while Tables 7 to 9 present the scores for ReTRiEVED dataset. After each set of tables, an analysis of the results is discussed. In all tables, the best score is in bold.

For the videos from VQEG Phase I dataset, in general, we can conclude that the VQM algorithm from NTIA [Pinson and Wolf 2004] is the best choice. Although there are other better results from SROCC, NTIA's VQM correlation is very close to these best values (stSSIM had the best score) at these few situations. When we go down to details and analyse each video, some aspects can be highlighted. Observing Pearson correlation of the degraded videos against the original videos, we can see that no metric has a good performance for the videos SRC16 and SRC19, especially, SSIM and its variations. Both videos have

Table 1: PLCC for full reference metrics on VQEG Phase I

Metric	50Hz	60Hz	All
SSIM (fast)	0.7796	0.8258	0.8034
SSIM (precise)	0.7892	0.7816	0.8087
SSIM (MatLab)	0.8114	0.8110	0.8164
GSSIM	0.7052	0.7461	0.7139
MS-SSIM (fast)	0.6598	0.7210	0.7342
MS-SSIM (precise)	0.6374	0.2479	0.3771
stSSIM	0.6983	0.7700	0.6651
3SSIM	0.8070	0.7410	0.7682
VQM MSU	0.4056	0.3780	0.3275
VQM NTIA	**0.8501**	**0.8853**	**0.8645**
PEVQ	0.7932	0.8037	0.7997

Table 2: SROCC for full reference metrics on VQEG Phase I

Metric	50Hz	60Hz	All
SSIM (fast)	-0.7723	-0.7135	-0.7591
SSIM (precise)	-0.7828	-0.7382	-0.7668
SSIM (MatLab)	0.8169	0.7143	0.7835
GSSIM	0.6577	0.6306	0.668
MS-SSIM (fast)	-0.7427	-0.5776	-0.6622
MS-SSIM (precise)	-0.6133	-0.1613	-0.3110
stSSIM	**0.8588**	**0.8588**	0.8102
3SSIM	-0.8104	-0.6213	-0.7141
VQM MSU	0.3822	0.2389	0.2124
VQM NTIA	0.8161	0.8151	**0.8301**
PEVQ	0.8068	0.7510	0.7904

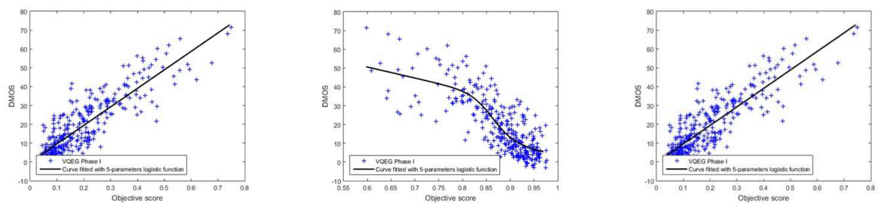
a lot of space activity and sudden movements (like scene cuts and fast camera movements).

In Figure 5, scatter plots for PEVQ, MatLab's SSIM and NTIA's VQM are shown.

At LIVE database, the best results came from PEVQ, NTIA's VQM and MS-SSIM. In this case, PEVQ has the best response for correlation and error. In a low level, most part of the metrics have low performance for the videos 'pa' (Pedestrian Area) and 'sf' (Sun Flower), with the worst Pearson correlation. The 'pa' video has a fixed camera recording people walking on a street. The 'sf'

Table 3: RMSE for full reference metrics on VQEG Phase I

Metric	50Hz	60Hz	All
SSIM (fast)	10.4150	7.6314	9.1581
SSIM (precise)	10.2129	13.1728	9.0459
SSIM (MatLab)	9.7213	7.9173	8.8814
GSSIM	11.7917	9.0095	10.7690
MS-SSIM (fast)	12.4974	9.3767	10.4414
MS-SSIM (precise)	12.8142	13.1096	14.2436
stSSIM	11.9045	8.6342	12.8657
3SSIM	9.8210	9.0863	9.8465
VQM MSU	15.2016	12.5279	14.5310
VQM NTIA	**8.7569**	**6.2933**	**7.7312**
PEVQ	10.1270	8.0520	9.2344



(a) (b) (c)

Figure 5: Scatter plots for: (a) PEVQ, (b) MatLab's SSIM and (c) NTIA's VQM; all of them applied to VQEG Phase I dataset.

video is focused on a bee flying close to a sun flower with soft movements of the camera.

Scatter plots for PEVQ, MatLab's SSIM and NTIA's VQM applied to LIVE dataset are presented in Figure 6.

The ReTRiEVED dataset is the most complex database among the three presented in this paper. In general, the worst results are from the videos 'Duckstakeoff' and 'ParkJoy'. According to the original paper[Seshadrinathan et al 2010], these videos are the ones which have more temporal information of the entire base, and also a high amount of spatial information. SSIM and SSIM-like metrics are the ones with worst results. NTIA's VQM presented the best results for almost all kinds of degradation in all the three measures (SROCC, PLCC and RMSE). As it was said before, the ReTRiEVED dataset is very difficult to deal due to its different types of degradations and very complex set of original videos.

Table 4: PLCC for full reference metrics on LIVE VQA

Metric	Wireless	IP	H.264	MPEG-2	All
SSIM (fast)	0.6146	0.6748	0.7275	0.6414	0.5614
SSIM (precise)	0.5454	0.5970	0.6957	0.5785	0.4994
SSIM (MatLab)	0.5233	0.5000	0.6066	0.5627	0.4776
GSSIM	0.6001	0.5347	0.7076	0.7328	0.4915
MS-SSIM (fast)	0.7022	**0.7570**	0.6182	0.6606	0.7542
MS-SSIM (precise)	0.6979	0.6725	0.6024	0.6573	-0.6876
stSSIM	0.7450	0.7129	0.7355	0.7055	0.7361
3SSIM	0.6644	0.6745	0.7199	0.6787	0.6366
VQM MSU	0.6211	0.5569	0.6473	0.50878	0.6231
VQM NTIA	0.7412	0.6729	0.6153	**0.7730**	0.7195
PEVQ	**0.8295**	0.7037	**0.8048**	0.7676	**0.7766**

Table 5: SROCC for full reference metrics on LIVE VQA

Metric	Wireless	IP	H.264	MPEG-2	All
SSIM (fast)	0.5910	0.5230	0.7066	0.5982	0.5998
SSIM (precise)	0.5133	0.4523	0.6531	0.5527	0.5220
SSIM (MatLab)	0.4619	0.3424	0.5553	0.5063	0.4507
GSSIM	0.5266	0.4727	0.6827	0.7159	0.4902
MS-SSIM (fast)	0.7377	**0.6436**	0.7537	0.6256	0.7358
MS-SSIM (precise)	0.7392	0.6263	0.7366	0.6347	0.7299
stSSIM	0.7407	0.6267	0.7148	0.6772	0.7129
3SSIM	0.6741	0.5858	0.7107	0.6298	0.6232
VQM MSU	0.6250	0.5226	0.4522	0.3924	0.5872
VQM NTIA	0.7236	0.6387	0.6385	**0.7525**	0.6875
PEVQ	**0.7880**	0.6392	**0.7895**	0.7482	**0.7556**

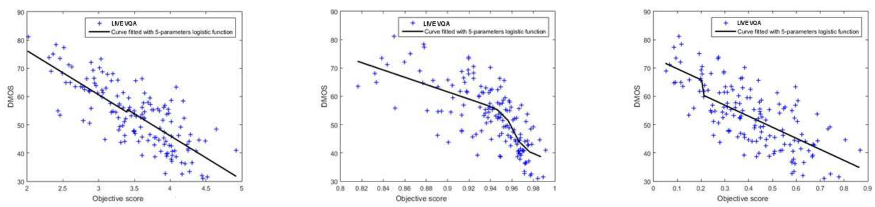
Scatter plots for PLCC for PEVQ, SSIM (fast) and 3SSIM applied to ReTRiEVED dataset are presented in Figure 7.

4.1.2 No-Reference Metrics

As expected, the correlation for no-reference metrics is still too low and the error too high for most part of the cases. If we compare Tables 10 to 18 with the previous tables, it is clear that there is still much to be done. There is just one remarkable case where both no-reference algorithms achieved good results

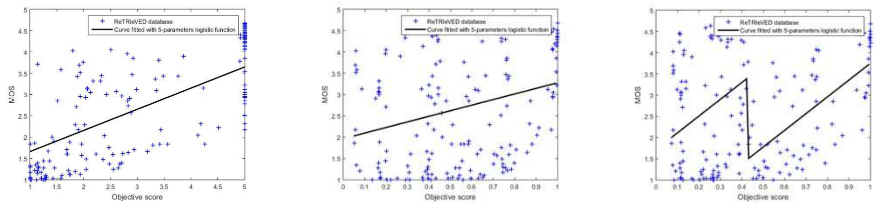
Table 6: RMSE for full reference metrics on LIVE VQA

Metric	Wireless	IP	H.264	MPEG-2	All
SSIM (fast)	8.1401	6.9199	7.4483	7.3177	9.0844
SSIM (precise)	8.6486	7.5248	7.7976	7.7769	9.5104
SSIM (MatLab)	8.7932	8.1060	8.6296	7.8819	9.6443
GSSIM	8.2540	10.3989	7.6714	6.4878	9.5599
MS-SSIM (fast)	11.3092	**6.1163**	8.5327	7.1579	7.2078
MS-SSIM (precise)	7.3901	6.9274	10.9430	7.1853	11.5596
stSSIM	6.8974	6.5761	7.3551	6.7571	7.4308
3SSIM	7.7121	6.9102	**5.5348**	7.0038	8.4658
VQM MSU	8.0872	7.7740	8.2742	8.2322	8.5901
VQM NTIA	6.9369	6.9238	8.5572	6.9487	7.6240
PEVQ	**5.7623**	6.6504	6.4424	**6.1110**	**6.9154**



(a) (b) (c)

Figure 6: Scatter plots for: (a) PEVQ, (b) MS-SSIM and (c) stSSIM; all of them applied to LIVE VQA dataset.



(a) (b) (c)

Figure 7: Scatter plots for: (a) PEVQ, (b) SSIM (fast) and (c) 3SSIM; all of them applied to ReTRiEVED dataset.

comparable to full reference algorithms: VIIDEO's PLCC score for delay degradation in ReTRiEVED dataset is almost 0.9, while BLIINDS's PLCC score is

Table 7: PLCC for full reference metrics on ReTRiEVED

Metric	Delay	Jitter	Throughput	PLR	All
SSIM (fast)	0.4547	0.5914	0.4716	0.1527	0.2772
SSIM (precise)	0.4593	0.5784	0.4937	0.1623	0.2744
SSIM (MatLab)	0.7098	0.4351	0.2726	0.2448	0.2179
GSSIM	0.6038	0.4211	0.4856	0.4935	0.3367
MS-SSIM (fast)	0.8299	0.6364	0.6910	0.4107	0.2883
MS-SSIM (precise)	0.7392	0.6459	0.6830	0.3947	0.2789
stSSIM	0.2868	0.6167	0.4396	0.4363	0.3992
3SSIM	0.7581	0.6498	0.6564	0.4430	0.4547
VQM MSU	0.6552	0.5654	0.6791	0.3594	0.4638
VQM NTIA	**0.8996**	**0.9034**	**0.9752**	**0.8144**	**0.9052**
PEVQ	0.8431	0.8118	0.9663	0.6019	0.7978

Table 8: SROCC for full reference metrics on ReTRiEVED

Metric	Delay	Jitter	Throughput	PLR	All
SSIM (fast)	0.3736	0.1617	0.5237	0.1877	0.2935
SSIM (precise)	0.4332	0.1283	0.5108	0.1880	0.2846
SSIM (MatLab)	0.6695	0.0759	0.2480	0.2112	0.2018
GSSIM	0.6690	0.0488	0.2281	0.4936	0.3039
MS-SSIM (fast)	0.2926	0.2932	0.5612	0.1832	0.3128
MS-SSIM (precise)	0.4332	0.1283	0.5108	0.1880	0.2846
stSSIM	-0.0225	0.4578	0.2853	0.3053	0.2774
3SSIM	0.2040	0.2299	0.5808	0.1797	0.2227
VQM MSU	-0.3156	-0.3598	-0.6995	-0.4050	-0.4936
VQM NTIA	0.7510	**0.9358**	**0.9574**	**0.8336**	**0.9219**
PEVQ	**0.7971**	0.6184	0.9407	0.5436	0.8150

0.7 (see Table 16). VIIDEO achieved a correlation higher than the best value of full reference metrics.

4.2 Cellular Network Simulation Scenario

The next experiment tried to create more realistic scenarios for mobile phone broadcasting. A cellular network is created and a connection between two mobile phones is established. Different settings of the cellular network are adjusted in order to create different scenarios for a video streaming service between both phones. The degraded videos received in each scenario are analysed.

Table 9: RMSE for full reference metrics on ReTRiEVED

Metric	Delay	Jitter	Throughput	PLR	All
SSIM (fast)	0.4653	0.7649	1.2221	0.8598	1.2407
SSIM (precise)	0.4640	0.7738	1.2052	0.8583	1.2418
SSIM (MatLab)	0.3679	0.8540	1.3334	0.8433	1.2603
GSSIM	0.4164	0.8603	0.7789	1.2053	1.2159
MS-SSIM (fast)	0.2914	0.7316	1.0018	0.7930	1.2365
MS-SSIM (precise)	0.3518	0.7241	1.0172	0.7992	1.2401
stSSIM	0.5004	0.7469	1.2448	0.7826	1.1840
3SSIM	0.3406	0.7210	1.0468	0.7798	1.1501
VQM MSU	0.3946	0.7824	1.0427	0.8117	1.1540
VQM NTIA	**0.2281**	**0.4068**	**0.2798**	**0.5047**	**0.5341**
PEVQ	0.2809	0.5539	0.3568	0.6946	0.8183

Table 10: PLCC for no-reference metrics on VQEG Phase I

Metric	50Hz	60Hz	All
VIIDEO	0.2128	0.0015	0.1415
BLIINDS	0.1282	0.2572	0.0644

Table 11: SROCC for no-reference metrics on VQEG Phase I

Metric	50Hz	60Hz	All
VIIDEO	0.0887	0.0605	0.0143
BLIINDS	0.2106	0.2228	0.1590

Table 12: RMSE for no-reference metrics on VQEG Phase I

Metric	50Hz	60Hz	All
VIIDEO	16.1481	13.5320	15.1878
BLIINDS	16.4984	13.0767	15.3473

Table 13: PLCC for no-reference metrics on LIVE VQA

Metric	Wireless	IP	H.264	MPEG-2	All
VIIDEO	0.5170	0.4500	0.5730	0.6870	0.6040
BLIINDS	0.5434	0.7731	0.6806	0.7103	0.6219

Table 14: SROCC for no-reference metrics on LIVE VQA

Metric	Wireless	IP	H.264	MPEG-2	All
VIIDEO	0.5100	0.4630	0.5520	0.6250	0.5690
BLIINDS	0.5583	0.7182	0.6248	0.7196	0.5958

Table 15: RMSE for no-reference metrics on LIVE VQA

Metric	Wireless	IP	H.264	MPEG-2	All
VIIDEO	8.4685	8.4051	7.8984	7.8173	8.7473
BLIINDS	8.6621	5.6887	7.9531	6.7115	8.5963

For the scenario creation, we have used Anritsu's MD8475A device. It is an all-in-one base station simulator that is able to create cellular networks connected to the Internet. The device was used with the following configuration: Intel i7 2GHz, 8GB RAM, and Microsoft Windows 7 Ultimate 64 bits installed. It creates the cellular network and also works as video transmitter. The videos are received by a smart phone Motorola Moto X Play, connected to the Internet.

There are several different parameters to be adjusted in Anritsu's MD8475A device for each type of network (2G, 3G or LTE - Long Term Evolution) As 2G and 3G networks are too slow for this kind of application (video streaming), resulting in videos with jitter, we have focused our attention to LTE networks (4G). It is also a network in a growing adoption rate.

To create the video streaming service between Anritsu's MD8475A and the mobile phone, VLC Media Player version 2.2.4 was used [VLC 2017]. It was installed in the Anritsu's equipment and in the cell phone (in this case, an Android app) as client. RTP (Real-time Transfer Protocol) was used without transcodification for video broadcasting. To capture the video in the cell phone, we have used the AZ Screen Recorder 4.1.1 app (from Google Store). VLC captures the screen with a default rate of 30 fps and stores it in a MP4 file. The proposed scheme can be seen in Figure 8.

Table 16: PLCC for no-reference metrics on ReTRiEVED

Metric	Delay	Jitter	Throughput	PLR	All
VIIDEO	0.8955	0.1743	0.1294	0.1501	0.1871
BLIINDS	0.7376	0.6745	0.7676	0.5570	0.6743

Table 17: SROCC for no-reference metrics on ReTRiEVED

Metric	Delay	Jitter	Throughput	PLR	All
VIIDEO	-0.1928	-0.3817	-0.1392	-0.1432	-0.2563
BLIINDS	0.3361	0.7312	0.4478	0.5411	0.5954

Table 18: RMSE for no-reference metrics on ReTRiEVED

Metric	Delay	Jitter	Throughput	PLR	All
VIIDEO	0.2325	0.9340	0.8599	1.3452	1.2610
BLIINDS	0.3527	0.7003	0.8770	0.7224	0.9517

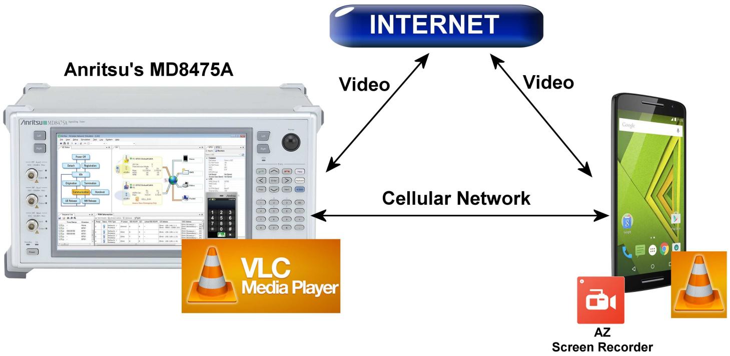


Figure 8: Scheme for the second experiment: Anritsu's MD8475A equipment simulates a cellular network that establishes a connection with the cellphone. VLC media player broadcast the video through the Internet, using the simulated cellular network from the base station simulator to the cellphone, which has also the VLC media player to receive the file. The streaming video is then captured using AZ Screen Recorder app.

There are two important observations about this experiment: the first is that, as many apps, the AZ Screen Recorder inserts the image of its main menu in the beginning of the video (with buttons of record, pause and stop). This creates a set of frames that are not related to the original video. To solve this, we have manually removed these frames from the received video and the same frames from the transmitted video, so that they have a perfect match. The second observation is that we have no information about the Mean Opinion Score (MOS) of the broadcasted videos; to generate these scores, we would have to replicate the exact same conditions that the others were produced in order to have a fair comparison. Thus, what we are trying to analyse, in this second experiment, is what is the behaviour of the video quality assessment metrics against the degradation of the channel. This means that the metrics are being presented in their real values and not as a correlation to the DMOS as presented before.

We have adjusted Anritsu's MD8475A to three different transfer rates: 5, 10 and 20 Mbps (Mega bits per second). Other parameters should have been changed as gain, but we have observed that the transfer rate is enough for what we are analysing. Figure 9 presents a sample frame of a video broadcasted with 5 Mbps and its original version.

(a) (b)

Figure 9: Cellular network experiment (sample frames): (left) frame from the original video and (right) the corresponding frame for the transmitted video transferred by a rate of 5 Mbps.

A subset of videos from ReTRiEVED dataset was used: six original videos produced 18 broadcasted videos. Some of them have severely lost their quality and were not considered in the analysis as they would not provide useful information. This was common in videos transmitted at 5 Mbps which presented strong jitter. The remained videos were analysed using PEVQ, NTIA's VQM, some MSU metrics (as seen in Table 19) and no-reference metrics (BLIINDS and VIIDEO). The results are presented in Tables 19 and 20. PEVQ results are presented in Table 21; in this case, PEVQ provides scores for MOS, Jerkiness, Blockiness, Blur, and PSNR for Y, Cb and Cr channels.

SSIM based algorithms have scores in the range between -1 and 1, where 1 happens when both images (or videos) are equal. This means that, the closer the scores are to 1, more similar are the videos. From Table 19, it is possible to see that the scores do not reflect the decrease of quality of the videos for some SSIM-based algorithms (SSIM fast, SSIM precise, 3SSIM and stSSIM);

Table 19: FR metrics average scores for streaming videos in cellular network experiment simulation. The average is taken for 5, 10 and 20 Mbps transfer rate.

Metric	5 Mbps	10 Mbps	20 Mbps
SSIM (fast)	0.2770	0.2143	0.2604
SSIM (precise)	0.3179	0.2547	0.4425
3SSIM	0.1404	0.1235	0.1352
MS-SSIM	0.0717	0.0495	0.0489
stSSIM	-6.5E-4	-0.12E-4	-7.2E-5
VQM	16.545	16.1137	15.8627

Table 20: NR metrics average scores for streaming videos in cellular network experiment simulation. The average is taken for 5, 10 and 20 Mbps transfer rate.

Metric	original video	5 Mbps	10 Mbps	20 Mbps
VIIDEO	0.094	0.1167	0.1009	0.1053
BLIINDS	1,394	1,380	1,500	1,498

just MSSIM had decreasing scores for decreasing quality. For VQM, the scores correctly represent the loss of quality. In this last case, the higher the score, the better the quality of the video in comparison to the pristine video.

The scores of VIIDEO and BLIINDS are very different in scale. Although there is a normalization scheme to put these scores in a MOS scale, we are presenting the original scores as it is very hard to find these algorithms applied directly to videos in literature. Again, our interest is to analyse what happens to the scores with the decreasing quality of the videos. In this case, for no-reference metrics, as there is no need of a reference video for comparison (a ground truth), the algorithms are applied to the distorted videos and to the original videos, aiming to create a reference value. For VIIDEO, the best videos have scores close to zero, while, for BLIINDS, the best videos have high scores. As it can be seen in Table 20, considering VIIDEO, the lowest score is for the original videos. The average score for 5 Mbps video is higher than the average score for 20 Mbps videos which should be expected. For BLIINDS, however, although the average score for the original videos is high, it is lower than the average scores for 10 and 20 Mbps videos, which are clearly lower quality videos. Thus, BLIINDS have failed in the evaluation of the videos in a real scenario.

Table 21 presents the results for Opticom's PEVQ applied to the sample videos. In the previous section, we have presented PEVQ results based on correlation or error. As said before, as we do not have the MOS of these broadcasted

Table 21: PEVQ average scores for streaming videos in cellular network experiment simulation. The average is taken for 5, 10 and 20 Mbps transfer rate.

Metric	5 Mbps	10 Mbps	20 Mbps
MOS	1	1.0525	1.01
Jerkiness	10	3.42	2.8
Blockiness	5.895	4.9175	4.704
Blur	7.96	8.04	7.082
PSNR Y	12.19	11.945	12.086
PSNR Cb	27.885	27.5925	28.194
PSNR Cr	29.285	32.5625	30.678

videos in the current experiment, it is not possible to make the same kind of comparison. Thus, we just compare the results looking for changes in the scores, as the level of degradation of the videos also changes. PEVQ evaluates: MOS, Jerkiness, Blockiness, Blur and PSNR (for each one of the color components in YCbCr color space). The mean opinion score (MOS) is a value between 1 and 5; Jerkiness, Blockiness and Blur are evaluated in a range between 0 an 10 (where 0 means the absence of that distortion); for PSNR (Peak Signal to Noise ratio), as it is known, the higher, the better.

From the Table 21, it is possible to see that, in a real scenario, PEVQ did not present a satisfactory result. Its most important feature is the MOS and the scores were too close to each other, which does not reflect the loss of quality of the videos, as can be seen in the sample frame in Figure 9. Even more, the scores are too low for high quality videos, as the ones streamed at 20 Mbps.

5 Conclusions

This paper analysed several objective video assessment metrics. They were divided into Full-Reference (FR) and No-Reference (NR) metrics and some state of the art approaches were tried. The evaluation was done in two different situations: the first was based on well-known datasets (VQEG Phase I, LIVE VQA and ReTRiEVED). For this case, the following FR metrics were used: SSIM (fast), SSIM (precise), SSIM (MatLab), GSSIM, MS-SSIM (fast), MS-SSIM (precise), stSSIM, 3SSIM, VQM (MSU), VQM (NTIA) and PEVQ. VQM (NTIA), an open source tool, achieved the best results in most part of the experiments. We highlight its performance on the ReTRiEVED dataset. This would be our recommendation for an application where a FR algorithm need to be used. The second set of experiments ran in a simulation of a cellular network. For this, we have used Anritsu's MD8475A device, which created the model of a real cellular

Table 22: Best NR metric for each experiment and each measure (PLCC, SROCC and RMSE).

Dataset (Degradation)	PLCC	SROCC	RMSE
VQEG Phase I (50Hz)	NTIA VQM	stSSIM	NTIA VQM
VQEG Phase I (60Hz)	NTIA VQM	stSSIM	NTIA VQM
LIVE VQA (Wireless)	PEVQ	PEVQ	PEVQ
LIVE VQA (IP)	MS-SSIM	MS-SSIM	MS-SSIM
LIVE VQA (H.264)	PEVQ	PEVQ	3SSIM
LIVE VQA (MPEG-2)	NTIA VQM	NTIA VQM	PEVQ
ReTRiEVED (Delay)	NTIA VQM	PEVQ	NTIA VQM
ReTRiEVED (Jitter)	NTIA VQM	NTIA VQM	NTIA VQM
ReTRiEVED (Throughput)	NTIA VQM	NTIA VQM	NTIA VQM
ReTRiEVED (PLR)	NTIA VQM	NTIA VQM	NTIA VQM

network with possibility of changing the settings. In our experiment, we have changed the transfer rate to 5, 10 and 20 Mbps. The videos were broadcasted to the Internet via the cellular network scenario and then back to the cell phone to be captured and analysed. For FR metrics, NTIA's VQM achieved the best results again. For No-Reference metrics, we have tried VIIDEO and BLIINDS for both experiments; VIIDEO performed better, with a closer representation of the changes in the degradation of the videos. NR metrics, however, are not reliable for practical applications yet. To summarize our experiments, Table 22 lists the best results for every FR metrics and every dataset. It does not worth to list what is the better application for each metric as most of them did not get closer to the best results in all experiments.

Acknowledgements

This work was supported by the research cooperation project between Motorola Mobility (a Lenovo Company) and CIn-UFPE.

References

[CDVL 2003] Consumer Digital Video Library: http://www.cdvl.org/

[EPFL 2009] EPFL database: http://vqa.como.polimi.it/

[Firsby and Stone 2010] Frisby, J.P., Stone, J.V.: "Seeing"; The MIT Press, Cambridge (2010).

[Li and Bovik 2009] Li, C., Bovik, A.C.: "Three Component Weighted Structural Similarity Index"; Proceedings of SPIE, vol. 7242 (2009).

[Mittal et al 2016] Mittal, A., Saad, M.A., Bovik, A.C.: "A Completely Blind Video Integrity Oracle"; IEEE Transactions on Image Processing, vol.25, (2016) 289-300.

[Moorthy and Bovik 2010] Moorthy, A.K., Bovik, A.C.: "Efficient Motion Weighted Spatio-Temporal Video SSIM Index"; Proceedings of SPIE-IS&T Electronic Imaging, USA (2010).

[MSU 2005] MSU Video Quality Measurement Tool: http://compression.ru/video/quality_measure

[Opticom 2017] Opticom: http://www.opticom.de/

[Pinson and Wolf 2004] Pinson, M.H., Wolf, S.: "A New Standardized Method for Objectively Measuring Video Quality"; IEEE Transactions on Broadcasting, vol. 50, (2004) 312-322.

[ReTRiEVED 2014] ReTRiEVED: http://www.comlab.uniroma3.it/retrieved.htm

[Saad et al 2014] Saad, M.A., Bovik, A.C., Charrier, C.: "Blind Prediction of Natural Video Quality"; IEEE Transactions on Image Processing, 23, (2014) 1352-1365.

[Seshadrinathan et al 2010] Seshadrinathan, K., Soundararajan, R., Bovik, A.C., Cormack, L.K.: "A subjective study to evaluate video quality assessment algorithms"; Proceedings of SPIE 7527, Human Vision and Electronic Imaging XV (2010).

[LIVE 2003] Sheikh, H.R., Wang, Z., Cormack, L., Bovik, A.C.: LIVE Image Quality Assessment Database: http://live.ece.utexas.edu/research/quality

[Seshadrinathan and Bovik 2010] Seshadrinathan, K., Bovik, A.C.: "Motion Tuned Spatio-temporal Quality Assessment of Natural Videos"; IEEE Transactions on Image Processing, vol.19, (2010) 335-350.

[Simone et al 2009] Simone, F., Naccari, M., Tagliasacchi, M., Dufaux, F., Tubaro, S., Ebrahimi, T.: "Subjective Assessment of H.264/AVC Video Sequences Transmitted over a Noisy Channel"; First International Workshop on Quality of Multimedia Experience, USA (2009).

[Simone et al 2010] Simone, F., Tagliasacchi, M., Naccari, M., Tubaro, S., Ebrahimi, T.: "H.264/AVC Video Database for the Evaluation of Quality Metrics"; 35th International Conference on Acoustics, Speech, and Signal Processing, USA (2010).

[VLC 2017] VLC: http://www.videolan.org/vlc/

[VQEG 2000] VQEG Final report from the video quality experts group on the validation of objective models of video quality assessment: http://www.vqeg.org, 2000.

[Wang and Bovik 2006] Wang, Z., Bovik, A.C.: "Modern Image Quality Assessment"; Morgan & Claypool Publishers; 1st edition (2006).

[Wang et al 2004a] Wang, Z., Liu, L., Bovik, A.C.: "Video Quality Assessment Based on Structural Distortion Measurement"; Signal Processing: Image Communication, 19, (2004) 1-9.

[Wang et al 2004b] Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: "Image Quality Assessment: From Error Measurement to Structural Similarity"; IEEE Transactions on Image Processing, 13, (2004) 1-14.

[Wang and Bovik 2002] Wang, Z., Bovik, A.C.: "A Universal Image Quality Index"; IEEE Signal Processing Letters, vol.9, (2002) 81-84.

[Wang et al 2003] Wang, Z., Simoncelli, E.P., Bovik, A.C.: "Multi-Scale Structural Similarity for Image Quality Assessment"; Proceedings of the 37th IEEE Asilomar Conference on Signals, Systems and Computers, USA, (2003) 1-5.

[Witkin 1983] Witkin, A.: "Scale-space filtering"; International Joint Conference in Artificial Intelligence, Germany, (1983) 1019-1021.