

# Improving the Prediction Accuracy of Objective Video Quality Evaluation

**Nenad Stojanović, Boban Bondžulić, Boban Pavlović,  
Marko Novčić, Dimitrije Bujaković**

Military Academy, University of Defence in Belgrade, Generala Pavla Jurišića  
Šturma 33, 11000 Belgrade, Serbia  
e-mails: nenad.m.stojanovic@vs.rs, boban.bondzulic@va.mod.gov.rs,  
boban.pavlovic@va.mod.gov.rs, marko.novcic@vs.rs,  
dimitrije.bujakovic@va.mod.gov.rs

---

*Abstract: Three different approaches for improvement of objective video quality evaluation are presented in this paper. Improvement is obtained through quality guided temporal pooling, information content weighted temporal pooling, and multiscale analysis. The analysis was performed using five objective video quality assessment measures on two publicly available datasets with subjective quality scores. Only the videos with H.264, H.265, and MPEG-2 types of compression from two datasets were considered. The level of agreement between the subjective and objective quality scores are given through the Spearman rank-order correlation coefficients on complete datasets and subsets of video sequences with the same type of compression. Obtained results show that the performance of objective measures is dependent on the choice of the dataset. The greatest improvement is given by multiscale analysis.*

*Keywords: information pooling; objective video quality assessment; temporal pooling; video compression; video resolution*

---

## 1 Introduction

In recent years there has been a rapid development of systems for digital processing, transmission and display of video content [1, 2]. This development has led to great interest in reliable, computationally efficient objective quality assessment measures. A subjective quality assessment is the most reliable way to determine the quality of video signals, but subjective tests are very expensive and time-consuming, and an alternative is sought in the form of objective quality assessment measures. There are three categories of objective quality assessment measures, No-Reference (NR), Full-Reference (FR), and Reduced-Reference (RR) [1-3]. This classification is based on the availability of the source signal on the receiving side. NR measures can be used in all applications where quality testing

is required because this type of metrics do not need knowledge of the source signal. FR metrics require full information of source signal and for that reason, this category cannot be used in some real-time applications where the knowledge of the original signal on the receiving side is not possible. RR techniques are between the two previously described categories and in these techniques only the most important part of the source signal is needed for quality evaluation. Objective image/video quality assessment measures have found numerous applications. Most applications are in situations where the quality of the modified version of the image/video needs to be evaluated.

Algorithms for video quality assessment usually have two phases. In the first, quality is evaluated on local spatial/temporal level, and in the second, spatial/temporal pooling of local scores produces a final value of quality [4]. Spatial and temporal integrations are closely related to visual significance. Estimation of visual significance identifies information on motion image which notably effect on observer during forming an impression of the quality. This allows for increasing the impact of essential information on the final score of the evaluation. Generally, strong degradation in space and/or time has a great effect on the final impression of quality. Strong distortions give low values of similarity between reference and test signals, so using the scores with the lowest quality, the final quality value can be formed. Also, the resolution of video during processing and display can have significant effect on final quality assessment.

The increasing number of video services and the increase in the resolution of the video display devices have led to the requirement for higher coding efficiency compared to the H.264 compression algorithm capabilities [5]. Therefore, a novel compression algorithm, H.265, was developed [6, 7], and a new compression standard is under progress [8, 9]. The goal of introducing the H.265/HEVC standard was to maintain subjective video quality by reducing the bit rate of 50% compared to H.264 [7].

The aim of this paper is to analyze the performance of objective quality assessment measures on sequences with MPEG-2, H.264 and H.265 compressions, using three different approaches for improving the prediction accuracy of objective video quality estimation. Objective quality assessment measures performance was analyzed on two publicly available, subject rated video datasets. H.264, H.265 and MPEG-2 compression algorithms are most commonly used algorithms in video systems, and therefore they are chosen for the analysis.

The quantitative measure adopted by the ITU [10] – Spearman's Rank Order Correlation Coefficient (SROCC) between subjective and objective quality scores, was used in the performance analysis of objective quality assessment measures.

In the second part of the paper are described used FR objective quality measures and the most important information of two datasets for video quality is provided. Three possible directions to improve the video quality estimation with results are given in the third part of the paper. The conclusions and further research directions are given at the end of the paper.

## 2 Overview of Objective Measures and Video Quality Datasets

The five objective video quality assessment measures were used in the analysis. Peak Signal to Noise Ratio (PSNR) [11], is the first measure. PSNR is an unavoidable measure in image/video quality analysis, although is often criticized [11, 12]. The Structural Similarity Index (SSIM) is the second measure, which is present in almost all tests of image/video quality measures [13].

Table 1  
Comparison of used video datasets

Video Dataset	FERIT-RTRK			CSIQ Video		
Number of original sequences	6			12		
Number of tested (distorted) sequences	H.264	30	90	H.264	36	72
	H.265	30		H.265	36	
	MPEG-2	30				
Number of degradation levels	5			3		
Degradation types	H.264, H.265, MPEG-2			H.264, H.265, MJPEG, SNOW, packet loss, AWGN		
Tested degradations	H.264, H.265, MPEG-2			H.264, H.265		
Resolution	1920x1080 pixels			832x480 pixels		
Length	5 seconds			10 seconds		
Frame rates	60 fps			24, 30, 50, 60 fps		
Number of observers	30			35		

SSIM has numerous modifications, such as GSM (average of local quality values of the gradient magnitude information preservation) and GMSD (standard deviation of local quality gradient magnitude similarity scores) [14]. GSM and GMSD are third and fourth used measures. The fifth objective video quality assessment measure is  $VQ^{AB}$  [15].  $VQ^{AB}$  is based on the analysis of the spatial information preservation (through the gradient magnitude and gradient orientation information preservation), the temporal information preservation (through the preservation of information on changes between frames) and the color information preservation.

The analysis was performed on two video quality datasets: FERIT-RTRK [16] and CSIQ Video [17]. Table 1 shows data comparison between two used video datasets. FERIT-RTRK dataset has more degradation levels and tested (degraded) sequences than the CSIQ dataset, but has less referent video sequences and shorter videos length.

### 3 Video Quality Analysis

Subject rated video quality datasets are of great importance because they can help in developing reliable objective measures. Quality guided temporal pooling, information content weighted temporal pooling and multiscale analysis are some of the approaches for improving the level of agreement between subjective and objective quality scores and performance of objective quality measures.

#### 3.1 Quality Guided Temporal Pooling

It has been shown that regions of poor image quality significantly affect on a human estimation of visual quality [18]. This fact is used for quality guided lowest percentile temporal pooling approach, where  $p\%$  ( $p$  percent) of the frames with the lowest quality scores are used. Parameter  $p$  represents a number of used frames in percent, in the step by 2%. The temporal pooling process is carried out in the following way. After determining objective quality scores on a frame-by-frame basis, their sorting is done in rising order, after which the final quality score is determined as the mean value of  $p\%$  of the lowest scores of frames quality. Values beyond this range are rejected. This approach is guided by the hypothesis that the frames with poor quality can have a dominant role in the subjective impression of quality [18]. Measure GMSD has an inverse scale, so the sorting is done in descending order, after which the final quality score is determined as the mean value of  $p\%$  of the highest scores of frames quality.

Figure 1 shows the normalized values of objective quality scores of frames for two sequences with H.265 compression. Graphics show significant quality variations during the lasting of the video. Also, from Figure 1, a periodic repetition of the local maximums of quality is observed, which is the consequence of the I frames present in the degraded sequences. In addition, it can be noticed that objective quality measures in different ways respond to changes that occur in video sequences. Thus, from the Figure 1 (b), between the 100th and 300th frame, can be noted that PSNR and SSIM objective values are increasing then decreasing, the values of GMSD and  $VQ^{AB}$  objective measures decrease, while GMSM values increase.

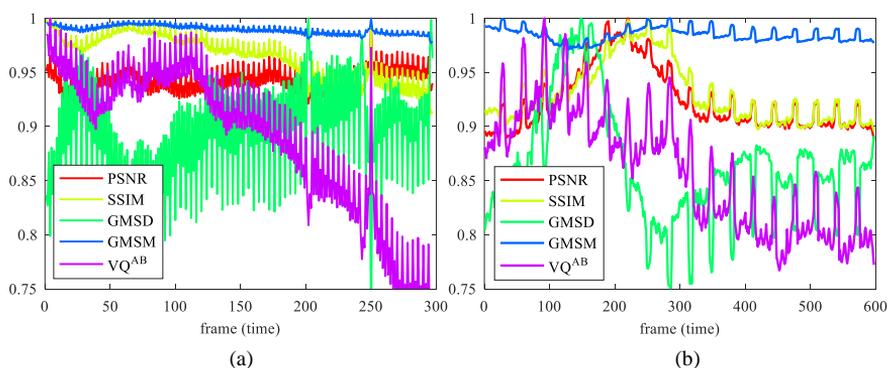


Figure 1

Objective quality scores of frames of the analyzed sequences: (a) sequence yac\_H265\_1 from FERIT-RTRK dataset and (b) sequence BQTerrace\_832x480\_dst\_18 from CSIQ dataset

Figure 2 shows the mean values of  $p\%$  of the lowest quality scores for two objective measures – PSNR and VQ<sup>AB</sup>. All test sequences from both video datasets are analyzed. From Figure 2 it is noticed that dynamic ranges of the objective values differ on analyzed datasets. Thus, the dynamic range of PSNR measure is 4 dB narrower on the FERIT-RTRK dataset than on CSIQ dataset. Furthermore, both measures have lower objective quality values on the sequences from the FERIT-RTRK dataset. This observation is also valid for other analyzed objective measures – SSIM, GMSM and GMSD (this measure has an inverse scale, so the higher quality scores are obtained on the FERIT-RTRK dataset).

Due to the content of the test videos, it can be explained why the results of the  $p\%$  of the lowest objective quality scores differ between these two datasets. Used video sequences should represent the real world images, i.e. datasets contain a wide range of content. A variety of content of a dataset can be characterized using Spatial Activity (SA), Temporal Activity (TA) and colorfulness index. In this work, the spatial complexity of video sequences, SA, was analyzed based on the mean values of the gradient magnitude of the frames. Sobel operator was used to determine the gradient magnitude. Figure 3 shows SA values per frame of all sequences in CSIQ and FERIT-RTRK datasets.

Dynamic ranges of the spatial activity of the distorted sequences on these two datasets are significantly different, which can be seen from Figure 3. The dynamic range of the spatial activity values is almost two times bigger in the CSIQ dataset than in the FERIT-RTRK dataset because video sequences from the CSIQ dataset are richer with details. This can be a consequence of the format of the delivered videos. Namely, CSIQ dataset sequences are delivered in raw format – YUV420, while all sequences (including reference) of the FERIT-RTRK dataset are delivered in the compressed format – mp4.

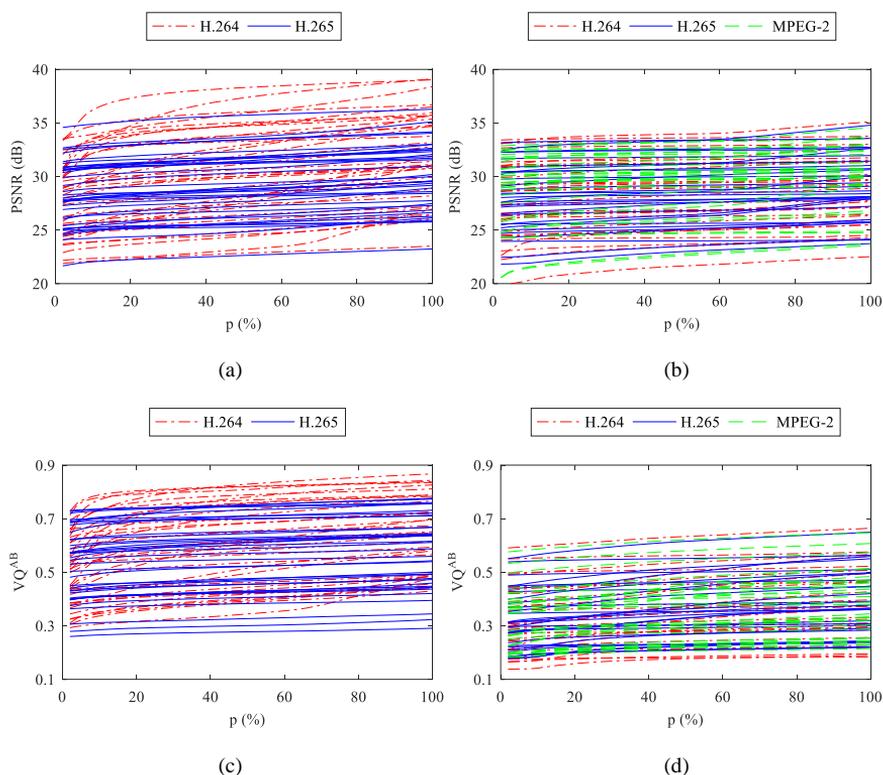


Figure 2

Mean values of the  $p\%$  of the lowest objective quality scores: (a) PSNR on CSIQ dataset, (b) PSNR on FERIT-RTRK dataset, (c)  $VQ^{AB}$  on CSIQ dataset and (d)  $VQ^{AB}$  on FERIT-RTRK dataset

A similar analysis was carried out in the analysis of the TA [15] per frame of the distorted sequences, whereby the conclusion that the dynamic ranges of the TA of the sequences from these two datasets are approximately the same.

The influence of the selection of the frames with poor quality on objective assessment was analyzed through the SROCC at the level of complete datasets and at the level of subsets of sequences with the same type of degradation. Results of the correlations with subjective quality impressions on the global level are shown in Figure 4.

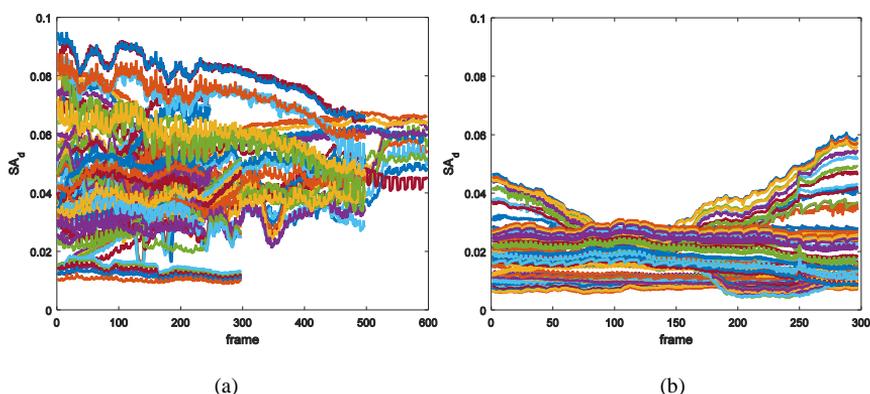


Figure 3

Spatial activity values per frame of the test (degraded) sequences on the: (a) CSIQ dataset and (b) FERIT-RTRK dataset

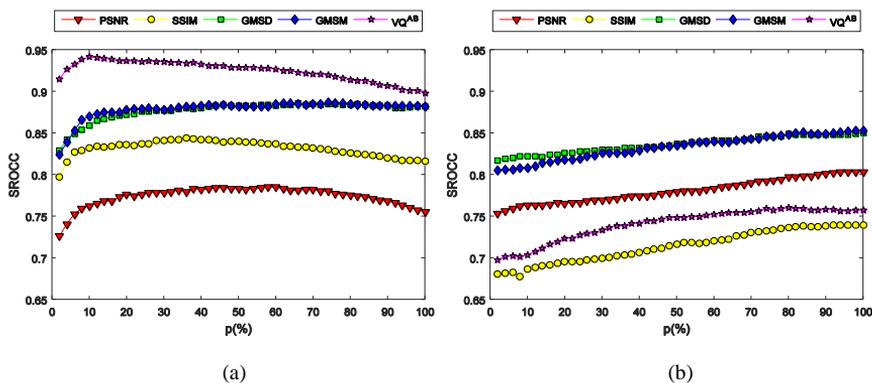


Figure 4

Rank order correlation (SROCC) between subjective and p% of the lowest objective quality scores on: (a) CSIQ dataset and (b) FERIT-RTRK dataset

From Figure 4 it is noticed different trends of SROCC values on these two analyzed datasets. Unlike the FERIT-RTRK dataset, where the best agreement of subjective and objective quality scores is achieved if all frames of the test sequences ( $p=100\%$ ) are used, in CSIQ dataset can be noted that using the lowest scores of frames quality can improve the performance of objective measures ( $VQ^{AB}$ , SSIM, and PSNR). In this dataset, the greatest gain would be obtained using 10% of the lowest scores of the  $VQ^{AB}$  objective measure (level of agreement by using all frames is 0.9, while the level of agreement by using 10% of the lowest quality scores is 0.94). Objective quality assessment measure,  $VQ^{AB}$ , has the best performance on the CSIQ dataset (in the entire range values of parameter  $p$ ). However, the performance of this measure is significantly worse on the FERIT-RTRK video dataset. The performance of all other tested objective measures is

worse on the FERIT-RTRK than on the CSIQ dataset, where the biggest performance drop is noticed at  $VQ^{AB}$  and SSIM measurements.

The performance of objective measures on subsets of videos with the same type of degradation is presented in Figure 5. From this figure, it can be noticed that on H.264 compressed videos, the performance of objective measures depends on the values  $p$ . On the other side, in a subset of sequences with H.265 compression, the performance of objective measures is almost independent of the choice of values  $p$ . The performance of objective measures on corresponding subsets of the FERIT-RTRK dataset is worse than on subsets of the CSIQ dataset.

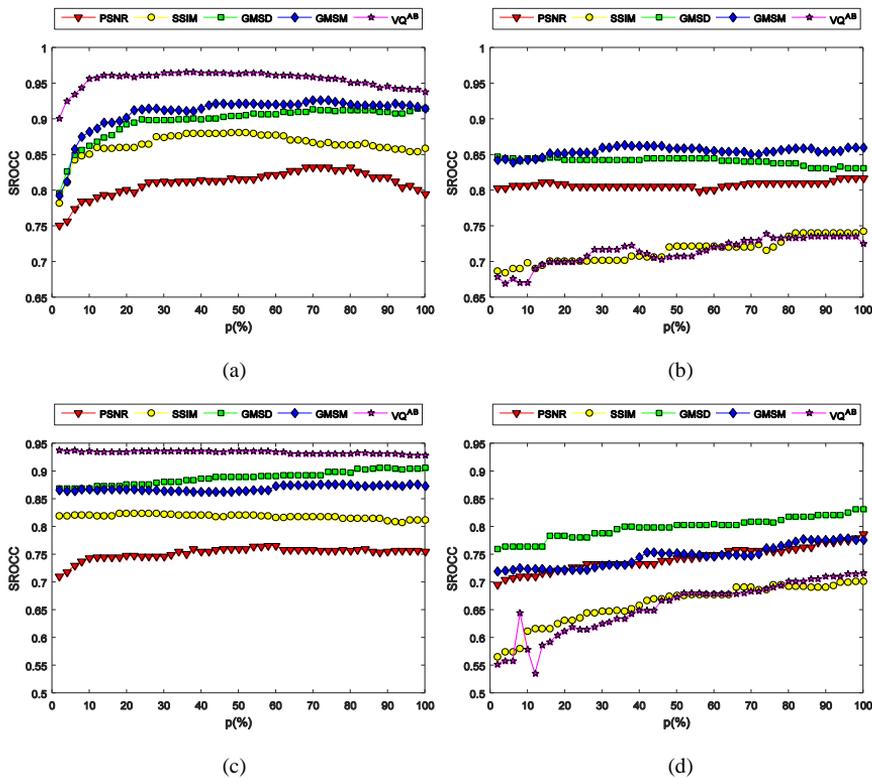


Figure 5

Rank order correlation (SROCC) between subjective and  $p\%$  of the lowest objective quality scores on subsets of video sequences: (a) H.264 subset of CSIQ dataset, (b) H.264 subset of FERIT-RTRK dataset, (c) H.265 subset of CSIQ dataset and (d) H.265 subset of FERIT-RTRK dataset

### 3.2 Information Content Weighted Temporal Pooling

The problem with non-uniform distribution of frames quality over time can be solved by assigning them a time-varying significance (weight) [19]. The general shape of a temporal weighting approach is given by:

$$Q_f = \frac{\sum_{i=1}^N w_i Q_i}{\sum_{i=1}^N w_i} \quad (1)$$

where  $w_i$  is the weight associated with the  $i$ -th temporal location (frame),  $N$  is the number of frames in the degraded/reference sequence and  $Q_i$  is the quality value at the  $i$ -th temporal location. The weights are determined by the frame information content (using reference or distorted frames or both of them).

A list of 18 weighted functions is given in Table 2. The significance associated with the estimates of the frames quality during the time is derived from the spatial activity of the reference ( $SA_r$ ) and distorted ( $SA_d$ ) video sequences, and from the temporal activity of the reference ( $TA_r$ ) and distorted ( $TA_d$ ) video sequences. Impact of the significance of the frames during the time is given through the rank order correlation between subjective and objective quality scores on a global level – on complete datasets. Spatial and temporal activities are combined in an additive and multiplicative manner or as their maximum value. In Table 2, for objective measures, with +/- are marked situations in which weighting led to an improvement/deterioration of the performance of the objective measure, while the value presents the gain/loss relative to the SROCC of the standard method of pooling the frames quality scores (averaging). All presented results are on the relation with correlation coefficients where 100% of the frames are included from the previous subchapter.

From the Table 2, it can be noticed that all measures have some improvement on the CSIQ dataset within all weighted functions, except the GMSD measure which has an improvement for only four weighted functions. In all objective measures, the use of temporal activities ( $TA_r/TA_d$ ) results in a higher gain than the use of spatial activities ( $SA_r/SA_d$ ). Also, it can be concluded that weighted functions in multiplicative form lead to a greater agreement between subjective and objective quality scores than weighted functions in additive form or weighted functions with the selection of maximum. The highest improvement on this dataset was obtained by using PSNR objective measure overall analyzed weighted functions. The greatest gain in the PSNR measure was achieved by applying multiplicative weighted functions  $TA_d \cdot TA_r$  (0.03),  $SA_r \cdot TA_r$  (0.024) and  $SA_d \cdot TA_d \cdot SA_r \cdot TA_r$  (0.022). These three weighted functions are suitable for the accuracy improvement of other objective measures.

Contrary, on the FERIT-RTRK dataset, there is no improvement except for the SSIM measure. The gain achieved for this measure on this dataset is slightly worse than the gain achieved on the CSIQ dataset. According to the achieved gain,

the multiplicative weighted functions stand out –  $SA_d \cdot TA_d \cdot SA_r \cdot TA_r$  (0.013),  $SA_d \cdot TA_d$  (0.01),  $TA_d \cdot TA_r$  (0.008) and  $SA_r \cdot TA_r$  (0.008). Objective measure SSIM is the only measure which has improvement on both used datasets and in all 18 used weighted functions.

Table 2

Improvement/deterioration of the performance of the objective measures on CSIQ and FERIT-RTRK datasets relative to the SROCC of all frames in all test sequences

Weight, $w_i$	PSNR		SSIM		GMSD		GMSM		VQ <sup>AB</sup>	
	CSIQ	FERIT	CSIQ	FERIT	CSIQ	FERIT	CSIQ	FERIT	CSIQ	FERIT
$SA_r$	+	-	+	+	-	-	+	-	+	-
	0.011	0.004	0.005	0.003	0.003	0.002	0.002	0.002	0.006	0.003
$SA_d$	+	-	+	+	-	-	0	-	+	-
	0.008	0.006	0.004	0.006	0.008	0.003	0	0.004	0.001	0.003
$TA_r$	+	-	+	+	+	-	+	-	+	-
	0.019	0.004	0.011	0.005	0.001	0.001	0.004	0.003	0.012	0.002
$TA_d$	+	-	+	+	-	-	+	-	+	-
	0.019	0.005	0.009	0.003	0.001	0.001	0.004	0.001	0.009	0.007
$SA_r \cdot TA_r$	+	-	+	+	+	-	+	-	+	-
	0.024	0.012	0.012	0.008	0.002	0.003	0.002	0.003	0.018	0.004
$SA_d \cdot TA_d$	+	-	+	+	-	-	+	-	+	-
	0.018	0.015	0.010	0.010	0.003	0.004	0.001	0.003	0.008	0.006
$\max(SA_r, TA_r)$	+	-	+	+	+	-	+	-	+	-
	0.014	0.003	0.011	0.005	0.001	0.001	0.005	0.003	0.007	0.001
$\max(SA_d, TA_d)$	+	-	+	+	-	-	+	-	+	-
	0.012	0.004	0.006	0.006	0.004	0.002	0.003	0.002	0.004	0.003
$SA_r + TA_r$	+	-	+	+	-	-	+	-	+	-
	0.015	0.003	0.009	0.004	0.001	0.002	0.005	0.002	0.009	0.002
$SA_d + TA_d$	+	-	+	+	-	0	+	-	+	-
	0.011	0.003	0.005	0.005	0.003	0	0.003	0.001	0.005	0.003
$SA_r + SA_d$	+	-	+	+	-	-	+	-	+	-
	0.009	0.004	0.005	0.004	0.006	0.002	0.001	0.002	0.002	0.002
$TA_r + TA_d$	+	-	+	+	-	-	+	-	+	-
	0.020	0.004	0.010	0.004	0.001	0.002	0.004	0.002	0.011	0.003
$SA_r + SA_d + TA_r + TA_d$	+	-	+	+	-	-	+	-	+	-
	0.015	0.004	0.008	0.005	0.002	0.002	0.005	0.001	0.007	0.002
$\max(SA_d, SA_r)$	+	-	+	+	-	-	+	-	+	-
	0.011	0.004	0.005	0.003	0.003	0.002	0.002	0.003	0.006	0.003
$\max(TA_d, TA_r)$	+	-	+	+	+	-	+	-	+	-
	0.020	0.004	0.011	0.004	0.001	0.002	0.005	0.003	0.012	0.002
$TA_d \cdot TA_r$	+	-	+	+	0	+	+	0	+	-
	0.030	0.011	0.014	0.008	0	0.005	0.002	0	0.014	0.006
$SA_d \cdot SA_r$	+	-	+	+	-	-	0	-	+	-
	0.010	0.015	0.006	0.008	0.006	0.004	0	0.006	0.006	0.005
$SA_d \cdot TA_d \cdot SA_r \cdot TA_r$	+	-	+	+	-	-	0	-	+	-
	0.022	0.022	0.019	0.013	0.008	0.006	0	0.007	0.011	0.008

### 3.3 Multiscale Analysis

In further analysis, the objective quality of video sequences in different scales (resolutions) is evaluated. The level of agreement between subjective and objective quality scores is analyzed in five scales for the FERIT-RTRK dataset, and in four scales for CSIQ video dataset, because the resolution of videos from

FERIT-RTRK dataset is greater than videos from CSIQ dataset. The decimation of the original and test sequences was made with scaling factors 1/2, 1/4, 1/8, and 1/16. In addition to decimation, bicubic interpolation was performed. Scale 2 corresponds to scaling factor 1/2, while scale 5 corresponds to scaling factor 1/16 [20-22].

The values of the  $VQ^{AB}$  measure are obtained by averaging of the lowest 20% frames quality scores in this analysis, as suggested in [15]. For all other measures, 100% of the frames are used.

Figure 6 shows the dependence of the level of agreement between subjective and objective quality scores (SROCC) over different scales (resolutions) with both used datasets. Correlation is calculated on complete datasets. From Figure 6 it is noted that the performance of objective measures significantly depends on the scale in which the original and compressed video was compared. The choice of the optimal scale depends on the objective measure, too. In this way, the observation from [20] that the assessment of image/video quality in different resolutions provides more flexibility in incorporating the variations of viewing conditions was confirmed.

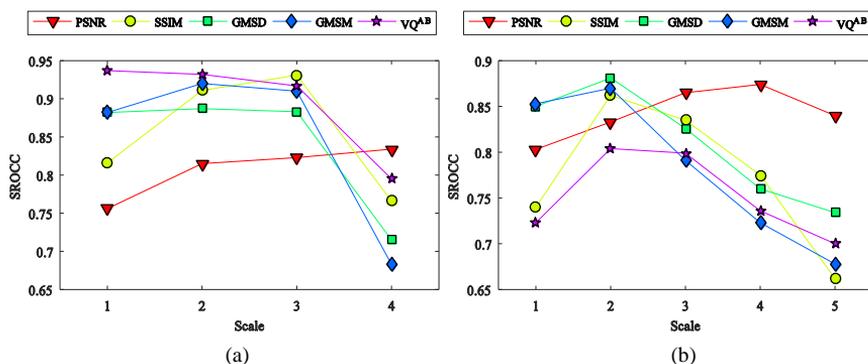


Figure 6

Rank order correlation (SROCC) between subjective and objective quality scores in different scales on:  
(a) CSIQ dataset and (b) FERIT-RTRK dataset

The highest degree of agreement for CSIQ dataset is between subjective and  $VQ^{AB}$  objective quality scores (original resolution,  $SROCC=0.936$ ). Applying the GMSD objective measure provides the highest degree of agreement of quality scores on the FERIT-RTRK dataset (scale 2,  $SROCC=0.881$ ).

In both video datasets, it is noticeable the performance improvement of the SSIM objective quality assessment measure, comparing with the original resolution; the comparison for the FERIT-RTRK dataset is carried out in scale 2 (Figure 6 (b) shows the SSIM's SROCC jump from 0.741 to 0.862), and for the CSIQ video dataset in scale 3 (Figure 6 (a) shows the SSIM's SROCC jump from 0.816 to 0.931).

Observing the results of the PSNR measure, a higher correlation with subjective scores is obtained by analyzing lower resolutions, while other objective measures have better performance in higher resolutions. In this way, it can be concluded that in the comparison of the signals in higher scales (lower resolution), the analysis of the energy preservation of the signal is important, while in the lower scales (higher resolution), the analysis of the preservation of the signal structure is more important. Furthermore, on FERIT-RTRK dataset it can be seen that the degree of agreement between subjective and PSNR objective quality scores in scale 4 is close to the results of the best GMSD measure (0.874 vs. 0.881).

## Conclusions

Three approaches for improving the performance of objective quality assessment measures are presented in the paper. The presented approaches are quality guided temporal pooling, information content-weighted temporal pooling, and multiscale analysis. The five objective video quality assessment measures and two publicly available video datasets with H.264, H.265, and MPEG-2 compressed video contents are used in the analysis. It has been shown that the performance of objective measures significantly depends on the choice of the dataset, which makes it necessary to use more reference video datasets in video quality analyzes. Since these datasets contain a relatively small number of test signals with H.264 and H.265 compressed contents (60+72 sequences), it can be concluded that there is a need for new datasets, which will contain a significantly larger number of compressed test signals.

In addition, from the analysis it can be concluded that the greatest potential for improving the performance of objective measures on both datasets has the multiscale approach, where the improvement depends on the choice of an objective measure. By applying this approach, the improvement of accuracy prediction achieved through the correlation of ranks was up to 0.12 (SSIM objective measure on both analyzed datasets). Quality guided temporal pooling, implemented through the use of the lowest quality scores, on the CSIQ dataset has led to the improvement of the performance of objective measures (rank correlation increased by up to 0.05), while on the FERIT-RTRK dataset the performance with such integration is in the level of performance without pooling. Information content-weighted temporal pooling does not give significant improvement (rank correlation increased by up to 0.03 on CSIQ dataset), and in this case, except for SSIM, information content-weighted temporal pooling did not lead to an improvement in the results of objective measures on the FERIT-RTRK dataset. The lack of improvement in the results of objective measures using temporal pooling on a FERIT-RTRK dataset is probably due to the format of the delivered sequences – the original and test sequences were delivered in a compressed mp4 format.

As these three approaches were analyzed separately, in further work we will analyze their combined effect in objective video quality assessment.

## Acknowledgment

This research has been a part of the project No. VA-TT/3/20-22 supported by the Ministry of Defence, Republic of Serbia.

## References

- [1] Bovik, A. C.: Automatic prediction of perceptual image and video quality, *Proceedings of the IEEE*, 2013, Vol. 101, No. 9, pp. 2008-2024
- [2] Baig, M. A., Moinuddin, A. A., Khan, E., Ghanbari, M.: Image fidelity estimation from received embedded bitstream, *Signal, Image and Video Processing*, 2020, Vol. 14, No. 3, pp. 465-472
- [3] Yu, X., Bampis, C. G., Gupta, P., Bovik, A. C.: Predicting the quality of images compressed after distortion in two steps, *IEEE Transactions on Image Processing*, 2019, Vol. 28, No. 12, pp. 5757-5770
- [4] Bampis, C. G., Li, Z., Moorthy, A. K., Katsavounidis, I., Aaron, A., Bovik, A. C.: Study of temporal effects on subjective video quality of experience, *IEEE Transactions on Image Processing*, 2017, Vol. 26, No. 11, pp. 5217-5231
- [5] Miličević, Z., Bojković, Z., Rao, K. R.: HEVC vs. H.264/AVC standard approach to coder's performance evaluation, *WSEAS Transactions on Signal Processing*, 2015, Vol. 11, pp. 272-279
- [6] Sullivan, G. J., Ohm, J.-R., Han, W.-J., Wiegand, T.: Overview of the high efficiency video coding (HEVC) standard, *IEEE Transactions on Circuits and Systems for Video Technology*, 2012, Vol. 22, No. 12, pp. 1649-1668
- [7] Tan, T. K., Weerakkody, R., Mrak, M., Ramzan, N., Baroncini, V., Ohm, J.-R., Sullivan, G. J.: Video quality evaluation methodology and verification testing of HEVC compression performance, *IEEE Transactions on Circuits and Systems for Video Technology*, 2016, Vol. 26, No. 1, pp. 76-90
- [8] Sullivan, G. J.: Video coding standards progress report: Joint video experts team launches the versatile video coding project, *SMPTE Motion Imaging Journal*, 2018, Vol. 127, No. 8, pp. 94-98
- [9] Takamura, S.: Versatile video coding: A next-generation video coding standard, *NTT Technical Review*, 2019, Vol. 17, No. 6, pp. 49-52
- [10] ITU Tutorial: Objective perceptual assessment of video quality – Full reference television, Geneva, ITU, 2004
- [11] Wang, Z., Bovik, A. C.: Mean squared error: Love it or leave it? A new look at signal fidelity measures, *IEEE Signal Processing Magazine*, 2009, Vol. 26, No. 1, pp. 98-117
- [12] Jakšić, B., Gara, B., Petrović, M., Spalević, P., Lazić, Lj.: Analysis of the impact of front and back light on image compression with SPIHT method

- during realization of the chroma key effect in virtual TV studio, *Acta Polytechnica Hungarica*, 2015, Vol. 12, No. 2, pp. 71-88
- [13] Wang, Z., Bovik, A. C., Sheikh, H. R., Simoncelli, E. P.: Image quality assessment: From error visibility to structural similarity, *IEEE Transactions on Image Processing*, 2004, Vol. 13, No. 4, pp. 600-612
- [14] Xue, W., Zhang, L., Mou, X., Bovik, A. C.: Gradient magnitude similarity deviation: A highly efficient perceptual image quality index, *IEEE Transactions on Image Processing*, 2014, Vol. 23, No. 2, pp. 684-695
- [15] Petrović, V., Bondžulić, B.: Objective assessment of surveillance video quality, *Sensor Signal Processing for Defence*, London, UK, IET, 2012, pp. 1-5
- [16] Bajčinovci, V., Vranješ, M., Babić, D., Kovačević, B.: Subjective and objective quality assessment of MPEG-2, H.264 and H.265 video, *International Symposium ELMAR, Zadar, Croatia, IEEE*, 2017, pp. 73-77
- [17] Vu, P. V., Chandler, D. M.: ViS3: An algorithm for video quality assessment via analysis of spatial and spatiotemporal slices, *Journal of Electronic Imaging*, 2014, Vol. 23, No. 1, pp. 1-24
- [18] Moorthy, A. K., Bovik, A. C.: Visual importance pooling for image quality assessment, *IEEE Journal of Selected Topics in Signal Processing*, 2009, Vol. 3, No. 2, pp. 193-201
- [19] Duanmu, Z., Ma, K., Wang, Z.: Quality-of-experience for adaptive streaming videos: An expectation confirmation theory motivated approach, *IEEE Transactions on Image Processing*, 2018, Vol. 27, No. 12, pp. 6135-6146
- [20] Wang, Z., Simoncelli, E. P., Bovik, A. C.: Multi-scale structural similarity for image quality assessment, *The Thirty-Seventh Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, USA, IEEE, 2003, pp. 1398-1402
- [21] Bondžulić, B., Petrović, V., Mitrović, S., Pavlović, B., Andrić, M.: Visual attention pooling and understanding the structural similarity index in multi-scale analysis, *Optica Applicata*, 2014, Vol. 44, No. 2, pp. 267-283
- [22] Pavlović, B., Bondžulić, B., Stojanović, N., Novčić, M., Bujaković, D.: Comments on objective quality assessment of MPEG-2, H.264 and H.265 videos, *New Trends in Signal Processing, Demanovska Dolina, Slovakia, IEEE*, 2018, pp. 156-163