|  |  |
| --- | --- |
| **Joint Collaborative Team on 3D Video Coding Extension Development**  **of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11**  2nd Meeting: Shanghai, CN, 13–19 Oct. 2012 | Document: JCT3V-B0180 |

|  |  |  |  |
| --- | --- | --- | --- |
| *Title:* | **AHG9: Accuracy of 2D metrics for 3D video quality assessment** | | |
| *Status:* | Input Document | | |
| *Purpose:* | For discussion | | |
| *Author(s) or Contact(s):* | Philippe Hanhart  Touradj Ebrahimi | Email: | philippe.hanhart@epfl.ch |
| *Source:* | Ecole Polytechnique Fédérale de Lausanne (EPFL), COST Action IC1003 - QUALINET | | |

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Abstract

In this contribution, the resolving power and classification errors of PSNR, SSIM, MS-SSIM, VIFp, and VQM are investigated following the methodology described in [ITU-T Recommendation J.149](#J149). A set of subjective data, which has been collected during the formal evaluation of the 3DVC proposals on the 2-view configuration, is used as ground truth. It is reported that MS-SSIM has a higher resolving power than the other considered metrics. Therefore, it is suggested to consider MS-SSIM as an alternative to PSNR in the Core Experiments.

# Introduction

In [M24807](#M24807), the correlation between different state-of-the-art 2D quality metrics and the perceived quality of stereo pairs formed from a decoded view and a synthesized view was investigated. It was reported that some of the considered metrics, such as VIFp, VQM, MS-SSIM, and SSIM, are more correlated with perceived quality than PSNR when the objective quality assessment is based on the synthesized view. In [JCT3V-A0150](#JCT3VA0150), a similar methodology was used to assess the performance of state-of-the-art 2D quality metrics on stereo pairs formed from two synthesized views. It was reported that the metrics identified in [M24807](#M24807) are more correlated with perceived quality than PSNR for quality assessment of synthesized views. In this contribution, the performance of SSIM, MS-SSIM, VIFp, and VQM in predicting subjective quality is further investigated following the methodology described in [ITU-T Recommendation J.149](#J149). This methodology defines a new accuracy measure, called resolving power, to evaluate and compare the performance of different metrics. As in the previous contributions, a set of subjective data, collected during the formal evaluation of the 3DVC proposals on the 2-view configuration, is used as ground truth. It is reported that MS-SSIM has a higher resolving power than the other considered metrics.

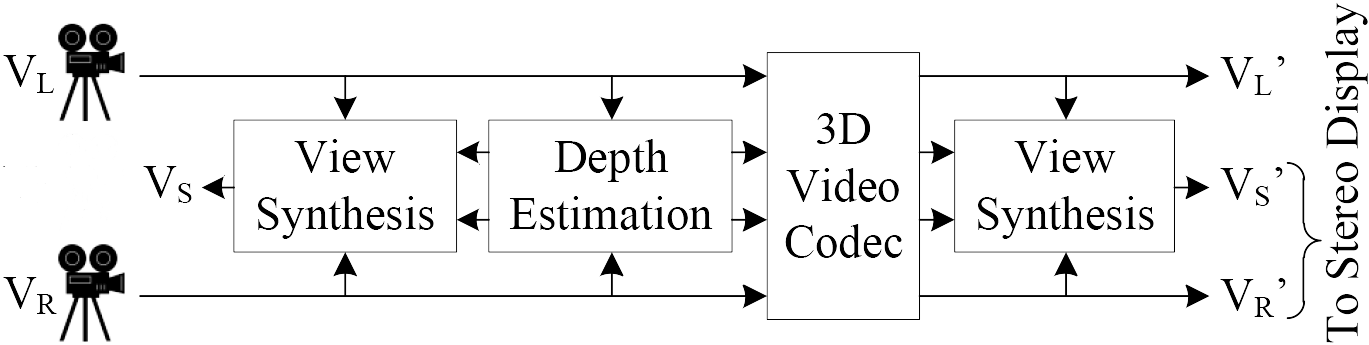
# Objective quality assessment

The accuracy of the following objective metrics (OM) is assessed:

1. PSNR: Peak Signal to Noise Ratio,
2. SSIM: Structural Similarity Index ([Wang *et al.*, 2004](#Wang2004)),
3. MS-SSIM: Multi-Scale Structural Similarity Index ([Wang *et al.*, 2003](#Wang2003)),
4. VIFp: Visual Information Fidelity ([Sheikh and Bovik, 2006](#Sheikh2006)), pixel domain version,
5. VQM: Video Quality Metric ([ITU-T Recommendation J.144, 2004](#J144)), NTIA General Model, no calibration.

All above metrics, except for VQM, were designed for quality assessment of still images. Thus, the metrics are computed on each frame and the resulting values are averaged across the frames to produce a global index for the entire video sequence.

In the 2-view configuration, as considered in the 3DVC Call for Proposals (CfP) ([N12036](#N12036)), a pair of cameras is used to produce the input views at the encoder side, as depicted in Figure 1. At the decoder side, the displayed stereo pair is formed from the decoded right view and a synthesized view, located in-between the input views.



*Figure 1 – Advanced stereoscopic processing with 2-view configuration*

In every objective metric, three different objective video quality models (VQRs) are considered:

1. Quality of the decoded view, calculated between the decoded view and the original view
2. Quality of the synthesized view, calculated between the synthesized view at the decoder side and the synthesized view at the encoder side
3. Average quality of the decoded view and the synthesized view, computed as the mean value of a) and b)

# Metric accuracy

In [(ITU-T Recommendation J.149, 2004)](#J149), the resolving power of a metric *OM* is defined as the difference in the metric values, , above which the conditional subjective-score distributions have means that are statistically different at a certain confidence or significance or probability, , level. The algorithm uses a one-tailed Z-test to determine the probability that, given a pair of objective scores, *OMi* and *OMj*, the greater score corresponds to the greater true underlying subjective mean opinion score (MOS). The details of the algorithm are described in [(ITU-T Recommendation J.149, 2004)](#J149).

The resolving power can be computed either in the native scale of the objective metric or in a transformed scale, common to different metrics. This common scale allows an easier comparison between metrics that have different ranges. A common scale can be obtained by fitting the objective scores to the subjective data.

By plotting the probability as a function of , one can determine the resolving power of each metric corresponding to a specific significance level (typically 0.68, 0.75, 0.90, or 0.95). By stacking curves corresponding to different metrics on the same graph, one can determine which metric reaches the highest significance level for a fixed resolving power (or vice versa, which metric has the smallest resolving power, i.e., the highest discriminability, for a fixed significance level).

In this contribution, as in the previous contributions ([M24807](#M24807) and [JCT3V-A0150](#JCT3VA0150)), a linear least squares regression is fitted between the objective and subjective scores for each sequence.

Then, the resolving power in the common scale, , is computed for each content separately using the Matlab script provided in [(ITU-T Recommendation J.149, 2004)](#J149). Finally, the p- curves have been built to compare the different metrics.

Since some metrics, such as PSNR, are very content dependent, the algorithm was run on each content separately and the results were averaged across the different contents. As the algorithm considers 19 bins of , intermediate values were interpolated using a linear interpolation.

# Metric classification errors

[(ITU-T Recommendation J.149, 2004)](#J149) also suggests another way to evaluate the performance of an objective metric: the classification errors. A classification error is made when the objective metric and subjective test lead to different conclusions on a pair of videos, A and B, for example. Three types of error can happen:

1. False Tie, the least offensive error, which occurs when the subjective test says that A and B are different while the objective scores say that they are identical,
2. False Differentiation, which occurs when the subjective test says that A and B are identical while the objective scores say that they are different,
3. False Ranking, the most offensive error, which occurs when the subjective test says that A is better than B while the objective scores say the opposite.

**Table 1: Classification errors**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | **Subjective** | | |
|  |  | **A > B** | **A = B** | **A < B** |
| **Objective** | **A > B** | Correct Decision | False Differentiation | False Ranking |
| **A = B** | False Tie | Correct Decision | False Tie |
| **A < B** | False Ranking | False Differentiation | Correct Decision |

The classifications errors are also computed using the Matlab script provided in [(ITU-T Recommendation J.149, 2004)](#J149). To determine whether two distributions of subjective scores are different or not, the one-tailed Z-test is used with a 95% confidence level. The relative frequencies of Correct Decision, False Tie, False Differentiation, and False Ranking are recorded from all possible distinct pairs of videos as a function of .

As increases, more pairs of data points are considered as equivalent by the objective metric. This reduces the occurrences of False Differentiations and False Rankings, but increases the occurrence of False Ties. As tends towards 0, the occurrence of False Tie will tend towards 0 and the occurrence of False Differentiation will tend towards the proportion of pairs of videos that were declared equivalent by the subjective test.

The relative frequencies are plot as a function of the . Ideally, the occurrence of Correct Decision should be maximized and the occurrence of False Ranking should be minimized when the tends towards 0. The occurrences of False Differentiations and False Rankings should decrease as fast as possible as the increases. Based on this, different graphs corresponding to different metrics can be compared to determine the best metric for the application under analysis.

Since some metrics, such as PSNR, are very content dependent, classification error rates were computed for each content separately and then averaged.

# Results and discussions

The *p* versus curves, as defined in Section 3, are reported in Figure 2 for each video quality model separately.

When the objective quality assessment is based on the measured quality of the decoded view, MS-SSIM shows a clear gain in terms of resolving power over the other metrics. For a fixed , the confidence than is better than is in average 3.11% higher for MS-SSIM.

If both views are taken into account, MS-SSIM shows an average significance level increase of 5.44% over PSNR.

When considering the video quality model based on the quality of the synthesized view, MS-SSIM shows, in general, better accuracy than the other metrics. Theoretically, the confidence than is better than should increase as the increases (monotonically increasing function). However, this is not the case for PSNR with this video quality model. The explanation for this behavior is the following: for some contents (Undo Dancer, Kendo, and Balloons), a few sequences have a low value for the PSNR of the synthesized view while the corresponding stereo pair has a high MOS ([Hanhart *et al.*, 2012](#Hanhart2012)). These data points are significantly distant from the trend in the scatter plot of the synthesized view. It is known that one proponent used a different view synthesis algorithm. Our hypothesis is that those results are from this specific proponent. This indicates that PSNR is not a trustfully indicator of quality in this case.

|  |  |
| --- | --- |
|  |  |
|  | |

*Figure 2 – Confidence versus*

The metric classification errors, as defined in Section 4, are reported in Figure 3 to Figure 5 for each metric and video quality model separately. Even though the results are reported in the native scale of the metric instead of the common scale, it is still possible to compare the classification errors of the different metrics by looking at the relative ratio ( divided by the maximum value of ) rather than the absolute .

When considering the video quality model based on the quality of the decoded view, a fast decaying False Ranking rate is observed for MS-SSIM while it is significantly slower for VQM. Similarly, the peak of the Correct Decision rate is reached at a lower ratio for MS-SSIM than VQM, meaning that MS-SSIM has a higher resolving power. The highest Correct Decision rate is obtained with MS-SSIM (0.7839) while SSIM has the lowest peak (0.7632).

|  |  |
| --- | --- |
|  | |
|  |  |
|  |  |

*Figure 3 – Frequencies of classification error: decoded view*

|  |  |
| --- | --- |
|  | |
|  |  |
|  |  |

*Figure 4 – Frequencies of classification error: synthesized view*

|  |  |
| --- | --- |
|  | |
|  |  |
|  |  |

*Figure 5 – Frequencies of classification error: decoded and synthesized views*

When the objective quality assessment is based on the measured quality of the synthesized view, PSNR shows a significantly lower Correct Decision rate and significantly higher False Ranking rate than the other metrics. Even for a 1dB difference in PSNR values, the False Ranking rate is around 7.8%. With this video quality model, MS-SSIM also shows better performance over the other metrics in terms of decaying False Ranking rate. However, the peak of Correct Decision rate is higher for VQM (0.7506) than for MS-SSIM (0.7262). Nevertheless, VQM has a significantly slower decaying False Ranking rate.

If both views are taken into account, the observations are similar to that of the video quality model based on the quality of the decoded view.

In many applications, the resolving power is not considered and even a small PSNR increase, such as 0.1dB, is considered as an improvement in perceived quality. Therefore, the Correct Decision and False Ranking values for a of zero are reported in Table 2.

SSIM, MS-SSIM, VIFp, and VQM show similar performance, with non-significant variations. However, PSNR shows a significantly lower reliability when considering the video quality model based on the quality of the synthesized view.

Table : Classification errors for

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Decoded | | Synthesized | | Decoded and synthesized | |
|  | Correct  Decision | False  Ranking | Correct  Decision | False  Ranking | Correct  Decision | False  Ranking |
| PSNR | 0.7300 | 0.0357 | 0.6331 | 0.1326 | 0.7094 | 0.0564 |
| SSIM | 0.7267 | 0.0391 | 0.6954 | 0.0703 | 0.7204 | 0.0453 |
| MS-SSIM | 0.7322 | 0.0335 | 0.6990 | 0.0667 | 0.7251 | 0.0406 |
| VIFp | 0.7311 | 0.0346 | 0.7061 | 0.0596 | 0.7270 | 0.0387 |
| VQM | 0.7260 | 0.0397 | 0.7096 | 0.0561 | 0.7231 | 0.0427 |

# Conclusion

In this contribution, the resolving power and classification errors of PSNR, SSIM, MS-SSIM, VIFp, and VQM were investigated following the methodology described in [ITU-T Recommendation J.149](#J149). It is reported that MS-SSIM has a higher resolving power than the other considered metrics. Therefore, it is suggested to consider MS-SSIM as an alternative to PSNR in the Core Experiments.

# Acknowledgement

We would like to thank Panos Nasiopoulos and Mahsa T. Pourazad from University of British Columbia (UBC), as well as Kjell Brunnström, Kun Wang, and Börje Andrén from Acreo AB for providing their subjective results for the 2-view case scenario.

This work was partially supported by the COST IC1003 European Network on Quality of Experience in Multimedia Systems and Services QUALINET and Swiss SER project Quality of Experience in 3DTV.

# References

P. Hanhart, F. De Simone and T. Ebrahimi. “Quality Assessment of Asymmetric Stereo Pair Formed From Decoded and Synthesized Views”. 4th International Workshop on Quality of Multimedia Experience (QoMEX), Yarra Valley, Australia, 2012.

ISO/IEC JTC1/SC29/WG11, "3DV: Alternative metrics to PSNR," Doc. M24807, Geneva, Switzerland, May 2012.

ISO/IEC JTC1/SC29/WG11, "Call Proposals on 3D Video Compression Technology", Doc. N12036, Geneva, Switzerland, March 2011.

ITU-T Recommendation J.144, “Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference,” ITU-T Telecommunication Standardization Bureau, March 2004.

ITU-T Recommendation J.149, “Method for specifying accuracy and cross-calibration of Video Quality Metrics (VQM),” ITU-T Telecommunication Standardization Bureau, March 2004.

JCT-3V, "3DV: Quality assessment of stereo pairs formed from two synthesized views," JCT3V-A0150, JCT-3V Meeting, Stockholm, Sweden, July 2012.

H.R. Sheikh and A.C. Bovik, “Image information and visual quality,” IEEE Transactions on Image Processing, vol. 15, no. 2, pp. 430-444, February 2006.

Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, April 2004.

Z. Wang, E.P. Simoncelli, and A.C. Bovik, “Multiscale structural similarity for image quality assessment,” in IEEE Asilomar Conference on Signals, Systems and Computers, November 2003, vol. 2, pp. 1398-1402.

# Patent rights declaration(s)

**EPFL does not have any current or pending patent rights relating to the technology described in this contribution.**