PEAK SIGNAL-TO-NOISE RATIO REVISITED: IS SIMPLE BEAUTIFUL?

Jari Korhonen

Technical Univ. of Denmark (DTU)
Dept. of Photonics Engineering
Kgs. Lyngby, Denmark

Junyong You

Norwegian Univ. of Science and Tech. (NTNU)
Dept. of Electronics and Telecommunications
Trondheim, Norway

ABSTRACT

Heavy criticism has been directed against using peak signal-to-noise ratio (PSNR) as a full reference quality metric for digitally processed images and video, since many studies have shown a weak correlation between subjective quality scores and the respective PSNR values. In this paper, we show that the low performance of PSNR is often related to a content dependent systematic shift of PSNR values. In scenarios with fixed content and distortion types that are typical for visual communications applications, PSNR may perform closely as well, or in some cases even better than the more complex objective quality models known from the literature. Therefore, the use of PSNR may be justified for comparative quality assessment with fixed content.

1. INTRODUCTION

The goal of image and video compression is to obtain the best possible visual quality in respect to the size of the compressed image file or bitrate of the encoded video sequence. In communications applications, video quality may be additionally distorted by transmission errors, such as packet losses. In order to compare the performance of different compression schemes and error resilience techniques, reliable methods for assessing video quality are essential. Unfortunately, quality is a subjective characteristic that cannot be measured directly as technical parameters, such as bitrate or packet loss rate.

To date, vast majority of research studies on digital image and video processing use PSNR as a metric for visual quality. The greatest advantage of PSNR is its simplicity: it can be derived directly from the mean squared error (MSE) between the pixels in the original and the distorted image or video sequence. Unfortunately, many studies have shown that PSNR often has a weak correlation with the mean opinion scores (MOS) obtained from subjective quality assessment [1,2]. PSNR is especially vulnerable to distortion types causing misalignment of pixels, such as spatial shift, rotation or resizing [1]. To overcome the limitations of PSNR, significant efforts have been invested on development of more reliable quality models exploiting our knowledge of the human visual system (HVS). However, objective image and video quality assessment has turned out to be a surprisingly challenging task, and results on methods statistically outperforming PSNR were not presented before the second phase evaluation study by video quality experts group (VQEG) in 2003.

Even though there are more accurate quality metrics available today, PSNR is still widely used. It has been reported that PSNR estimates the relative perceptual quality reasonably well when fixed content and codec is concerned across different test cases [3]. On the other hand, comparison of PSNR values obtained from different content types is misleading, since the range of meaningful PSNR values varies heavily between contents. However, few studies have been conducted to explore the meaningful scope of use for PSNR. In this paper, we use data from well-known annotated video quality databases to compare the performance of PSNR against other quality models for each original source sequence separately.

2. QUALITY METRIC PERFORMANCE ANALYSIS

Different quality models are typically evaluated by using root mean squared error (RMSE), Spearman rank order correlation coefficient (SROCC), Pearson correlation coefficient (PCC) and outlier ratio between the quality estimates produced by the model and the MOS values obtained from subjective assessment. Since different metrics operate with different score scales that are not necessarily linear, nonlinear regression is typically performed to find mapping between the objective video quality rating *VQR* and predicted MOS_p values. Different functions can be used for this purpose. In this paper, we use the logistic function with three parameters a_1, a_2 and a_3, as suggested by VQEG:

$$MOS_P = a_1 / (1 + \exp(-a_2 \cdot (VQR - a_3))) \quad (1)$$

There are several image and video databases accompanied with MOS data that can be used for validation purpose. For this study, we have identified two databases with a reasonable mix of different contents and relevant distortion types including source and channel artifacts: LIVE video quality database [2] and EPFL-PoliMi database [4].

Typically, the studies comparing different quality models operate with integrated data sets covering all the sequences with different contents and distortion types. In that type of studies, PSNR usually performs inferior to other more advanced metrics. However, when the analysis is performed separately for different contents, the situation is essentially different. The data points for each content tend to form clusters and significantly better correlation between predicted and actual MOS can be obtained within each content separately than with the entire database. We did not observe similar clusters with VQM and MOVIE.

In Table I, we show PCC, SROCC, and RMSE computed for the whole LIVE database and each content separately. To compensate bias from direct averaging of correlation coefficients, the average values for PCC and SROCC have been computed after applying Fisher's Z-transformation to individual coefficients, and then the inverse transformation to the resulting average. In Table II, the respective results with EPFL-PoliMi database are shown. Due to the lack of space, results for individual contents are omitted. The results with both databases show that PSNR performs poorly with the entire database, but with individual contents the average performance of PSNR is on a par with VQM and MOVIE. This observation supports the hypothesis that the weak performance of PSNR reported in many studies can be mostly explained by different content dependent shift of PSNR values. The LIVE database contains several different types of compression and channel artifacts, and therefore our results are assumed valid when distortions typically observed in communications applications are concerned. However, it should be noted that the results cannot be generalized to all types of distortions.

3. CONCLUSIONS

In this paper, we have used two different annotated video quality databases to demonstrate that in scenarios with fixed content distorted by typical compression and channel artifacts, PSNR predicts the perceived subjective quality nearly as well as more complex quality models representing the state-of-the-art. Our results suggest that PSNR may be used as a quality metric for video, assuming that the analysis is performed separately for each content (source sequence), or the content based bias in PSNR values is otherwise taken into account. Methods for analyzing and compensating the content bias will be explored in the future work.

4. REFERENCES

[1] Z. Wang, and A. C. Bovik, “Mean Squared Error: Love It or Leave It? A New Look at Signal Fidelity Measures,” *IEEE Signal Proc. Magazine,* vol. 26, no. 9, pp. 98-117, Jan. 2009.

[2] K. Seshadrinathan, R. Soundararajan, A. C. Bovik and L. K. Cormack, "Study of Subjective and Objective Quality Assessment of Video", *IEEE Trans. on Image Proc.*, vol. 19, no.6, pp.1427-1441, Jun. 2010.

[3] Q. Hyunh-Thu, and M. Ghanbari, “Scope of Validity of PSNR in Image/Video Quality Assessment,” *Electronics Letters*, vol. 44, no. 13, pp. 800-801, Jun. 2008.

[4] F. De Simone, M. Naccari, M. Taggliasacchi, F. Dufaux, S. Tubaro, and T. Ebrahimi, "Subjective Quality Assessment of H.264/AVC Video Streaming with Packet Losses," *EURASIP J. on Image and Video Processing*, 2011.

Table I. Performance comparison per original source sequence between PSNR, VQM and MOVIE, using LIVE database.

	PSNR			VQM			MOVIE		
Source sequence	**PCC**	**SROCC**	**RMSE**	**PCC**	**SROCC**	**RMSE**	**PCC**	**SROCC**	**RMSE**
a. Pedestrian	**0.914**	0.871	**4.37**	0.843	**0.879**	5.80	0.837	0.821	5.90
b. Riverbed	**0.947**	**0.942**	**2.81**	0.862	0.853	4.41	0.756	0.710	5.70
c. Rush Hour	0.922	0.896	3.98	0.828	0.846	5.78	**0.936**	**0.911**	**3.63**
d. Tractor	0.971	**0.971**	2.85	0.905	0.861	5.03	**0.974**	0.964	**2.71**
e. Station	**0.793**	**0.675**	**5.01**	0.692	0.661	5.93	0.708	0.486	5.81
f. Sunflower	0.662	0.693	6.18	0.599	0.625	6.60	**0.689**	**0.729**	**5.97**
g. Blue Sky	**0.873**	0.818	4.96	0.925	**0.961**	**3.88**	0.871	0.864	5.01
h. Shield	0.878	0.850	5.87	0.777	0.793	7.72	**0.910**	**0.889**	**5.10**
i. Park Run	0.891	0.857	5.50	**0.940**	**0.929**	**4.13**	0.935	0.889	4.30
j. Mobile Cal.	0.814	0.804	5.90	0.758	0.804	6.61	**0.931**	**0.950**	**3.72**
Average a-j	**0.890**	**0.864**	**4.74**	0.838	0.848	5.59	0.886	0.862	4.78
All contents	0.560	0.540	9.09	0.686	0.663	7.98	**0.810**	**0.789**	**6.44**

Table II. Performance comparison between PSNR, VQM and MOVIE using EPFL-PoliMi database.

	Average of all contents separately				All contents integrated			
Metric	**PCC**	**SROCC**	**RMSE**	**Outlier**	**PCC**	**SROCC**	**RMSE**	**Outlier**
PSNR	0.978	0.983	0.29	0.08	0.795	0.800	0.84	0.37
VQM	**0.982**	**0.986**	**0.26**	**0.06**	**0.957**	**0.953**	**0.41**	**0.16**
MOVIE	0.971	0.984	0.34	0.11	0.921	0.920	0.55	0.25