https://doi.org/10.2352/lim.2024.5.1.5

Colourlab Image Database: Optical Aberrations

Raed Hlayhel, Mobina Mobini, Bidossessi Emmanuel Agossou, Marius Pedersen, and Seyed Ali Amirshahi
Norwegian University of Science and Technology, Department of Computer Science, Gjøvik, Norway.

Abstract

For image quality assessment, the availability of diverse databases is vital for the development and evaluation of image quality metrics. Existing databases have played an important role in promoting the understanding of various types of distortion and in the evaluation of image quality metrics. However, a comprehensive representation of optical aberrations and their impact on image quality is lacking. This paper addresses this gap by introducing a novel image quality database that focuses on optical aberrations. We conduct a subjective experiment to capture human perceptual responses on a set of images with optical aberrations. We then test the performance of selected objective image quality metrics to assess these aberrations. This approach not only ensures the relevance of our database to real-world scenarios but also contributes to ensuring the performance of the selected image quality metrics. The database is available for download at `https://www.ntnu.edu/colourlab/software`.

Introduction

The measurement and assessment of an image's perceived quality is known as Image Quality Assessment (IQA). It is an essential part of most everyday applications such as multimedia, computer vision, and image processing tasks. To ensure that photos are technically appropriate in terms of quality for their applications, evaluating image quality is vital [1]. The two ways to evaluate image quality are subjective and objective assessments. The subjective assessment of image quality uses the subjective judgment of human observers to evaluate the quality of images. Objective assessments use algorithms and computational techniques to measure image quality.

Objective assessment, with so-called Image Quality Metrics (IQMs), requires evaluation against ground-truth data to ensure the metric's performance and consistency. Most IQMs aim to predict perceived image quality, and their performance is determined by the agreement with perceived quality ratings using, for example, correlation measures. There is a wide range of databases created with the intention of evaluating IQMs, such as Kadid-10K [2], TID2008 [3], TID2013 [4], SEID [5], LIVE [6], CID:IQ [7], CID:GD [8], and CID image demosaicing [9]. These databases have been crucial for benchmarking IQMs, however, they mostly contain similar distortions such as JPEG compression, blur, and noise. By analyzing these databases we have identified a gap in the distortions, namely optical aberrations. Optical aberrations are present in all imaging systems in varying degrees of intensity, and are therefore relevant in the context of IQA.

With the goal of evaluating the performance of IQMs on images with optical aberrations, in this paper, we introduce a new image quality database called "Colourlab Image Database: Optical Aberrations (CID: OA)". To the best of our knowledge, defocus, astigmatism, and spherical aberrations have not been incorporated in existing IQA databases. The utilization of these aberrations contributes to a more comprehensive understanding of image quality and the creation of IQMs.[1]

This paper is organized as follows: the methodology followed to simulate the aberrant images is presented in "Methodology: Optical aberrations", details regarding the assessment of the simulated database, by subjects and the evaluations of IQMs, are presented in "Methodology: Subjective Experiment" and "Methodology: Objective Metrics" respectively. The results of the two assessments and discussion are presented in "Results and discussion". In the final section, we conclude the paper.

Methodology

The work consists of different parts, first, we will introduce the aberration types and how these have been simulated, followed by the subjective experimental setup to evaluate the quality of the aberrant images, and lastly, the IQMs used and how we assess their performance on the aberrant images.

Optical aberrations

The core of an imaging tool is its optical system. The role of the optical system is to focus and optimise the light distribution onto the sensor for maximum information recording from the object. For a perfect optical system, light passes in perfectly linear "wave-fronts" and is focused on the sensor [10, 11]. However, a perfect optical system does not exist and real optical systems do not function in this manner. Optical elements of the system (such as lenses, mirrors, and/or apertures) introduce wavefront distortions. These wavefront distortions are responsible for what is described as optical aberrations. These optical aberrations lead to the dispersion of light rays instead of focusing them onto the sensor, which results in a lower image quality output (blurry or size-distorted image). There are multiple aberration categories, each depending on the type of wavefront distortion introduced by the optical system. Also, aberrations can be: achromatic (does not depend on the wavelength) and chromatic (depends on the wavelength). Aberrations can be visualised through the Point Spread Function (PSF).

The resulting image (I) captured by an optical system is the convolution of the object to capture (O) and the PSF of the optical system. Equation 1 shows how this image is formed:

$$I(x,y) = O(x,y) \circledast PSF(x,y) \tag{1}$$

Therefore, knowing the PSF can be very useful to determine the captured image quality. Also, the shape of the PSF allows the identification of the aberrations in play in the optical system and their intensity. Knowing this, aberrations can be classified into different categories following the shapes of PSFs produced at the output of the optical system. These different categories of aberrations are referred to using Zernike polynomials. These polynomials relate the PSF shape and intensity to physical quantities proper to the optical system, mainly the radial distance of

[1] Contrary to image quality literature, we refer to the simulated images as *aberrant* images instead of *distorted*. This is done to avoid confusion between the two different meanings of the term "Distortion" in image quality and optical engineering, where distortion is a type of optical aberration.

the lens and the azimuthal angle of incoming light rays onto the optical system.

Hence, optical aberrations are always present in real optical systems. Consequently, it would be interesting to simulate their effects on captured images and see the output of these aberrant optical systems. There are a multitude of aberrations that can be considered. However, in this paper, we implemented three achromatic aberrations. These aberrations are all characterized by a blurry image output. The simulated aberrations are presented below:

Defocus: A defocused system is a system where the incoming light rays are focused close to, but not on the focal plane. This is usually due to a miscalculation of the focal distance, or a slight misplacement of the optical elements (lenses, mirrors) or the sensor.

Astigmatism: An astigmatic optical system is one where the sagittal and tangential axes of the lens have different focus points i.e. the lens has two focal planes, a sagittal and a tangential. This is a more complex situation than the defocus case and requires optical correction and optimization of the optical system.

Spherical aberration: Spherical aberrations occur in optical systems containing un-optimized spherical elements such as spherical lenses and mirrors. It is caused by a difference in refraction (or reflection in the case of mirrors) power between light rays coming from the lens's extremities and rays passing through the center of the lens. This difference in refraction causes each light ray to focus on different points from one another instead of focusing all on the focal plane.

ISETcam, an open-source MATLAB toolbox developed by the Stanford Center for Image Systems Engineering [9], was used for the simulation. We simulate these aberrations by generating, using ISETcam, aberrant PSF profiles that get convolved with the reference images, resulting in aberrant images. We simulate the PSF profiles and their intensity by first selecting the Zernike polynomial proper to each aberration type we want to implement, and then setting the polynomial's value to the wanted intensity.

We used 23 reference images from the CID:IQ database [7] because they cover a wide range of characteristics, mainly colourfulness and spatial information. Then we applied the aberrations on each reference image at four levels: defocus (levels 0.5, 1, 1.5, and 2), astigmatism (levels 0.5, 1, 1.5, and 2), and spherical aberration (levels 0.2, 0.4, 0.6, and 0.9). Aberration levels were chosen in such a way as to allow observers to perceive visual differences between the aberrant images. We have 12 aberrant images per reference image, therefore overall, the created database contains 299 images (276 aberrant images and 23 reference images). All images are in PNG format with a size of 800×800 pixels. An illustration of the simulation process is presented in Figure 1.

Subjective Experiment

After generating the aberrant images, we conducted a subjective categorical judgment experiment in a controlled environment to evaluate the quality of the images. We calibrated the evaluation room and screen to achieve a standard viewing space for all of the observers which was done in the following way:

1. The room light was dimmed. We used a Luxmeter to measure the illuminance in the experiment room. We placed it near the keyboard of the computer where the experiment was taking place.
2. An Eizo CG279X monitor (resolution: 2560 x 1440, pixel density: 109 ppi) was calibrated to sRGB viewing standard (Brightness: 80 cd/m, Temperature: D65, Gamma: 2.2, ColourGamut: sRGB).
3. A fixed viewing distance from the screen is set to 58 cm. We used a headrest to make sure that the distance is stable throughout the experiment and across all experiments (for all of the observers).
4. We held the observers in the environment for two minutes before starting the experiment to allow the observer's vision to adapt to the room lighting.

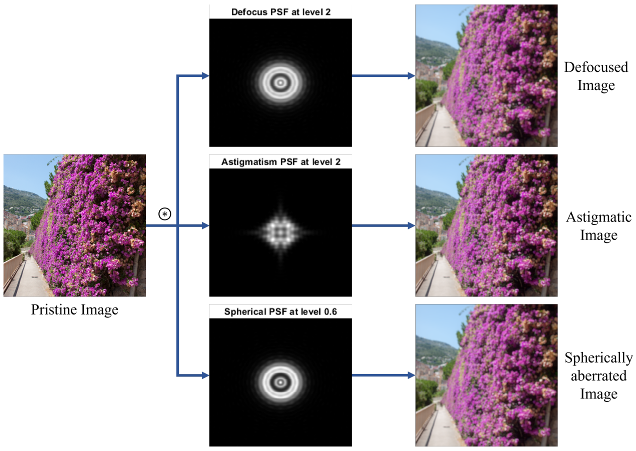

Figure 1. An illustration summarizing the aberration simulation process. The PSFs are generated in ISETcam and then convolved with the reference image to obtain the aberrant images

The experiment was carried out by 15 observers (40% female, 60% male, average age 30 years) on the QuickEval online platform [12]. The instructions given to the observers are as follows: "Please rate the overall technical quality of the images. The scale is from 1 to 5, where 1 is 'Very Low Quality' and 5 is 'Very High Quality' ". Before the start of each experiment, we ensured that each observer had 20/20 visual acuity using a Snellen chart. We also tested the observers for colour deficiency via Ishihara Test, even though colour deficiency would not affect the results because the focus in the aberrant images was mainly on structural changes rather than chromatic changes. Therefore, observers passed the following two tests before entering the experiment room: (a) reading from 3 meters distance the "Sloan Letters In LogMAR SIZES" (b) reading the HRR pseudoisochromatic plate. The subjective ratings were processed to obtain Mean Opinion Scores (MOS).

Objective Metrics

We tested the database on 19 state-of-the-art IQMs. The IQMs are Structural Similarity (SSIM) [13], Multiscale Structural Similarity (MSSIM) [14], Peak-Signal-to-Noise Ratio (PSNR), Mean Squared Error (MSE) [15], Blind Image Quality Assessment through Anisotropy (BIQAA) [16], BLUR Metric [17], HUE angle [18], Spatial Hue Angle Metric (SHAMEII) [19], Adaptive Bilateral Filter (ABF) [20], Local Phase Coherence Sharpness Index (Lpc-si) [21], Cumulative Probability of Blur Detection (CPBD) [22], COLORPSNRHVMA [23], Feature SIMiliarity Index (FSIM) [24], EntropyDiff [25], Just Noticeable Blur Metric (JNBM) [22], Most Apparent Distortion (MAD) [26], Spatial-CIELAB (S-CIELAB) [27], Total Variation of Difference (TVD) [28], and Visual Information Fidelity (VIF) [29]. These IQM provide a good selection in terms of measuring image quality in the spatial and frequency domains.

The performance of the IQMs was assessed using the most common performance criteria which are, Spearman Rank-Order

Correlation Coefficient (SROCC), Pearson Linear Correlation Coefficient (PLCC), and Kendall Rank Correlation Coefficient (KROCC). These are shown with a 95% confidence interval. In addition, we have calculated Root-Mean-Squared-Error and outlier ratio [30].

Results and discussion

Results of our experiment are found in Figures 2, 3, and 4.

Analysis of the subjective scores

From Figure 2, we can see that the subjective results are coherent: reference images have the highest opinion scores overall with an average score of 4.6 out of 5, and MOS decreases when the level of aberration increases.

Also, there is a clear preference shown by observers for some aberrations over others. Vertical astigmatism aberrant images are rated higher than other aberrations for the same aberration level, followed by defocus and lastly spherical aberration. This may be explained by the fact that, although the output of the three aberrations is characterized by blur, vertically astigmatic images are blurry mostly following the x and y axis, whereas defocused and spherically aberrant are blurry throughout the picture. This indicates that observers were able to differentiate between the aberration types and have a clear preference for one type over others.

We also analyzed the inter-rater reliability using Fleiss Kappa for multiple observers [31]. The analysis shows that for all images the observers have a fair agreement with a Fleiss Kappa of 0.28. Investigation of the Fleiss Kappa for each of the 23 contents (reference images and their corresponding aberrant images) shows that all Kappa values are between 0.22 and 0.33, meaning that the agreement for all contents are fair. For all tests the p-values are close to 0, so the observed agreement is not accidental.

Analysis of the objective scores

We analysed the performance of the state-of-the-art IQMs using the aforementioned correlation measures. Figure 3 shows the performance of the tested IQMs on our database. We observe that the Lpc-si metric has the highest overall correlation. For Lpc-si, the SROCC, PLCC, and KROCC values are 86%, 87%, and 68% respectively. We also observe that there are several IQMs that perform quite well (correlation coefficient higher than 80%), such as FSIM and MS-SSIM. On the other hand, there is a group of IQMs that perform quite poorly (correlation coefficient lower than 30%), such as TVD, S-CIELAB, and ABF. We also analysed the performance of Lpc-si for each of the aberrations, and the SROCC for each of the aberrations ranges from 0.835 to 0.845 while the overall SROCC is 0.862. This indicates that Lpc-si performs similarly for each of the aberrations.

First, when comparing IQMs, we notice a disparity between different correlation coefficients. For example, MS-SSIM shows a 20% difference between the SROCC and PLCC values, which indicates that MS-SSIM has a non-linear behaviour. On the contrary, Lpc-si shows similar PLCC and SROCC values (87% and 86% respectively), and its point distribution is highly linear.

Our analysis also reveals that some IQMs, such as SHAMEII, ABF, and TVD, are impacted by the content of the images and that they can perform well for some images while having difficulty with other images. These IQMs seem to have problems with scale differences between images, which has been previously reported for other metrics in [28].

Although not shown in this paper because of page limitation, the analysis of the RMSE shows similar information about the performance of each IQM. Similarly to what was observed in Figure 3, the separation in two groups of IQMs is also visible. However, according to our experiments with RMSE, the IQM with the lowest RMSE is FSIM and not the Lpc-si, and the highest RMSE value is that of the MSE and not that of TVD.

Figure 4 shows the outlier ratio of the evaluated IQMs. The outlier ratio is defined as the percentage of the number of predictions outside the interval of ± 2 times the standard error of the MOS. LPC-Si has the lowest outlier ratio, but on average many of the IQMs have a high outlier ratio, indicating a low prediction consistency. Overall, there is a lack of consistency in terms of correlation, linearity, and outlier ratio among all the tested metrics and there is no perfect metric that succeeds in accurately predicting the subjective scores. This shows that the dataset is challenging for IQMs, which contributes to a more detailed evaluation of IQMs.

By examining the results of the IQMs, we see that the best-performing IQMs are structural difference-based, while the worst-performing metrics are colour-difference-based. This distinction in the performance of different IQMs can be explained by the fact that the aberrations applied are characterized by the blur they add to the images, although each aberration blurs the images differently. When applied, blur affects the structure of the images, but does not necessarily modify colour components. Knowing this, we can see why the structural difference-based IQMs perform the best on our database. Also, this explains why the Lpc-si metric has the highest correlation coefficient, since it is used specifically for sharpness and blur characterization.

However, this does not mean that the Lpc-si performance can be generalized to all optical aberrations because other types of aberrations are not characterized by uniform blur, or by blur at all. Also, aberrations can be both "achromatic" and "chromatic". It may be the case that structural difference-based IQMs perform better for achromatic aberrations, while colour-difference-based metrics perform better for chromatic aberrations, but this needs to be further evaluated.

Also, in a real optical system, these two types of aberrations are present simultaneously. So, we cannot completely ignore one type and only treat the other. Therefore, concluding that structural-difference-based IQMs are better suited for aberrant image quality assessment becomes even less convincing.

Conclusion

Databases for the evaluation of image quality metrics have many types of distortions applied, but none, to our knowledge, has implemented optical aberrations. Therefore, we sought to fill the gap in the literature and create a new database with optical aberrations for evaluating the performance of image quality metrics. We successfully simulated defocus, vertical astigmatism, and spherical aberrations on a set of images with four different levels each, using an open-source code on MATLAB called ISETcam. This was achieved by producing aberrant PSFs and applying a convolution of them with reference images to simulate aberrant images. After simulating the aberrant images, we measured the subject's opinion on quality, through a psychometric experiment. We tested the performance of IQMS on the created images. In the subjective experiments, observers preferred astigmatic images over defocused and spherically aberrant images. Objectively, the Lpc-si IQM had the highest PLCC correlation coefficient which was 87%. The database is available for download at `https://www.ntnu.edu/colourlab/software`.

Future research could build upon this work by simulating

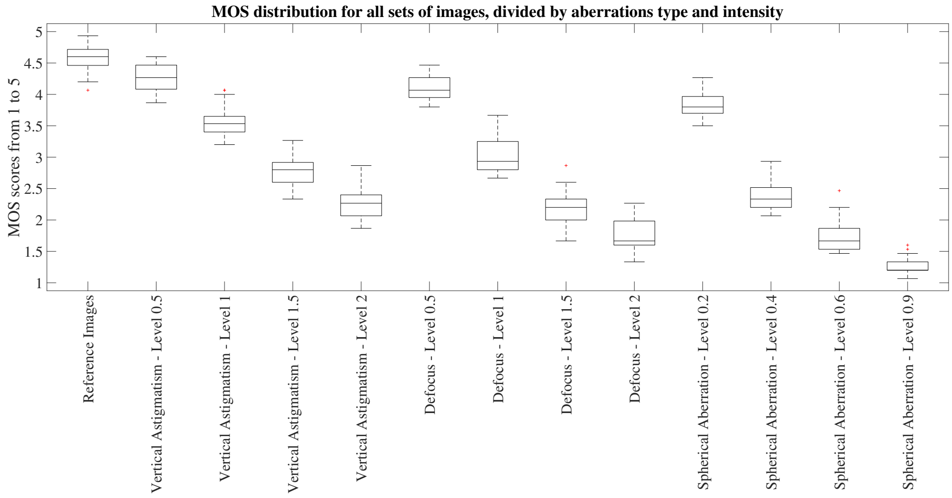

Figure 2. *Illustration of the subjective data in the form of MOS box-plots for each aberration type - level. Box-plots illustrate the quartile distribution of each category, along with their averages and outliers.*

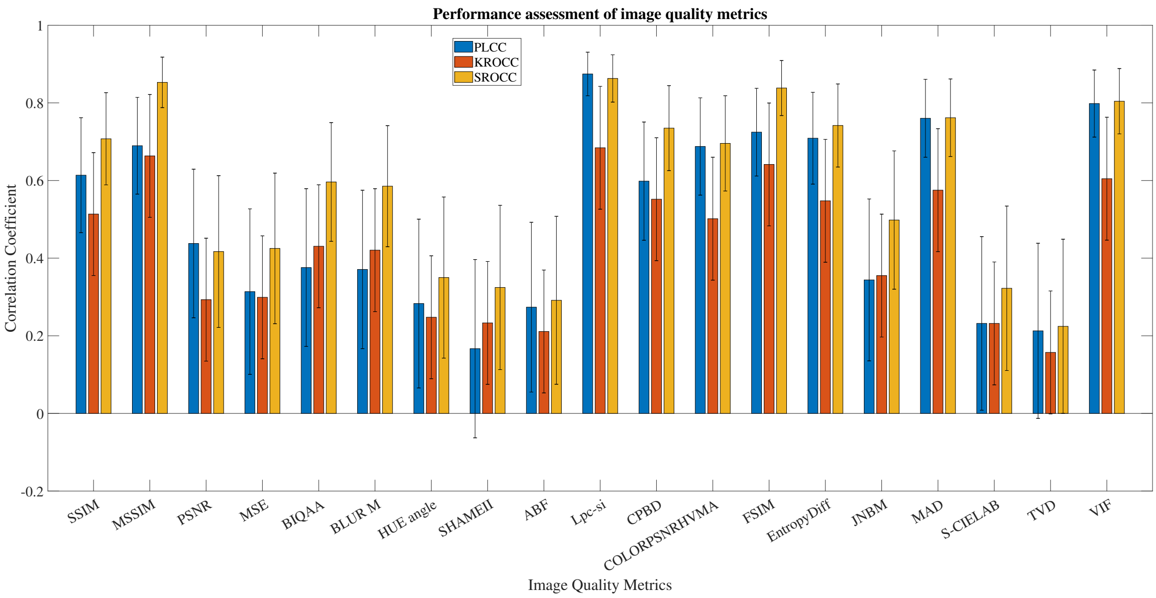

Figure 3. *Performance comparison of different IQMs applied on our database, along with their 95% confidence intervals.*

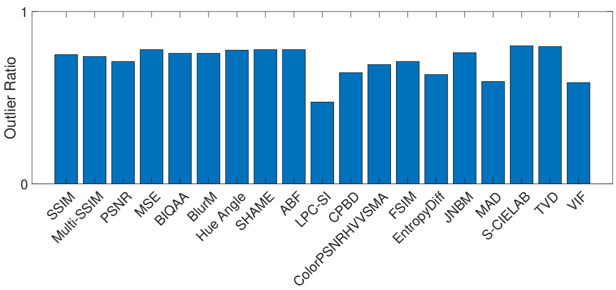

Figure 4. *The outlier ratios of the different IQMs on our database.*

more optical aberration types, such as field curvature, distortion, etc. combined with using optical simulation software.

Acknowledgements

Marius Pedersen and Mobina Mobini are supported by the Research Council of Norway through the “Quality and Content” project (grant number 324663), while Bidossessi Emmanuel Agossou has been supported by the Research Council of Norway through the “CapsNetwork” project (grant number 322600). Seyed Ali Amirshahi is supported by the project “VQ4MedicS: Video Quality Assessment and Enhancement for Pre-Hospital Medical Services” (grant number 329034) from the Research Council of Norway.

References

[1] Damon M Chandler. “Seven challenges in image quality assessment: past, present, and future research”. In: *International Scholarly Research Notices* 2013 (2013).

[2] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. “KADID-10k: A large-scale artificially distorted IQA database”.

In: *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE. 2019, pp. 1–3.

[3] Nikolay Ponomarenko et al. "TID2008-a database for evaluation of full-reference visual quality assessment metrics". In: *Advances of modern radioelectronics* 10.4 (2009), pp. 30–45.

[4] Nikolay Ponomarenko et al. "Image database TID2013: Peculiarities, results and perspectives". In: *Signal processing: Image communication* 30 (2015), pp. 57–77.

[5] Sahar Azimian, Farah Torkamani-Azar, and Seyed Ali Amirshahi. "How good is too good? a subjective study on over enhancement of images". In: (2021).

[6] Hamid R Sheikh, Muhammad F Sabir, and Alan C Bovik. "A statistical evaluation of recent full reference image quality assessment algorithms". In: *IEEE Transactions on image processing* 15.11 (2006), pp. 3440–3451.

[7] Xinwei Liu, Marius Pedersen, and Jon Yngve Hardeberg. "CID: IQ–a new image quality database". In: *Image and Signal Processing: 6th International Conference, ICISP 2014, Cherbourg, France, June 30–July 2, 2014. Proceedings 6*. Springer. 2014, pp. 193–202.

[8] Marius Pedersen and Seyed Ali Amirshahi. "Colourlab Image Database: Geometric Distortions". In: *Color and Imaging Conference* 29.1 (2021), pp. 258–258.

[9] Tawsin Ahmed, Seyed Ali Amirshahi, and Marius Pedersen. "Image demosaicing: Subjective analysis and evaluation of image quality metrics". In: *Electronic Imaging* 35 (Jan. 2023), pp. 301–1. DOI: 10.2352/EI.2023.35.8.IQSP-301.

[10] W.T. Welford. *Aberrations of Optical Systems*. Nov. 2017, pp. 1–284. ISBN: 9781315136530. DOI: 10.1201/9781315136530.

[11] H. Gross. *Handbook of Optical Systems*. Sept. 2005, pp. 1–826. DOI: 10.1002/9783527699223.

[12] Khai Van Ngo et al. "Quickeval: a web application for psychometric scaling experiments". In: *Image Quality and System Performance XII*. Vol. 9396. SPIE. 2015, pp. 212–224.

[13] Zhou Wang et al. "Image quality assessment: from error visibility to structural similarity". In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612.

[14] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. "Multiscale structural similarity for image quality assessment". In: *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*. Vol. 2. Ieee. 2003, pp. 1398–1402.

[15] Zhou Wang and Alan C Bovik. "Mean squared error: Love it or leave it? A new look at signal fidelity measures". In: *IEEE signal processing magazine* 26.1 (2009), pp. 98–117.

[16] Salvador Gabarda and Gabriel Cristóbal. "Blind image quality assessment through anisotropy". In: *JOSA A* 24.12 (2007), B42–B51.

[17] Frederique Crete et al. "The blur effect: perception and estimation with a new no-reference perceptual blur metric". In: *Human vision and electronic imaging XII*. Vol. 6492. SPIE. 2007, pp. 196–206.

[18] Guowei Hong and Ming Ronnier Luo. "Perceptually-based color difference for complex images". In: *9th Congress of the International Colour Association*. Vol. 4421. SPIE. 2002, pp. 618–621.

[19] Marius Pedersen and Jon Yngve Hardeberg. "A new spatial hue angle metric for perceptual image difference". In: *Computational Color Imaging: Second International Workshop, CCIW 2009, Saint-Etienne, France*. Springer. 2009, pp. 81–90.

[20] Zhaohui Wang and Jon Yngve Hardeberg. "An adaptive bilateral filter for predicting color image difference". In: *Color and Imaging Conference*. Vol. 2009. 1. Society for Imaging Science and Technology. 2009, pp. 27–31.

[21] Rania Hassen, Zhou Wang, and Magdy MA Salama. "Image sharpness assessment based on local phase coherence". In: *IEEE Transactions on Image Processing* 22.7 (2013), pp. 2798–2810.

[22] Rony Ferzli and Lina J Karam. "A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB)". In: *IEEE transactions on image processing* 18.4 (2009), pp. 717–728.

[23] Nikolay Ponomarenko et al. "Modified image visual quality metrics for contrast change and mean shift accounting". In: *2011 11th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM)*. IEEE. 2011, pp. 305–311.

[24] Lin Zhang et al. "FSIM: A feature similarity index for image quality assessment". In: *IEEE transactions on Image Processing* 20.8 (2011), pp. 2378–2386.

[25] Aristides D Vaiopoulos. "Developing Matlab scripts for image analysis and quality assessment". In: *Earth Resources and Environmental Remote Sensing/GIS Applications II*. Vol. 8181. SPIE. 2011, pp. 60–71.

[26] Eric C Larson and Damon M Chandler. "Most apparent distortion: full-reference image quality assessment and the role of strategy". In: *Journal of electronic imaging* 19.1 (2010), pp. 011006–011006.

[27] Xuemei Zhang, Brian A Wandell, et al. "A spatial extension of CIELAB for digital color image reproduction". In: *SID international symposium digest of technical papers*. Vol. 27. Citeseer. 1996, pp. 731–734.

[28] Marius Pedersen and Ivar Farup. "Improving the robustness to image scale of the total variation of difference metric". In: *2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN)*. IEEE. 2016, pp. 116–121.

[29] Hamid R Sheikh and Alan C Bovik. "Image information and visual quality". In: *IEEE Transactions on image processing* 15.2 (2006), pp. 430–444.

[30] Telecommunication Standardization Sector of ITU. *Objective perceptual assessment of video quality: Full reference television*. 2004.

[31] Cardillo G. *Fleiss'es kappa: compute the Fleiss'es kappa for multiple raters.* 2007. URL: http://www.mathworks.com/matlabcentral/fileexchange/15426.