|  |  |
| --- | --- |
| **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**  1st Meeting: Stockholm, SE, 16–20 July 2012 | Document: JCT2-A0150 |

|  |  |  |  |
| --- | --- | --- | --- |
| *Title:* | **3DV: Quality assessment of stereo pairs formed from two synthesized views** | | |
| *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 objective quality assessment of stereo pairs formed from two synthesized views is investigated. State-of-the-art 2D quality metrics are benchmarked using as ground truth a set of subjective data, which has been collected during the formal evaluation test campaign of the proposals on 3D Video Compression Technology. It is reported that some of the considered metrics, such as VIFp, VQM, MS-SSIM, and SSIM, are more correlated with perceived quality than PSNR for the scenario under analysis.

# Introduction

In [M23908](#M23908), the PSNR-based quality assessment of stereo pairs formed from a decoded view and a synthesized view having different visual quality, as in the 2-view configuration, was investigated. It was reported that the PSNR of the decoded view had the highest correlation with perceived quality, while the PSNR of the synthesized view had a significantly lower correlation. In [M24807](#M24807), the performance of state-of-the-art 2D quality metrics, other than PSNR, was assessed following a similar procedure as in [M23908](#M23908). 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 this contribution, a similar methodology as in [M24807](#M24807) is used to assess the performance of state-of-the-art 2D quality metrics on stereo pairs formed from two synthesized views, as in the stereoscopic evaluations of the 3-view configuration. It is reported that the metrics identified in our previous contribution are more correlated with perceived quality than PSNR for quality assessment of synthesized views.

# Objective quality assessment

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. At the decoder side, the displayed stereo pair is formed from the decoded right view and a synthesized view, located in-between the decoded views, as specified in Table 1. In the 3-view configuration, as considered in the CfP, three cameras are used to produce the input views at the encoder side, as depicted in Figure 1. The 3-view configuration was evaluated on auto-stereoscopic multiview display and stereoscopic display. In the latter case, the displayed stereo pair is formed from two synthesized views, as specified in Table 1. For this configuration, a fixed stereo pair was defined in the CfP. This stereo pair is centered on the central decoded view and has a baseline equals to half the baseline of the stereo pair from the 2-view configuration. The test coordinator randomly selected a second stereo pair for each Class C content ([N12347](#N12347)). This random stereo pair has the same baseline as the fixed stereo pair but is located in-between two decoded views rather than centered on the central decoded view. Both stereo pairs were evaluated during the evaluations of the proposals.

OSX:Users:philippe:Desktop:3-view.pdf

*Figure 1 – Stereoscopic and auto-stereoscopic output with 3-view configuration*

Table : Input views and displayed stereo pairs

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Seq. ID** | **Test Sequence** | **Test Class** | **2-view configuration** | | **3-view configuration** | | |
| **Input views** | **Stereo pair** | **Input views** | **Fixed stereo pair** | **Random stereo pair** |
| S01 | Poznan\_Hall2 | A | 7-6 | 6.5-6 | 7-6-5 | 6.125-5.875 | - |
| S02 | Poznan\_Street | 4-3 | 3.5-3 | 5-4-3 | 4.125-3.875 | - |
| S03 | Undo\_Dancer | 2-5 | 3-5 | 1-5-9 | 4.5-5.5 | - |
| S04 | GT\_Fly | 5-2 | 4-2 | 9-5-1 | 5.5-4.5 | - |
| S05 | Kendo | C | 3-5 | 4-5 | 1-3-5 | 2.75-3.25 | 2.25-2.75 |
| S06 | Balloons | 3-5 | 4-5 | 1-3-5 | 2.75-3.25 | 4.375-4.875 |
| S07 | Lovebird1 | 6-8 | 7-8 | 4-6-8 | 5.75-6.25 | 4.0833-4.5833 |
| S08 | Newspaper | 4-6 | 5-6 | 2-4-6 | 3.75-4.25 | 4.3333-4.8333 |

Similarly to [M24807](#M24807), the performance 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. VSNR: Visual Signal to Noise Ratio ([Chandler and Hemami, 2007](#Chandler2007)),
5. VIFp: Visual Information Fidelity ([Sheikh and Bovik, 2006](#Sheikh2006)), pixel domain version,
6. WSNR: Weighted Signal to Noise Ratio[[1]](#footnote-1) ([Damera-Venkata *et al.*, 2000](#Damera2000)),
7. PSNR-HVS: PSNR Human Visual System ([Egiazarian *et al.*, 2006](#Egiazarian2006)),
8. PSNR-HVS-M: PSNR Human Visual System Masking ([Ponomarenko *et al.*, 2007](#Ponomarenko2007)),
9. VQM: Video Quality Metric ([ITU-T Recommendation J.144, 2004](#J144)) (NTIA General Model, no calibration).

All above metrics, except for VQM, are computed on the luma component of each frame and the resulting values are averaged across the frames to produce a global index for the entire video sequence.

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

1. Quality of the left view, calculated between the synthesized view at the decoder side and the synthesized view at the encoder side:
2. Quality of the right view, calculated between the synthesized view at the decoder side and the synthesized view at the encoder side:
3. Average quality of both views, computed as the mean value of a) and b)

The following properties of the VQR estimation of Mean Opinion Scores (MOS) are considered: accuracy, monotonicity and consistency. A linear least squares regression is fitted for each sequence to each [VQR, MOS] data set:

The Pearson linear correlation coefficient (PCC) and the root-mean-square error (RMSE) are used between and to estimate the accuracy of the VQR. To estimate monotonicity and consistency, the Spearman rank order correlation coefficient (SCC) is used between and . The Pearson linear correlation coefficient, Spearman rank order correlation coefficient, and root-mean-square error are finally averaged across the different sequences.

# Results and discussions

The accuracy, monotonicity, and consistency indexes of the objective video quality models, as defined in Section 2, are reported for each metric separately in Table 2 and Table 3 for the fixed and random stereo pairs, respectively. The metrics are ranked for each objective video quality model and the ranking number is specified below each performance index value.

The fixed stereo pair is centered on the central decoded view and both views are equidistant from the central decoded view. Thus, both views should have the same amount of disocclusion and the same strength of view synthesis artifacts. The random stereo pair is located in-between two decoded views; one view of the stereo pair is always located closer to one of the decoded views than the other view of the stereo pair. Thus, we denote them as closer and farther views rather than left and right views. For example, for content S05, view 2.75 is the closer view while view 2.25 is the farther view. The closer view has a lower amount of disocclusion than the farther view. Thus, the closer view should contain less view synthesis artifacts than the farther view. However, there is no significant difference between the results for the closer and farther views. In general, the objective video quality model based on the average quality of both views has the highest correlation with perceived quality, but the difference between the models is not significant.

In [M24807](#M24807), PSNR, PSNR-HVS, PSNR-HVS-M, WSNR, and VSNR had significantly lower correlation with perceived quality than VIFp, VQM, SSIM, and MS-SSIM when using the synthesized view. Two hypotheses were raised to explain our observations:

1. In terms of perceived quality, the higher quality of the decoded view, which does not contain view synthesis artifacts, tends to mask the lower quality of the synthesized view
2. Most of the considered objective metrics do not predict well perceived quality of synthesized views

In this contribution, a lower performance is reported for PSNR, PSNR-HVS, PSNR-HVS-M, WSNR, and VSNR, than for VIFp, VQM, SSIM, and MS-SSIM, especially for the random stereo pair. In this case, PSNR (PCC=0.7122, SCC=0.7415) has a significantly lower correlation with perceived quality than VIFp (PCC=0.9434, SCC=0.9511). Similar results are obtained for the fixed stereo pair.

The above-mentioned hypotheses were raised in [M24807](#M24807) because, for some metrics, the model based on the quality of the decoded view had a significantly higher correlation with perceived quality than the model based on the quality of the synthesized view. Based on the results reported in this contribution, it is suggested to reject hypothesis a) and accept hypothesis b). This conclusion is in line with the results from ([Bosc *et al.*, 2011](#Bosc2011)).

Table : 3-view configuration, fixed stereo pair

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Pearson linear correlation coefficient (PCC) | | | Spearman rank order correlation coefficient (SCC) | | | Root-mean-square error (RMSE) | | |
|  | Left view | Right view | Average | Left view | Right view | Average | Left view | Right view | Average |
| PSNR | 0.7891 | 0.8084 | **0.8086** | 0.7957 | 0.8095 | **0.8096** | 1.3581 | 1.3053 | **1.3015** |
| 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 |
| SSIM | 0.9189 | 0.9205 | **0.9215** | 0.9295 | 0.9311 | **0.9324** | 0.8857 | 0.8769 | **0.8721** |
| 3 | 4 | 3 | 4 | 5 | 4 | 3 | 4 | 4 |
| MS-SSIM | **0.9074** | 0.9046 | 0.9073 | 0.9374 | 0.9359 | **0.9388** | **0.9429** | 0.9574 | 0.9449 |
| 4 | 5 | 5 | 1 | 2 | 1 | 5 | 5 | 5 |
| VSNR | 0.9050 | **0.9281** | 0.9267 | 0.9168 | **0.9339** | 0.9324 | 0.9313 | **0.8274** | 0.8399 |
| 5 | 1 | 1 | 5 | 3 | 5 | 4 | 1 | 2 |
| VIFp | 0.9214 | 0.9241 | **0.9245** | 0.9366 | 0.9362 | **0.9382** | 0.8563 | **0.8376** | 0.8377 |
| 1 | 2 | 2 | 2 | 1 | 2 | 1 | 2 | 1 |
| WSNR | 0.8373 | **0.8587** | 0.8586 | 0.8386 | 0.8526 | **0.8536** | 1.2087 | **1.1310** | 1.1318 |
| 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| PSNR-HVS | 0.7995 | 0.8190 | **0.8190** | 0.8038 | 0.8167 | **0.8179** | 1.3304 | 1.2746 | **1.2725** |
| 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 | 8 |
| PSNR-HVS-M | 0.8016 | 0.8208 | **0.8210** | 0.8043 | 0.8175 | **0.8187** | 1.3274 | 1.2711 | **1.2689** |
| 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 | 7 |
| VQM | 0.9196 | **0.9210** | 0.9208 | 0.9335 | 0.9318 | **0.9337** | 0.8613 | **0.8528** | 0.8542 |
| 2 | 3 | 4 | 3 | 4 | 3 | 2 | 3 | 3 |

Table : 3-view configuration, random stereo pair

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Pearson linear correlation coefficient (PCC) | | | Spearman rank order correlation coefficient (SCC) | | | Root-mean-square error (RMSE) | | |
|  | Closer view | Farther view | Average | Closer view | Farther view | Average | Closer view | Farther view | Average |
| PSNR | 0.7077 | 0.7082 | **0.7122** | 0.7390 | 0.7400 | **0.7415** | 1.5903 | 1.6041 | **1.5880** |
| 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 | 9 |
| SSIM | 0.9307 | **0.9404** | 0.9384 | 0.9338 | **0.9452** | 0.9427 | 0.8452 | **0.7949** | 0.8056 |
| 3 | 2 | 2 | 4 | 2 | 3 | 3 | 2 | 2 |
| MS-SSIM | 0.9092 | 0.9050 | **0.9099** | 0.9338 | 0.9326 | **0.9369** | 0.9711 | 0.9945 | **0.9702** |
| 4 | 4 | 4 | 3 | 4 | 4 | 4 | 4 | 4 |
| VSNR | 0.8368 | 0.8514 | **0.8517** | 0.8495 | 0.8419 | **0.8569** | 1.1637 | 1.1674 | **1.1436** |
| 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 5 |
| VIFp | 0.9373 | 0.9425 | **0.9434** | 0.9442 | 0.9500 | **0.9511** | 0.8098 | 0.7727 | **0.7693** |
| 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 |
| WSNR | 0.7569 | 0.7587 | **0.7633** | 0.7735 | 0.7652 | **0.7784** | 1.4721 | 1.4777 | **1.4609** |
| 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 | 6 |
| PSNR-HVS | 0.7216 | 0.7216 | **0.7265** | 0.7442 | 0.7452 | **0.7480** | 1.5599 | 1.5754 | **1.5564** |
| 8 | 8 | 8 | 8 | 7 | 8 | 8 | 8 | 8 |
| PSNR-HVS-M | 0.7256 | 0.7262 | **0.7309** | 0.7456 | 0.7452 | **0.7497** | 1.5542 | 1.5663 | **1.5484** |
| 7 | 7 | 7 | 7 | 8 | 7 | 7 | 7 | 7 |
| VQM | 0.9314 | 0.9294 | **0.9324** | **0.9466** | 0.9392 | 0.9453 | 0.8364 | 0.8448 | **0.8279** |
| 2 | 3 | 3 | 1 | 3 | 2 | 2 | 3 | 3 |

# Conclusion

It is recommended that alternative metrics to PSNR, such as VIFp, VQM, MS-SSIM, or SSIM, also be used in the Core Experiments, where the performance of the proposed coding tools are assessed using the objective metric on a set of synthesized views, as it is reported that, for a stereo pair formed from two synthesized views, the above alternative metrics are more correlated with perceived quality than PSNR.

# Acknowledgement

This work was partially supported by the COST IC1003 European Network on Quality of Experience in Multimedia Systems and Services - QUALINET.

# References

ISO/IEC JTC1/SC29/WG11, "3DV: Objective quality measurement for the 2-view case scenario," Doc. M23908, San Jose, USA, February 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.

ISO/IEC JTC1/SC29/WG11, "Report of Subjective Test Results from the Call for Proposals on 3D Video Coding Technology", Doc. N12347, Geneva, Switzerland, November 2011.

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.

D.M. Chandler and S.S. Hemami, “VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images,” IEEE Transactions on Image Processing, vol. 16, no. 9, pp. 2284-2298, September 2007.

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.

N. Damera-Venkata, T.D. Kite, W.S. Geisler, B.L. Evans, and A.C. Bovik, “Image quality assessment based on a degradation model,” IEEE Transactions on Image Processing, vol. 9, no. 4, pp. 636-650, April 2000.

K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, and M. Carli, “New full-reference quality metrics based on HVS,” in Proceedings of the Second International Workshop on Video Processing and Quality Metrics, 2006.

N. Ponomarenko, F. Silvestri, K. Egiazarian, M. Carli,J.Astola, and V. Lukin, “On between-coefficient contrast masking of DCT basis functions,” in Proceedings of the Third International Workshop on Video Processing and Quality Metrics for Consumer Electronics, January 2007.

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.

E. Bosc, M. Köppel, R. Pépion, M. Pressigout, L. Morin, P. Ndjiki-Nya, and P. Le Callet, "Can 3D synthesized views be reliably assessed through usual subjective and objective evaluation protocols?," in International Conference on Image Processing, 2011, pp. 2597-2600.

# Patent rights declaration(s)

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

1. This metric should not be confused with the weighted sum of the PSNR of the luma and chroma components. [↑](#footnote-ref-1)