

Stereoscopic quality assessment of 1,000 VR180 videos using 8 metrics

Sergey Lavrushkin, Ivan Molodetskikh, Konstantin Kozhemyakov, Dmitriy Vatolin;
Lomonosov Moscow State University, Moscow, Russian Federation

Abstract

In this work we present a large-scale analysis of stereoscopic quality for 1,000 VR180 YouTube videos. VR180 is a new S3D format for VR devices which stores the view for only a single hemisphere. Instead of a multi-camera rig, this format requires just two cameras with fisheye lenses similar to conventional 3D-shooting, resulting in cost reduction of the final device and simplification of the shooting process. But as in the conventional stereoscopic format, VR180 videos suffer from stereoscopy-related problems specific to 3D shooting. In this paper we analyze videos to detect the most common stereoscopic artifacts using objective quality metrics, including color, sharpness and geometry mismatch between views and more. Our study depicts the current state of S3D technical quality of VR180 videos and reveals its overall poor condition, as most of the analyzed videos exhibit at least one of the stereoscopic artifacts, which shows a necessity for stereoscopic quality control in modern VR180 shooting.

Introduction

VR180 is a stereoscopic virtual-reality video format that Google introduced in 2018 [1]. The main difference relative to conventional 360-degree or spherical video is that the new format uses only one hemisphere to store a view. In typical 360-degree video, viewers can look in any horizontal direction, but the action is usually in one particular direction. In this case, however, the display device receives the entire stream, leading to transmission and storage of redundant information. Therefore, the vast majority of cases have no need to implement a view over the entire sphere — one hemisphere is sufficient to achieve the same viewer immersion.

The VR180 specification has additional benefits over conventional spherical video. First, VR180 videos are much easier to shoot than 360-degree videos. Creating spherical videos usually involves a special rig with multiple cameras that simultaneously film at different angles around the viewing point, causing a range of problems including the large size of the captured video, potential failure or overheating of some cameras, and unstable focus. The situation becomes even more complicated for stereoscopic spherical videos, which is why nearly all spherical videos are 2D. VR180, however, only requires two cameras with fisheye lenses, supporting stereoscopy by default. It also considerably reduces the cost of the filming apparatus and simplifies the recording process, because all conventional-camera methods remain applicable. As a result, this format is accessible to a wider group of people beyond just professionals. Furthermore, it eliminates the need for stitching — a problem that remains unsolved for 360-degree videos, leading to visual artifacts where images from two cameras merge. All of these benefits suggest VR180 has a promising

future.

As with conventional stereoscopic videos, however, VR180 videos can exhibit distortions between two stereoscopic views—also called stereoscopic artifacts. Shooting S3D video is more complicated than shooting traditional 2D video because it requires control of additional technical parameters. For example, color and sharpness mismatches as well as geometry distortions regularly occur when capturing S3D, potentially causing viewers discomfort, even including headaches, and thus dissuading some users from using VR180. In this paper, we assess the state of the new format's stereoscopic quality by conducting a large-scale analysis of VR180 videos from YouTube. We identify major objective-quality trends for several video characteristics. The study provides a reference point for further VR180-video analysis and demonstrates the need for quality control in these types of videos.

Related Work

All work on evaluating stereoscopic-video quality is divisible into two categories: objective quality assessment and subjective quality assessment, both of which consider viewer discomfort. Only a few works on discomfort evaluation analyzed stereoscopic distortions [2–4]; most examined asynchronous coding errors in stereoscopic video, data loss in one view during transmission, the effect of scene depth budget, and the speed of object movement [5–7]. These methods, however, targeted prediction of the discomfort viewers experienced while watching stereoscopic videos and did not assess objective distortions.

A number of proposed methods attempt to perform objective quality assessment by evaluating stereoscopic distortions: color mismatch [8, 9], sharpness mismatch [10, 11], geometry distortions [9, 12] and channel mismatch [13, 14]. Most of these methods, however, have only been tested on a few stereoscopic frames and have not been applied in practice to stereoscopic movies.

There are only a few objective-quality metrics for spherical images or videos. Most assessment methods are based on standard image-quality metrics for full images or for image patches [15]. The situation is similar for stereoscopic 360-degree content. Current metrics for S3D images or videos apply to predetermined patches in the views and are then aggregated to a single score for the full image [16, 17]. To our knowledge, no metrics apply specifically to VR180 videos, nor have any studies addressed the problem of measuring their stereoscopic quality.

This paper employs an approach that adapts our stereoscopic-quality-assessment methods to VR180. We previously employed these methods to analyze full-length stereoscopic movies [18, 19], proving their practicality.



Figure 1: Preprocessing of VR180 frames. The red square on the right highlights the front edge of the cube-map projection.

Method

Our proposed method for evaluating VR180 video comprises three main steps:

- VR180-video preprocessing;
- Stereo matching with confidence estimation;
- Metric estimation.

The following sections describe each step.

VR180-Video Preprocessing

All VR180 video frames initially appear in an equirectangular projection. The preprocessing step remaps them into a cube-map projection (Figure 1). Because the original video has a maximum 180-degree field of view, the cubemap-projection edges contain information from the initial frame as follows: side, top and bottom — only half, front — the whole, and the back is not filled at all. Then we select only the front edge for further processing. This edge contains the most information about the frame, it lacks any areas obstructed by black borders, and it has the mildest geometric distortions. All of these factors make applying stereoscopic metrics and getting valid results easier. Furthermore, we process the views in the same way we process conventional stereoscopic frames.

Stereo Matching With Confidence Estimation

We compute disparity maps between stereoscopic views using fast local block matching [20]. Since the result can contain errors, we construct corresponding confidence maps based on the LRC criterion [21] and block RGB variance. Since the left and right views display the same scene, the disparity values of the left-view pixels should be equal in magnitude but opposite in sign compared with the disparity values of the corresponding right-view pixels. More formally, we calculate the LRC criterion as follows. If a pixel with coordinates $x = (x_1, x_2)$ in one view corresponds to a pixel with coordinates $x' = (x'_1, x'_2) = x + v_x$ in the other view, its confidence measure is:

$$\text{lrc} = \frac{\text{dif}_1^2}{h} + \frac{\text{dif}_2^2}{w}, \quad (1)$$

$$\text{dif} = (\text{dif}_1, \text{dif}_2) = v'_x + v_x, \quad (2)$$

where v_x is the disparity vector of a pixel with coordinates x in the first view; v'_x is the disparity vector of a pixel with coordinates x' in the second view; h is the view height and w is the view width.

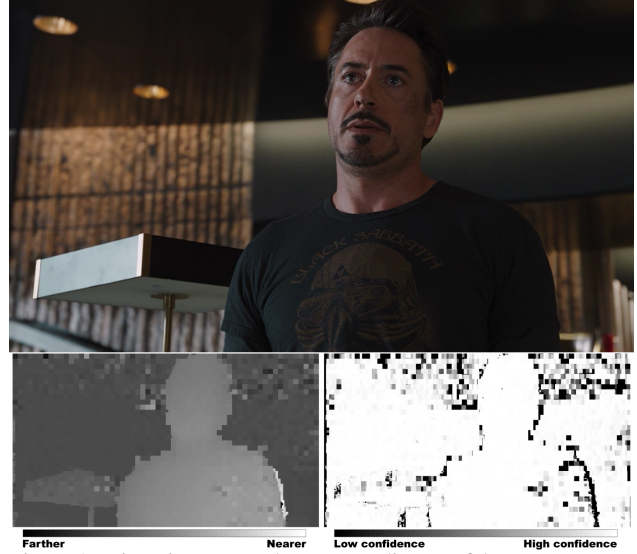


Figure 2: Disparity map and corresponding confidence map computed for the right view of a stereopair. The scene is from *The Avengers*.

The block variance for each block in the corresponding view is the sum of the variances for each RGB color component in the block:

$$\text{var} = \text{var}^R + \text{var}^G + \text{var}^B, \quad (3)$$

$$\text{var}^i = \frac{1}{s} \sum_{p \in \text{block}} p_i^2 - \left(\frac{1}{s} \sum_{p \in \text{block}} p_i \right)^2, \quad (4)$$

where p is a pixel from the image block and i is an RGB color channel.

The final confidence value is the following, taking into account the two characteristics described above:

$$\text{conf}_i = \min(1 - a \times \text{lrc}_i, b \times \text{var}_i), \quad (5)$$

where a and b are positive real coefficients and i is the pixel index. An example of a computed disparity map and corresponding confidence map appears in Figure 2.

Metrics Estimation

To analyze the technical quality of VR180 videos, we estimate the following stereoscopic parameters:

1. Positive parallax;
2. Negative parallax;
3. Color mismatch;
4. Vertical parallax;
5. Rotation mismatch;
6. Scale mismatch;
7. Sharpness mismatch;
8. Channel mismatch.

Our estimation of these parameters employs the following metrics.

We use a modified version of the metric described in [22] to find extreme positive and negative disparity values. The main difference is the calculation of a weighted histogram of disparity

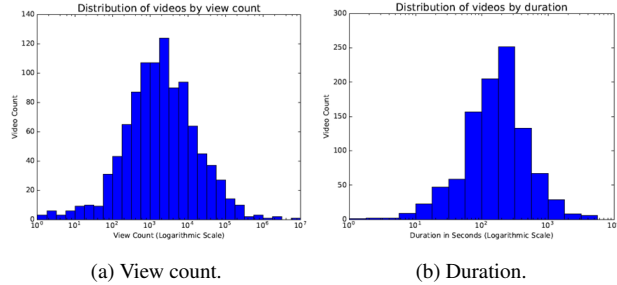


Figure 3: VR180-video statistics.

values using corresponding confidence values. We then use the histogram we constructed to estimate positive and negative parallaxes.

To reduce the impact of unwanted outliers (in occlusions, for example), we calculate color mismatch as the sum of absolute differences between the left and right views, interpolated to the left view using the computed disparity map, weighted by the disparity-map confidence and then normalized by the sum of the confidence-map values:

$$\frac{\sum_{i \in [h \times w]} \text{conf}_i * (|L_i^r - IR_i^r| + |L_i^g - IR_i^g| + |L_i^b - IR_i^b|)}{\sum_{i \in [h \times w]} \text{conf}_i}, \quad (6)$$

where L and IR are the left view and the interpolated right view, respectively, in the RGB color space. Additionally, before calculating the metric, we apply a weighted median filter [23] to each color channel and to the corresponding confidence maps to further reduce the influence of undesirable regions on the final metric value. This filter allows us to preserve object contours during image processing, considerably increasing the metric's accuracy.

To estimate geometry-distortion parameters we follow an approach similar to [18], calculating rotation, scale and vertical shift on the basis of a simplified affine-transform model. Our calculation of the sharpness difference between stereoscopic views employs the algorithm described in [19] and detection of the channel mismatch — method in [24].

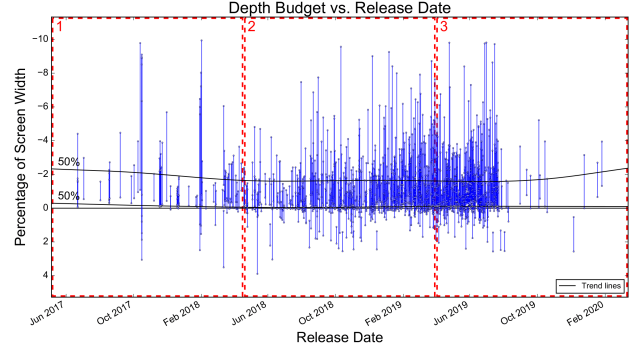
VR180 Analysis Results

VR180-Video Dataset

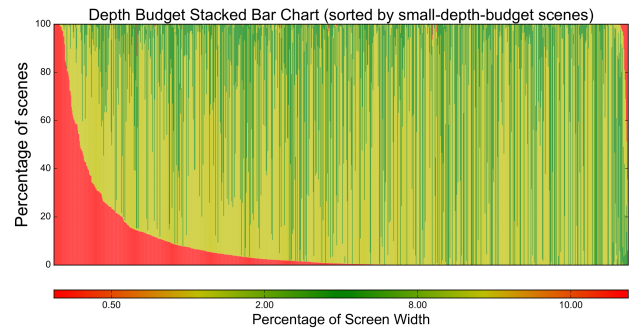
To conduct a large-scale VR180-video analysis, we collected 1,000 videos from YouTube. To reduce bias in our collection, we conducted 36 searches in total: 26 English letters one by one and 10 digits one by one. We set the filter to VR180 videos and collected the first 5 to 10 pages of results together with their statistics. We excluded videos that were unavailable, low in resolution or nonstereoscopic. Figure 3 presents the video distribution by number of YouTube views and by duration (in seconds). The x -axis in both charts is logarithmic. Most of the videos have 10,000 to 100,000 views, but a few have several million. Most are 5 to 10 minutes long. We finished collecting videos early in 2020 and subsequently focused on processing and evaluation them.

Depth-Budget Analysis

Figure 4a shows the average disparities of the closest and farthest objects in the videos we evaluated. Positive values correspond to objects behind the screen plane, while negative values



(a) Overall results relative to video release date.



(b) Depth budget of scenes for each video.

Figure 4: Results of VR180-video depth-budget analysis.

correspond to objects in front of the screen plane. The figure represents individual videos using lines that go from their highest positive parallax to their lowest negative parallax, measured in percentage of screen width. These lines correspond approximately to the range between objects farthest away from the viewer and objects closest to the viewer. The longer the line, the greater the video's depth budget. Figure 4a shows that some videos have a tiny depth budget, whereas others have an enormous one. It also includes trend lines for positive and negative disparities, revealing the average depth budget among the videos. Those videos with a larger-than-average depth budget are likely to be uncomfortable to watch on some VR headsets — a problematic result, because depth budget is difficult to fix in postproduction. Additionally, several videos have significant positive parallax. Because the left and right views are already in front of the viewer's corresponding eyes in a VR headset, a zero parallax (0%) corresponds to an object's depth being “at infinity” (by comparison, zero parallax in S3D means the object appears at the screen plane). Thus, VR180 videos should have minimal positive parallax, because a positive value means the object is “beyond infinity” — an unrealistic situation for the viewer's brain. The red rectangles in Figure 4a depict areas that we magnify in the final technical report for closer inspection.

Figure 4b shows breakdowns of scene depth budget for each video. The x -axis represents the different videos, and the y -axis uses bars to represent how many shots in each video have a good, average or bad depth budget. For videos on the far left, nearly all shots have a small depth budget. Videos in the middle have average depth budgets, but the depth budget in many scenes is still a little too low or high. A few videos on the right mostly con-

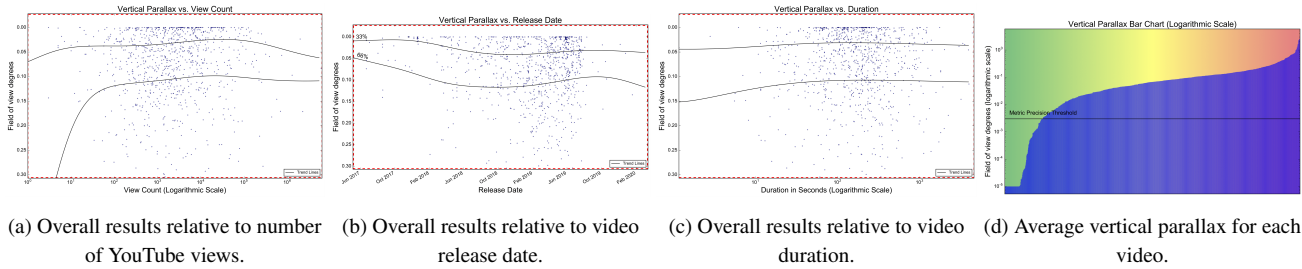


Figure 5: Results of vertical-parallax analysis for VR180 videos.

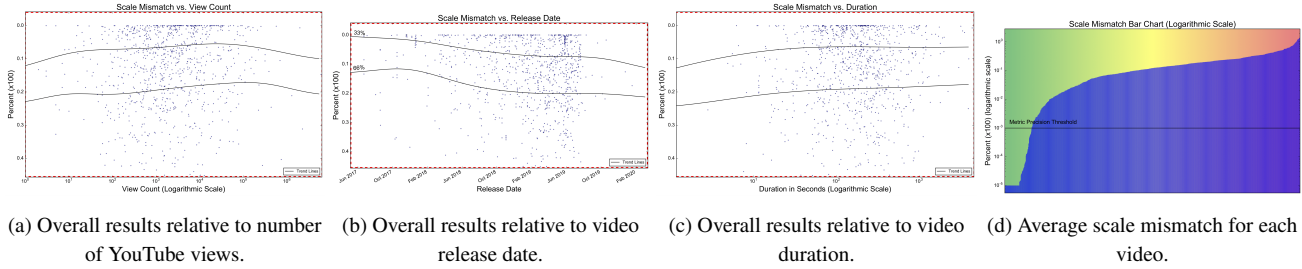


Figure 6: Results of scale-mismatch analysis for VR180 videos.

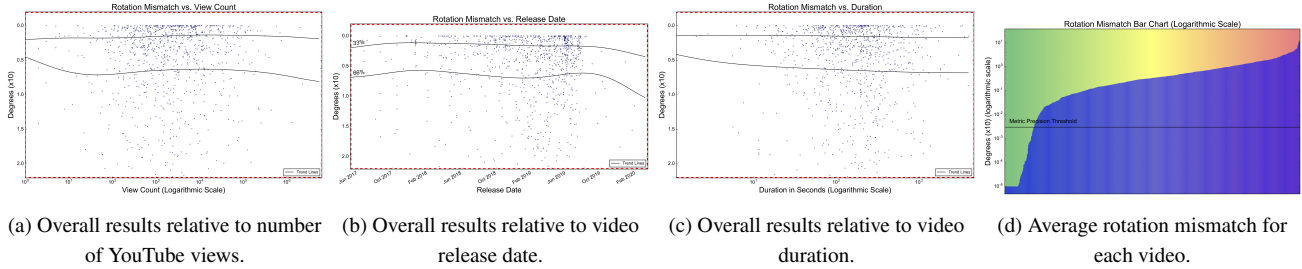


Figure 7: Results of rotation-mismatch analysis for VR180 videos.

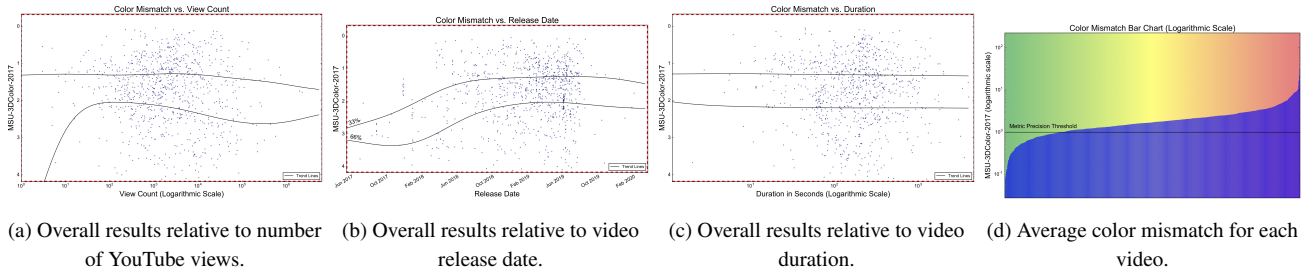


Figure 8: Results of color-mismatch analysis for VR180 videos.

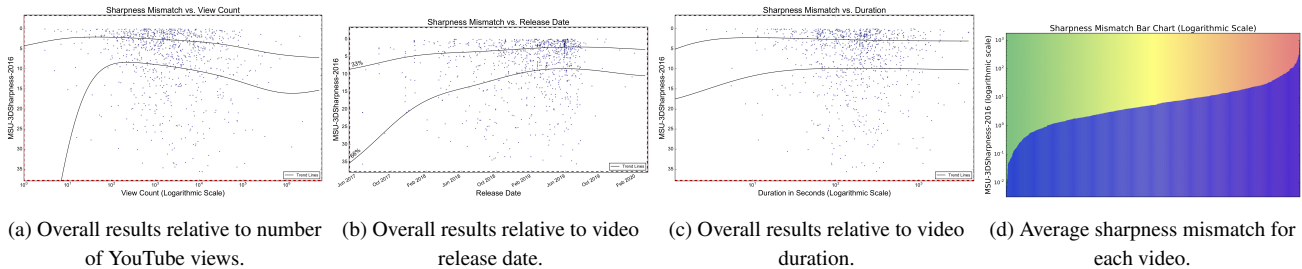


Figure 9: Results of sharpness-mismatch analysis for VR180 videos.

tain shots with huge depth budgets. According to this chart, the overall situation is better than the previous chart indicates: most shots have an average depth budget, and extremely wide shots are infrequent. A large percentage of shots are too flat, however.

Analysis of Common 3D-Shooting Artifacts

For all 1,000 VR180 videos, we calculated scores for each geometric distortion: vertical shift (Figure 5), scale mismatch (Figure 6), rotation mismatch (Figure 7); color mismatch (Figure 8) and sharpness mismatch (Figure 9). We plotted the overall results relative to the number of YouTube views (a), video release date (b) and video duration (c). The *x*-axis in these charts corresponds to the particular video statistic, and the *y*-axis corresponds to the S3D artifact. The blue dots represent individual videos. We also included two trend lines: the top one is for the 33rd percentile and the bottom one is for the 66th percentile. None of the stereoscopic artifacts or video statistics we consider exhibits any significant trend: some charts reveal slight decreases or increases, but the average estimated distortion values change little. Sudden descents and ascents emerge on the left and right sides of some charts, but they are mostly due to a small number of videos with the corresponding statistics. The plots allow us to make the following conclusions:

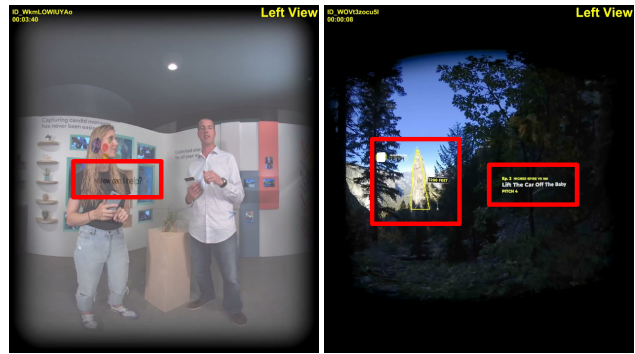
- Videos with many YouTube views have, on average, estimated distortion values similar to those of videos with few views;
- That situation has remained unchanged over time, as videos released later have average scores similar to those released earlier;
- The artifact values are independent of video duration.

But a substantial number of the VR180 videos exhibit at least one S3D artifact from the group we analyzed. Figure (d) shows the average metric values (*y*-axis) for each video (*x*-axis). Many videos have slight stereoscopic distortions, but several extreme cases appear, too. The left sides of the charts with geometric artifacts also contain flat regions indicating no geometric distortions. These areas correspond either to flat videos whose views are the same or to CGI videos.

Channel-Mismatch Analysis

Channel mismatch is a stereoscopic distortion that occurs when the left view of a 3D video replaces the right view and vice versa. This S3D artifact is rare, but even one scene with swapped views can cause viewers to suffer serious discomfort [25]. Swapped views can additionally result from incorrect editing of video content — for example, when adding titles and CGI elements with the wrong depth.

We analyzed the 50 most viewed VR180 videos to find channel mismatch. Using our channel-mismatch metric, we detected 21 scenes with swapped views (a result we verified manually) in 10 videos. The probability that a video contains a scene with channel mismatch is therefore 20%, according to this result. But most of the scenes we identified with channel mismatch have incorrectly placed titles and CGI. Figure 10 presents examples. These mistakes probably owe to amateur videographers with little or no knowledge of 3D-scene composition, along with failure to check the resulting depth of added objects because a special tool for that task was unavailable.



(a) Left views for scenes with channel mismatch. Red rectangles highlight objects with incorrect depth.



(b) Anaglyph for scenes with channel mismatch.

Figure 10: Examples of VR180 scenes with channel mismatch.

Conclusion

In this paper we presented the results of a large-scale stereoscopic-quality analysis of 1,000 VR180 videos from YouTube using eight objective metrics. To summarize, we observed no significant trend in any measured technical-quality parameter relative to several video statistics, such as number of YouTube views, release date and duration. Most of the videos we analyzed contain at least one severe stereoscopic distortion, meaning some viewers will probably experience discomfort after watching several such videos. This situation reveals the need to develop quality-control and correction tools to help creators improve their video content.

We plan to soon release a full technical report based on our analysis. The report will include approximately 400 pages with overall charts, S3D-artifact examples and video statistics. It will be available on the main VQMT3D project page: http://videoprocessing.ml/stereo_quality/.

Acknowledgments

This work was supported by the START program of the State Fund for Support of Small Enterprises in the Scientific-Technical Fields under the project “Development of a system for automatic objective quality assessment and correction of stereoscopic video and video in VR180 format.”

This work was partially supported by the Russian Foundation for Basic Research under Grant 19-01-00785 a.

References

- [1] “VR180.” available online: <https://arvr.google.com/vr180/>.
- [2] D. Khaustova, J. Fournier, E. Wyckens, and O. Le Meur, “An objective method for 3d quality prediction using visual annoyance and acceptability level,” in *Stereoscopic Displays and Applications XXVI*, vol. 9391, p. 93910P, International Society for Optics and Photonics, 2015.
- [3] E. Dumić, S. Grgić, K. Šakić, P. M. R. Rocha, and L. A. da Silva Cruz, “3d video subjective quality: a new database and grade comparison study,” *Multimedia tools and applications*, vol. 76, no. 2, pp. 2087–2109, 2017.
- [4] J. Yang, Y. Zhu, C. Ma, W. Lu, and Q. Meng, “Stereoscopic video quality assessment based on 3d convolutional neural networks,” *Neurocomputing*, vol. 309, pp. 83–93, 2018.
- [5] A. Banitalebi-Dehkordi, M. T. Pourazad, and P. Nasiopoulos, “An efficient human visual system based quality metric for 3d video,” *Multimedia Tools and Applications*, vol. 75, no. 8, pp. 4187–4215, 2016.
- [6] Y. Han, Z. Yuan, and G.-M. Muntean, “Extended no reference objective quality metric for stereoscopic 3d video,” in *2015 IEEE International Conference on Communication Workshop (ICCW)*, pp. 1729–1734, IEEE, 2015.
- [7] P. Aflaki, M. M. Hannuksela, and M. Gabbouj, “Subjective quality assessment of asymmetric stereoscopic 3d video,” *Signal, Image and Video Processing*, vol. 9, no. 2, pp. 331–345, 2015.
- [8] S. Winkler, “Efficient measurement of stereoscopic 3D video content issues,” in *Image Quality and System Performance XI*, vol. 9016, p. 90160Q, International Society for Optics and Photonics, 2014.
- [9] Q. Dong, T. Zhou, Z. Guo, and J. Xiao, “A stereo camera distortion detecting method for 3DTV video quality assessment,” in *2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pp. 1–4, IEEE, 2013.
- [10] F. Devernay and S. Pujades, “Focus mismatch detection in stereoscopic content,” in *Stereoscopic Displays and Applications XXIII*, vol. 8288, p. 82880E, International Society for Optics and Photonics, 2012.
- [11] M. Liu and K. Müller, “Automatic analysis of sharpness mismatch between stereoscopic views for stereo 3D videos,” in *2014 International Conference on 3D Imaging (IC3D)*, pp. 1–6, IEEE, 2014.
- [12] I. Rocco, R. Arandjelović, and J. Sivic, “End-to-end weakly-supervised semantic alignment,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6917–6925, 2018.
- [13] M. Knee, “Getting machines to watch 3d for you,” *SMPTE Motion Imaging Journal*, vol. 121, no. 3, pp. 52–58, 2012.
- [14] J. Bouchard, Y. Nazzar, and J. J. Clark, “Half-occluded regions and detection of pseudoscopy,” in *2015 International Conference on 3D Vision*, pp. 215–223, IEEE, 2015.
- [15] S. Chen, Y. Zhang, Y. Li, Z. Chen, and Z. Wang, “Spherical structural similarity index for objective omnidirectional video quality assessment,” in *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, IEEE, 2018.
- [16] S. Croci, S. Knorr, L. Goldmann, and A. Smolic, “A framework for quality control in cinematic vr based on voronoi patches and saliency,” in *2017 International Conference on 3D Immersion (IC3D)*, pp. 1–8, IEEE, 2017.
- [17] R. Dudek, S. Croci, A. Smolic, and S. Knorr, “Robust global and local color matching in stereoscopic omnidirectional content,” *Signal Processing: Image Communication*, vol. 74, pp. 231–241, 2019.
- [18] D. Vatolin, A. Bokov, M. Erofeev, and V. Napadovsky, “Trends in s3d-movie quality evaluated on 105 films using 10 metrics,” *Electronic Imaging*, vol. 2016, no. 5, pp. 1–10, 2016.
- [19] D. Vatolin and A. Bokov, “Sharpness mismatch and 6 other stereoscopic artifacts measured on 10 chinese s3d movies,” *Electronic Imaging*, vol. 2017, no. 5, pp. 137–144, 2017.
- [20] K. Simonyan, S. Grishin, D. Vatolin, and D. Popov, “Fast video super-resolution via classification,” in *2008 15th IEEE International Conference on Image Processing*, pp. 349–352, IEEE, 2008.
- [21] G. Egnal and R. P. Wildes, “Detecting binocular half-occlusions: Empirical comparisons of five approaches,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 8, pp. 1127–1133, 2002.
- [22] A. Voronov, D. Vatolin, D. Sumin, V. Napadovsky, and A. Borisov, “Methodology for stereoscopic motion-picture quality assessment,” in *Stereoscopic Displays and Applications XXIV*, vol. 8648, p. 864810, International Society for Optics and Photonics, 2013.
- [23] Q. Zhang, L. Xu, and J. Jia, “100+ times faster weighted median filter (wmf),” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2830–2837, 2014.
- [24] S. Lavrushkin and D. Vatolin, “Channel-mismatch detection algorithm for stereoscopic video using convolutional neural network,” in *2018-3DTV-Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON)*, pp. 1–4, IEEE, 2018.
- [25] D. Vatolin and S. Lavrushkin, “Investigating and predicting the perceptibility of channel mismatch in stereoscopic video,” *Moscow University Computational Mathematics and Cybernetics*, vol. 40, no. 4, pp. 185–191, 2016.

Author Biography

Sergey Lavrushkin received his M.S. degree in computer science from the Moscow State University (2017), where he is currently a Ph.D. student. His research interests include S3D video processing and quality assessment, machine learning and neural networks. His contact email is sergey.lavrushkin@graphics.cs.msu.ru.

Ivan Molodetskikh received his M.S. degree in computer science from the Moscow State University (2020), where he is currently a Ph.D. student. His research interests include VR180 quality assessment, image inpainting, semantic video matting and machine learning. His contact email is ivan.molodetskikh@graphics.cs.msu.ru.

Konstantin Kozhemyakov received his B.S. degree in computer science from the Moscow State University (2020), where he is currently a Master student. His research interests include applications of neural networks and machine learning to video processing and quality assessment. His contact email is konstantin.kozhemiakov@graphics.cs.msu.ru.

Dmitriy Vatolin received his Ph.D. in 2000 from Moscow State University. Currently he is head of the Video Group at the CS MSU Graphics & Media Lab. His research interests include compression methods, video processing, 3D video techniques (depth from motion, focus and other cues, video matting, background restoration, high-quality stereo generation), as well as 3D video quality assessment (metrics for 2D-to-3D conversion artifacts, temporal asynchrony, swapped views and many others). He is a key organizer of the 3D video quality measurement project VQMT3D (http://videoprocessing.ml/stereo_quality/). His contact email is dmitriy@graphics.cs.msu.ru.

JOIN US AT THE NEXT EI!

IS&T International Symposium on

Electronic Imaging

SCIENCE AND TECHNOLOGY

Imaging across applications . . . Where industry and academia meet!



- **SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •**
- **INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •**

www.electronicimaging.org

