Perceptual Image Quality Assessment for Various Viewing Conditions and Display Systems

Andrei Chubarau[1], Tara Akhavan[2], Hyunjin Yoo[2], Rafal K. Mantiuk[3] and James Clark[1]

[1] Department of Electrical and Computer Engineering, McGill University, Montreal, Canada
[2] IRYStec Software Inc., Montreal, Canada
[3] Department of Computer Science and Technology, University of Cambridge, UK
andrei.chubarau@mail.mcgill.ca, {tara, hyunjin.yoo}@irystec.com, rafal.mantiuk@cl.cam.ac.uk, clark@cim.mcgill.ca

Abstract

From complete darkness to direct sunlight, real-world displays operate in various viewing conditions often resulting in a non-optimal viewing experience. Most existing Image Quality Assessment (IQA) methods, however, assume ideal environments and displays, and thus cannot be used when viewing conditions differ from the standard. In this paper, we investigate the influence of ambient illumination level and display luminance on human perception of image quality. We conduct a psychophysical study to collect a novel dataset of over 10000 image quality preference judgments performed in illumination conditions ranging from 0 lux to 20000 lux. We also propose a perceptual IQA framework that allows most existing image quality metrics (IQM) to accurately predict image quality for a wide range of illumination conditions and display parameters[1]*. Our analysis demonstrates strong correlation between human IQA and the predictions of our proposed framework combined with multiple prominent IQMs and across a wide range of luminance values.*

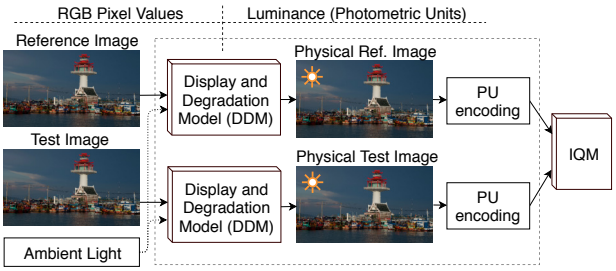

Figure 1. *Our proposed perceptual IQA framework. We estimate the visual signal that reaches the observer's eye given the information about the environment and the used display system, and apply existing IQA metrics on the resulting stimuli. A display and degradation model converts digital inputs from gamma-corrected pixel values to the physical luminance space and simulates the influence of the ambient illumination level. The resulting signals are then linearized with perceptually uniform (PU) encoding [17] to account for luminance masking and, finally, IQA is computed.*

Introduction

For any display system with the ultimate goal of accurately communicating visual information to its viewer, image quality and legibility are key performance metrics that must be assured. This is especially prominent in safety critical display systems, for instance those found in the automotive industry. While the exact methodology and criteria of objective IQA can vary – one can evaluate a variety of properties, such as presence of digital degradations, level of contrast, overall aesthetic quality, content visibility, image naturalness, etc. - the consensus is to provide an estimate based on a mathematical metric related to the perceived image quality and given a standard predefined set of assumptions about the viewer and the viewing conditions [1]. A reference and a distorted image are often compared to compute an associated image quality prediction based on some perceptual difference metric between the two inputs. The underlying content can be compared using various strategies: mean-squared difference (MSE), signal-to-noise ratio (PSNR), structural similarity (SSIM [2], MS-SSIM [3], IW-SSIM [4], TMQI [5]), naturalness statistics and information content (VIF [6, 7]), contrast visibility (HDR-VDP [8, 9]), low-level feature similarity (FSIM [10]), visual saliency (VSI [11]), and more. Multiple recent attempts have also put machine learning (ML) in the context of IQA [12, 13, 14, 15], a notable example of which is LPIPS [16], a perceptual distance metric achieving state-of-the-art IQA performance.

Although existing computational IQA methods are well correlated with human predictions of image quality, they are typically designed for digital gamma-corrected images assumed to be displayed in ideal viewing conditions. Moreover, the major subjective IQA datasets either use standardized viewing conditions [18] or are collected in crowdsourcing experiments where conditions are not controlled [16, 19, 20]. Real world environments, however, expose the viewer to a wide range of ambient illumination conditions, from nighttime darkness to direct sunlight and everything in between. The same digital content, which when displayed under regular office lighting will have good apparent image quality, in extremely bright or dark conditions, will be less visible to the viewer and its apparent image quality will be reduced [21]. As most existing IQA metrics do not model the effect of non-ideal ambient conditions, they lack the ability to predict the associated effect on image quality.

It is well documented that the Human Visual System (HVS) is characterized by visual adaptation based on ambient light levels. As the perceived illumination level decreases, the HVS shifts from photopic (daytime) to scotopic (nighttime) vision [22, 23]. With this comes a decrease in contrast sensitivity and perception of color; in dark conditions, the human eye thus sees less detail and colors appear desaturated [21, 24]. Meanwhile, at high ambient

[1] Project details can be found at https://github.com/ch-andrei/L-IQA

illumination levels, although the contrast sensitivity of the eye does not vary much, the eye is exposed to stronger reflection and glare from the environment. The perception of visual content is then proportionately adversely affected [21, 25].

Many perceptual phenomena influenced by a variety of viewing conditions can be predicted by Color Appearance Models (CAM) [26]. For instance, CIECAM97 [27] and its later revision CIECAM02 [28] describe the appearance of colored stimuli given the surrounding environment and the ambient illumination, and correlate multiple aspects of color appearance such as brightness, colorfulness, hue, lightness, chroma, and saturation. These methods, however, have several key limitations: they are designed for small patches of color and not complex spatially varying stimuli such as images or video; they lack spatial contrast consideration; and they operate in predominantly photopic conditions (cone-mediated vision). The iCAM06 model [29] partially addresses these issues by considering an extended dynamic range and modelling spatial color appearance parameters, but it is mainly devised for tone-mapping of HDR imagery and not predicting colour differences.

Similarly to CAMs, the luminance retargeting (LRT) algorithm proposed in [25] models the appearance of color and contrast for the full range of real-world luminance values. LRT is presented as a method of simulation of or compensation for the ambient illumination conditions; the algorithm modifies the perceived contrast and colors of an image in an attempt to match the appearance of image content between different luminance levels. This involves finding an optimal tone-curve and spatial contrast processing to account for the changes to eye's contrast sensitivity, as well as modeling of hue and saturation shifts to ensure similar color perception. One practical application of the LRT algorithm is to process images intended to be displayed under non-ideal viewing conditions to match the ideal condition appearance. Such processing techniques are capable of preventing apparent image quality degradation due to unfavorable viewing conditions [21].

The dynamic range and other relevant parameters of the used display system also have an effect on the perception of imagery and hence must be considered when assessing image quality. Plasma (500 cd/m^2) and HDR (3000 cd/m^2) display systems were shown to provide higher overall image quality than the previously typical CRT displays (100 cd/m^2) due to wider luminance range and more accurate color reproduction [30]. On the other hand, many modern consumer display systems implement adaptive screen luminance profiles to reduce energy consumption and lessen eye strain; these systems are equipped with a light sensor and dim the display to better match dark ambient illumination conditions. Decreasing the screen's maximum luminance modifies its dynamic range (the ratio between largest and smallest value) and further reduces the perceived contrast, which often degrades apparent image quality.

In their work on Perceptually Uniform (PU) encoding [17] for luminance signals, Aydin et al. observed that humans tend to rate image quality distortions with the same type and magnitude more harshly when displayed on brighter displays. This is a case of luminance masking: an overall brighter stimuli with higher dynamic range makes the severity of displayed degradations more easily observable. Most popular IQA metrics cannot directly predict this effect; PU encoding was designed as their extension to ensure that "the distortion visibility is approximately uniform along all encoded values". Given display system parameters, gamma-corrected pixel values are converted to physical luminance space and then linearized with respect to human perception. With the PU encoding applied to the inputs, PSNR and SSIM could more accurately predict the change in human quality preferences for brighter displays.

The recent broadcasting industry standard for evaluating the visibility of colour differences between displays is given by the metric ΔI_{ETP} [31]. Suitable for workflows involving display calibration and characterization, this metric is computed using display-referred stimuli (acquired, for instance, using an imaging colorimeter). Color difference visibility between two stimuli is estimated by computing Euclidean distance in the IC_TC_P color space, which is designed as a successor to YC_BC_R to offer a more perceptually uniform color representation with improved decorrelation of saturation, hue, and intensity. The ΔI_{ETP} metric, however, has several drawbacks: it assumes the most sensitive state of adaptation, which ensures that it will not under-predict color difference (but may over-predict them); it requires physically measured (or simulated) display response; and it does not directly consider viewing conditions.

Our current paper explores the influence of various viewing conditions, namely ambient illumination level and display luminance, on human perception of image quality. In Section Subjective IQA Experiments, we present a novel IQA dataset collected during our psychophysical study assessing human perception of image quality as the ambient light level varies from very dark (0 *lux*) to very bright (20000 *lux*). The study allowed us to determine the image quality trend for a wide range of illumination conditions. Additionally, we investigated how the LRT algorithm influences perceived image quality in non-ideal viewing conditions, since conventional IQA methods are incapable of assessing this effect.

Furthermore, we propose an IQA framework that can extend most existing IQA metrics to predict image quality for custom display systems and non-ideal viewing conditions (Section Perceptual IQA Framework). We simulate the physical signal that reaches the observer's eye given display system parameters and ambient illumination level; IQA is then computed using the resulting physical stimuli as opposed to the original gamma-corrected image inputs. We test our framework with multiple prominent IQMs, namely PSNR, SSIM, MS-SSIM, HDR-VDP-2, TMQI, FSIM, VSI, LPIPS, and MDSI [32], and our evaluation demonstrates the effectiveness of our method (Section Results and Discussion).

Subjective IQA Experiments

Quantitatively rating image quality is difficult as it requires training the participants and in general results in higher variance. To avoid this, we employed a pairwise comparison approach, whereby we presented the participants with two image stimuli and tasked them with selecting the one with higher perceived image quality. The two stimuli were physically displayed in two separate environments with potentially different illumination levels and thus the associated effect on perceived image quality could be assessed.

We built a 'light-box' consisting of two isolated compartments each individually illuminated by several remotely controllable 5000 *lumen* LED lights. A single display system was placed inside the box; we used a Samsung TM-800 tablet with a 260 cd/m^2 maximum screen luminance OLED display. As suggested in [25], to minimize the time required for eye adaptation between illumination conditions, our setup implemented a haploscopic sep-

aration of the observer's eyes such that each eye was exposed to a different stimuli and was adapted to a different luminance level. The separation on two environments naturally split the display into two equal halves; lensless binocular goggles were used as a visor through which the viewer inspected the scenes. We only displayed one image stimuli at any given time; the viewer toggled which side displayed the stimuli by tapping on the screen. There was no time limit, and the viewer could inspect both signals interchangeable until they decided which stimuli appeared to have higher image quality. Note that we automatically modified display brightness as per the ambient viewing conditions in the selected environment, and displayed a black image on the other side to avoid confusing the viewer. Further note that we instructed the viewer to keep both eyes open throughout the experiment.

The study was run in a dark room as per the recommendations in [18]. The viewer was given about 10 minutes for the initial adaptation to the dark environment of the room and one minute to adapt to each of the changing illumination levels inside the light-box. The viewing distance was 50 *cm* and the display was angled perpendicularly to the viewing direction. The observers took about 5-10 seconds to pick the preferred image quality winner, and a session lasted 30-60 minutes, depending on the subject's speed. After the first 30 minutes, we initiated a short intermission to avoid potential eye fatigue [18], after which the session continued.

Dataset Description

Our dataset contained 12 reference images classified under three main content categories with distinctive characteristics: i) natural-indoors (6 images), ii) social media (3 images), and iii) automotive (3 images). Separation into categories was motivated by our goal of evaluating image quality for several different real-world applications of physical display systems. The images had 3:4 aspect ratio, which was required to display several images simultaneously on the same screen.

Since human perception of light intensity is better approximated on a logarithmic scale, we considered illumination levels that roughly double at each step, namely: 0, 20, 50, 100, 250, 500, 1000, 2500, 5000, 10000, 20000 *lux*, for a total of 11 conditions. In addition to varying the ambient illumination level, we also included a second dimension in our comparison space. We used the luminance retargeting (LRT) image processing algorithm proposed in [25] as a means of digital compensation for the ambient illumination level. We wanted to evaluate the effectiveness of LRT

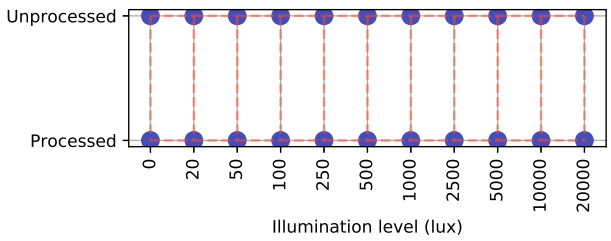

Figure 2. *The comparison space of our experiment. Blue dots are the conditions, red links imply adjacency and thus a comparison in the context of our study. For an incomplete study design, given 11 illumination levels and two image processing types (22 conditions in total), there are 31 adjacent condition pairs, which corresponds to 31 required comparisons.*

processing as well as to verify if the objective IQA metrics were capable of accurately predicting the associated quality difference. In our case, we applied LRT processing[2] to match the perception of images displayed in the 11 selected conditions to the assumed ideal condition.

Our comparison space is thus formulated with two dimensions: illumination level (in *lux*) and image processing type (unprocessed, processed). For 11 illumination level values and two processing types, we thus have a total of 22 *conditions*, as depicted in Figure 2. Although a full study would test all combinations, this is impractical; as per the suggestions in [33], we ran an incomplete study design, omitting comparisons between non-adjacent conditions. The study was run with 25 observers and a total of over 10000 image quality preference judgments were collected, with approximately 400 judgments carried out by each observer.

Acquiring Quality Scores

In order to obtain relative IQA scores from the pairwise comparisons, we used the *pwcmp* scaling software [34] to perform psychometric re-scaling of the collected data. This approach is warranted by the analysis done in [33], which shows that pairwise comparison data can be effectively reinterpreted via psychometric re-scaling to obtain quality scores with a strong linear relation to Mean Opinion Score (MOS). The *pwcmp* algorithm builds a statistical model treating preference judgments as noisy samples of an underlying quality difference distributions, and transforms these into relative quality scores $\hat{q}$ in Just-Objectionable-Differences (JOD) units. JOD is a unit of probabilistic measure of preference, similar to the difference-mean-opinion-score (DMOS), which quantifies the difference between a test signal and some reference. A difference of 1 JOD unit implies that 75% of the observers prefer one stimuli over the other; increasing difference in JODs represents asymptotically higher probability of preference.

We applied the *pwcmp* scaling algorithm on a per-image basis, i.e. our data is treated and scaled separately for each of the 12 images in our dataset. This essentially lowers the accuracy of each scaling, as it limits the amount of data each computation operates with, but better captures content dependency of human IQA, as each image is treated separately and thus no averaging across content occurs. For illustrative purposes, instead of presenting 12 different quality trends, we also applied *pwcmp* on the entirety of our data as depicted in Figure 3. While specific content-dependent image quality particularities are not observed in the combined plot, this illustration is a good summary of the results of our experiment.

For all images, humans give preference to the illumination range near the "ideal" conditions around 200-600 lux, where the display luminance is near its maximum and the effects of environmental glare and reflection are not as strong; image quality falls off as the illumination level deviates from the "ideal". Contrast distortions seem to be the most prevalent source of image quality degradation as ambient illumination level differs from the ideal. Under dim illumination, lower display luminance and the visual adaptation of the HVS result in decreased perceived contrast and color sensitivity of the eye; in bright conditions, ambient light reflection lower the physical contrast of the observed image. In this way, non-ideal viewing conditions result in poor visibility and

[2] We used an existing implementation of the LRT algorithm described in [21].

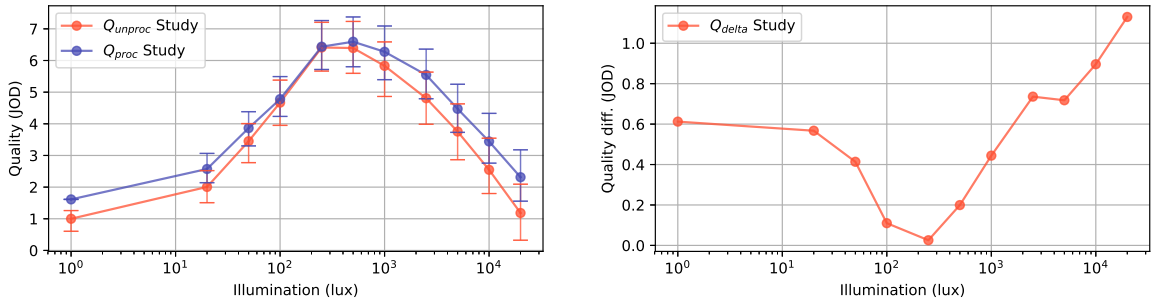

Figure 3. *Image quality trends obtained by psychometric scaling of the pairwise comparison data acquired in our study. On the left, quality trend for unprocessed and processed images, Q_{unproc} and Q_{proc}, respectively; on the right, the difference in quality between unprocessed and processed, $Q_{delta} = Q_{proc} - Q_{unproc}$. Error bars represent confidence intervals as reported by the scaling software.*

a decrease of the perceived image quality. Lastly, our subjects perceived the quality of processed images to be higher, thus Q_{delta} is nearly always positive. Note that there is essentially no visual difference due to LRT processing around the ideal conditions as that illumination range is the target of compensation (in theory, this implies that Q_{delta} should be minimal for that illumination, which is captured in our results).

Perceptual IQA Framework

We propose a generic IQA framework (see Figure 1 for a block-diagram) that can extend most existing IQA metrics to a wider range of luminance values, supporting a variety of display systems and illumination conditions. We accomplish this by incorporating a display and degradation simulation in the IQA pipeline to estimate the visual stimuli that physically reaches the viewer's eye. A physical comparison space for the reference and the test images involved in IQA is depicted in Figure 4. In order to assess the image quality degradation associated with displaying an image in particular viewing conditions, i.e. including the adversary effect of non-ideal illumination level, we estimate the difference between the reference image displayed in *ideal* conditions and the reference (or test) image content displayed in *test* conditions. While our current work mainly focuses on the comparison shown in red, Figure 4 presents other image pairs for which IQA can be computed. Each pair has a different interpretation for IQA since the reference and test images are defined differently; for instance, both the reference and the test images can be simulated in equivalent viewing conditions.

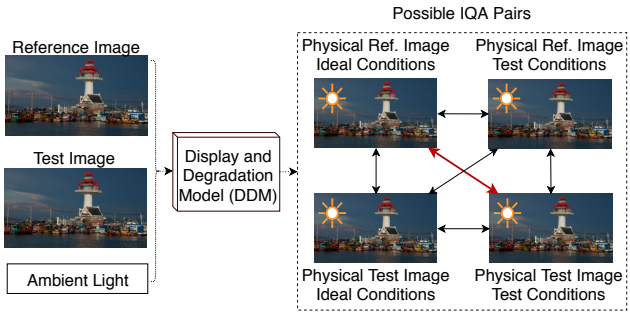

Figure 4. *Different combinations of physical image pairs result in different IQA interpretations. In the above, six physical image pairs are generated by our DDM; we focus on the comparison shown in red, where a physical reference image simulated in ideal conditions is compared to the physical reference or test images simulated in test conditions. Comparing these pairs will assess the perceptual difference between how the image should appear in ideal conditions and how it truly appears in non-ideal conditions; the associated image quality degradation can thus be evaluated.*

Display and Degradation Simulation

Our simulation approach closely follows the Gamma-Gain-Offset [35] display model with an extension accounting for ambient light as described in [36]; we will refer to this as Display and Degradation Model (DDM). We simulate adaptive brightness of the display by controlling the maximum luminance parameter according to the ambient illumination level. Given an input image as gamma-corrected pixel values, the parameters of the display, and the ambient illumination level, we simulate the signal that reaches the observer's eye as physical luminance in photometric units of cd/m^2 as per Equation 1:

$$L(V) = (L_{max} - L_{blk})V^{\gamma} + L_{blk} + L_{refl} \quad (1)$$

In the above, L_{max} corresponds to the maximum display luminance in cd/m^2, L_{blk} is the display's black level luminance in cd/m^2, V is the input signal luma in the range 0-1, and L_{refl} is the reflected luminance approximated as shown in Equation 2, where k is the reflectivity factor (typically 0.01-0.05 for common displays) and E_{amb} is the ambient illumination level in lux. Our Display and Degradation Model (DDM) thus consists of two components: i) an internal change in the display's light emission parameters via an adaptive screen dimming profile controlled by L_{max}, and ii) an external influence via injected reflection in the form of L_{refl}.

$$L_{refl} = \frac{k}{\pi} E_{amb} \quad (2)$$

We simulate adaptive brightness by varying maximum screen luminance L_{max} as a function of ambient illumination conditions and as per device characteristics. Various devices have different schemes and displays; such profiles can be empirically determined or specified by the manufacturer. For example, a typical mobile device display can operate at L_{max} between 2 cd/m^2 in dark ambiance

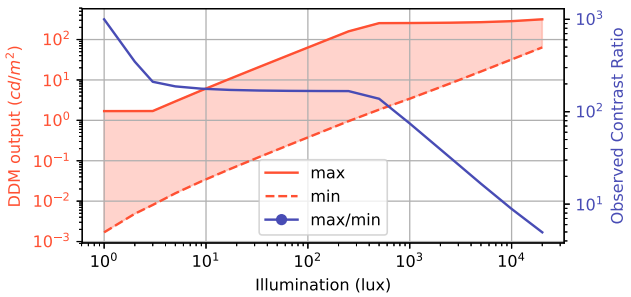

Figure 5. *Display and degradation model (DDM) output range in varying illumination conditions for a display with L_{max} ranging between 2 to 400 cd/m^2, physical contrast ratio of 1000:1, and some sample image content. In red, we show the range for minimum and maximum output values; in blue, we plot the associated observed contrast ratio (maximum / minimum). The change in the range is due to i) display dimming, and ii) ambient reflection; these two effects vary with ambient illumination level. The same digital image content will be reproduced and observed differently based on the ambient conditions.*

to 400 cd/m^2 in bright conditions, while its L_{blk} is determined according to the desired physical contrast ratio. As a result of adaptive brightness, the physical dynamic range of the images that are displayed in significantly different illumination conditions often is largely distinct. For instance, when comparing conditions of 0 lux versus 1000 lux, the resulting display simulations will produce values with a difference of several orders of magnitude (see Figure 5 for a graphical illustration). This is even more pronounced for HDR displays with larger L_{max}.

Display and Illumination Aware IQA

We begin with the assumption that both images are displayed on the same screen, but in different illumination conditions. For most common displays, the display brightness is typically nearing its maximum for ambient conditions above 500 lux; this illumination level nears the typical recommendation for office room lighting and is usually considered to be the *ideal* viewing condition. As such, we simulate the physical reference image displayed in ideal viewing conditions with no degradations: we approximate this by computing the DDM (see Equation 1) with maximum display brightness (given display parameters) and minimal environmental reflection (L_{refl} is set to 0, or simulated normally for the ideal illumination level). The physical test image, on the other hand, is simulated using L_{max} and L_{refl} terms computed as per the queried illumination level.

In the quality assessment stage that follows, a given IQA metric is used on the PU-encoded DDM outputs to provide the final image quality prediction. It must be emphasized that PU encoding is a critical component of our model, as it linearizes the physical luminance values from our DDM with respect to human perception and accounts for luminance masking. PU encoding was designed to account for luminance masking as signal strength varies; we exploit this to allow IQA metrics to operate on a wider range of luminance values.

Lastly, since our DDM converts RGB inputs to the luminance domain, the inputs to the final IQA stage are "grayscale" and do not have a color component - for metrics enforcing RGB inputs, we stack the PU-encoded luminance in three channels. While this results in a possible loss of overall performance for metrics that rely on chromaticity, we do not, as of writing this, have a proven combination of the display model and PU encoding that operates with RGB channels separately.

Results and Discussion

We evaluated our proposed IQA framework in combination with various prominent IQA metrics, namely PSNR, SSIM, MS-SSIM, HDR-VDP-2, TMQI, LPIPS, FSIM, VSI, and MDSI. The performance was validated against our own subjective study. We used four common performance metrics to evaluate the predictions of our model. To assess the level of correlation between subjective and predicted IQA, we computed Spearman rank-order correlation coefficient (SROCC) and Kendall rank-order correlation coefficient (KROCC). We also evaluated linear correlation between the two trends by computing Pearson linear correlation coefficient (PLCC) and root mean squared error (RMSE). Note that, as recommended in [37], we pass the scores through a logistic non-linearity before computing PLCC and RMSE.

For a more fair comparison, we employed five fold cross-validation across the available data. The logarithmic fitting was optimized using the training set, while the performance scores were computed on the test sets after logarithmic remapping was applied. This procedure was performed 100 times to minimize bias from randomized splitting of the data. The results across runs were then averaged (see Table 1); Figure 6 shows the tested IQA metrics ranked by their performance. We observe that many of the common IQA metrics can be effectively extended to operate in a wider range of luminance values using our proposed perceptual IQA framework. Strong correlations between the data from our subjective study and our model's prediction are observed across the entirety of the considered illumination range. Among the tested metrics, we notice that LPIPS, HDR-VDP-2, FSIM, VSI, and MDSI are the most effective at predicting the overall quality trend. Furthermore, the aforementioned assess the difference between content - the difference in image quality between unprocessed and processed images - the most effectively. While the overall trend along the considered illumination range for most metrics matches well with the human predictions, the difference between processed and unprocessed content is often not accurately predicted. For instance, SSIM and PSNR yield very poor performance for this type of image difference. Curiously, TMQI fails for the overall trend but is accurate in predicting content difference.

Conclusion

In this paper, we investigated the effect of ambient illumination level on human perception of image quality and introduced a new IQA dataset consisting of human IQA in various non-ideal illumination conditions, ranging from very dark to bright bright. We also proposed a novel perceptual framework for image quality assessment, extending most common IQA metrics to non-ideal illumination levels and to a wider luminance range than originally intended. Our approach is based on simulating the physical visual signal that reaches the human eye and performing IQA on the resulting stimuli as opposed to the originally strictly digital content; we implemented a display and degradation simulation and thus modelled the adverse effect associated with physically displaying visual content in non-ideal viewing conditions. We evaluated mul-

Table 1: Performance evaluation of our IQA framework with various popular IQA metrics on our subjective study. The scores represent the average of 100 runs of randomized five-fold cross-validation. Correlation and goodness of fit measures between subjective and objective IQA are shown for a) the overall quality trends ($Q_{unproc.}$ and $Q_{proc.}$), b) IQA prediction difference (Q_{delta}) between unprocessed and processed images. Best scores are emphasized in bold.

Label	Metric	PSNR	SSIM	MSSSIM	TMQI	HDRVDP2	LPIPS	FSIM	VSI	MDSI
a) Quality Trend	SROCC	0.6846	0.6291	0.8695	0.4857	0.8638	0.9057	0.9071	0.9211	**0.9233**
	KROCC	0.4927	0.4507	0.6829	0.3444	0.6716	0.7325	0.7351	0.7596	**0.7633**
	PLCC	0.6905	0.6547	0.8806	0.4424	0.8712	0.9124	0.9117	0.9267	**0.9291**
	RMSE	0.2332	0.2583	0.1494	0.3038	0.1536	0.1289	0.1329	0.1257	**0.1162**
b) Quality Delta	SROCC	0.0243	0.0467	0.4922	0.6109	0.5358	0.5435	**0.6245**	0.5819	0.5923
	KROCC	0.0126	0.0285	0.3573	0.4390	0.3826	0.3958	**0.4534**	0.4188	0.4232
	PLCC	0.0230	0.0455	0.3667	0.5739	0.5207	0.4929	**0.5787**	0.5496	0.5660
	RMSE	0.4006	0.4679	0.3639	0.2899	0.2841	0.3221	0.3039	0.2875	**0.2730**

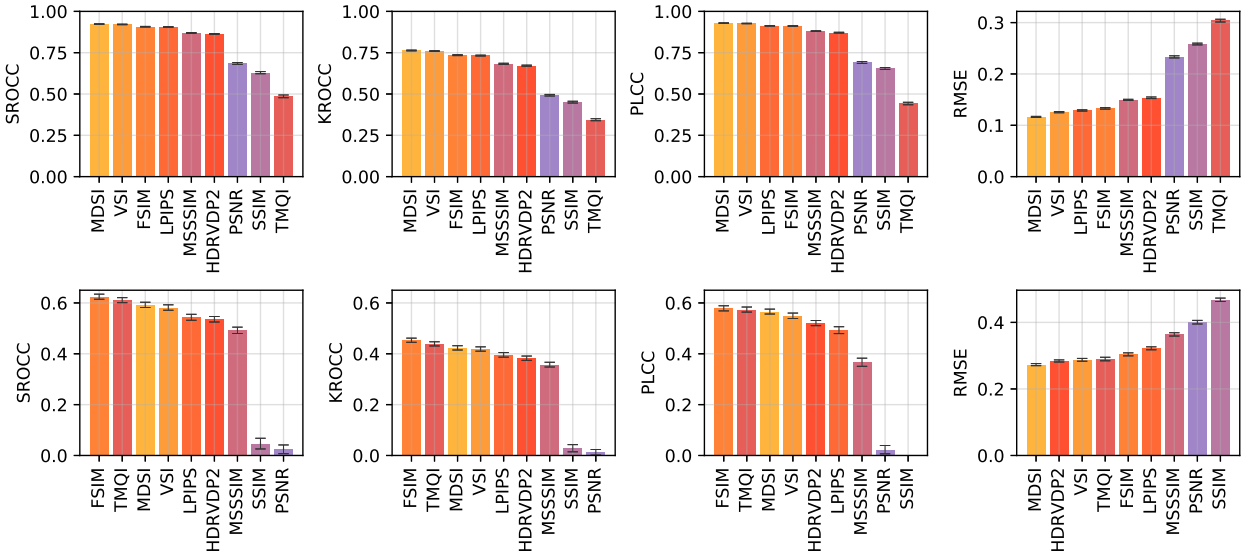

Figure 6. *Tested IQA metrics ranked by correlation and goodness of fit measures between subjective and objective IQA for the overall quality trends ($Q_{unproc.}$ and $Q_{proc.}$; top row) and IQA prediction difference (Q_{delta}) between unprocessed and processed images (bottom row). The scores represent the average of 100 runs of five-fold cross-validation results and the error bars depict the associated standard error. Note that the IQMs are color coded.*

tiple IQA metrics in combination with our model and determined that the corresponding IQA predictions strongly correlate with human judgments of image quality across a wide spectrum of illumination conditions and for multiple display configurations. Our framework can be extended to more applications, for instance more complex comparison schemes in the proposed physical photometric space. The individual components of the framework can also be customized; any IQA metric can be plugged in and its performance verified against our dataset; a more advanced simulation of the display and degradation can also be implemented. A version of our framework supporting chromaticity is also desirable as many modern IQA methods rely on color information.

References

[1] Z. Wang and A. Bovik. *Modern Image Quality Assessment*. Morgan & Claypool Publishers, 1st edition, 2006.

[2] 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*, 13(4):600–612, April 2004.

[3] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, volume 2, pages 1398–1402 Vol.2, Nov 2003.

[4] Z. Wang and Q. Li. Information content weighting for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 20(5):1185–1198, May 2011.

[5] H. Yeganeh and Z. Wang. Objective quality assessment of tone-mapped images. *IEEE Transactions on Image Processing*, 22(2):657–667, Feb 2013.

[6] H. R. Sheikh, A. C. Bovik, and G. de Veciana. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Transactions on Image Processing*, 14(12):2117–2128, Dec 2005.

[7] H. R. Sheikh and A. C. Bovik. Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2):430–444, Feb 2006.

[8] R.K. Mantiuk, S. J. Daly, K. Myszkowski, and H.-P. Seidel. Predicting visible differences in high dynamic range images: model and its calibration, 2005.

[9] R. K. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich. HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Trans. Graph.*, 30(4):40:1–40:14, July 2011.

[10] L. Zhang, L. Zhang, X. Mou, and D. Zhang. FSIM: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, Aug 2011.

[11] L. Zhang, Y. Shen, and H. Li. VSI: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 23(10):4270–4281, Oct 2014.

[12] S. Bosse, D. Maniry, T. Wiegand, and W. Samek. A deep neural network for image quality assessment. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3773–3777, Sep. 2016.

[13] W. Xue, L. Zhang, and X. Mou. Learning without human scores for blind image quality assessment. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 995–1002, June 2013.

[14] K.-Y. Lin and G. Wang. Hallucinated-IQA: No-reference image quality assessment via adversarial learning. *CoRR*, abs/1804.01681, 2018.

[15] H. T. Esfandarani and P. Milanfar. NIMA: neural image assessment. *CoRR*, abs/1709.05424, 2017.

[16] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, June 2018.

[17] T. O. Aydın, R. K. Mantiuk, and H.-P. Seidel. Extending quality metrics to full luminance range images. In *Human Vision and Electronic Imaging*, pages 68060B–10. Spie, 2008.

[18] Recommendation 500-13: Methodology for the subjective assessment of the quality of television pictures. ITU-R Rec. BT.500, 2012.

[19] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, M. Carli, and F. Battisti. TID2008 - a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, 10:30–45, 01 2009.

[20] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. Jay Kuo. Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30:57 – 77, 2015.

[21] T. Akhavan, H. Yoo, and A. Chubarau. Solving challenges and improving the performance of automotive displays. *Information Display*, 35(1):13–27, 2019.

[22] B. A. Wandell. *Foundations of Vision*. Sinauer Associates, 1995.

[23] D. Cao, J. Pokorny, V. C. Smith, and A. J. Zele. Rod contributions to color perception: Linear with rod contrast. *Vision Research*, 48:2586–2592, 11 2008.

[24] P. G. J. Barten. Contrast sensitivity of the human eye and its effects on image quality. 1999.

[25] R. Wanat and R. K. Mantiuk. Simulating and compensating changes in appearance between day and night vision. *ACM Trans. Graph.*, 33(4):147:1–147:12, July 2014.

[26] M.D. Fairchild. *Color Appearance Models*. The Wiley-IS&T Series in Imaging Science and Technology. Wiley, 2013.

[27] The CIE 1997 interim colour appearance model (simple version), CIECAM97s, cie 131-1998. *Color Research & Application*, 23(6):431–431, 1998.

[28] N. Moroney, M. Fairchild, R. Hunt, L. Changjun, M. Luo, and T. Newman. The CIECAM02 color appearance model. volume 10, pages 23–27, 01 2002.

[29] J. Kuang, G. M. Johnson, and M. D. Fairchild. iCAM06: A refined image appearance model for HDR image rendering. *J. Visual Communication and Image Representation*, 18(5):406–414, 2007.

[30] A. Choudhury and S. Daly. HDR display quality evaluation by incorporating perceptual component models into a machine learning framework. *Signal Processing: Image Communication*, 74:201 – 217, 2019.

[31] Recommendation 2124-0: Objective metric for the assessment of the potential visibility of colour differences in television. ITU-R Rec. BT.2124, 2019.

[32] H. Z. Nafchi, A. Shahkolaei, R. Hedjam, and M/ Cheriet. Mean deviation similarity index: Efficient and reliable full-reference image quality evaluator. *CoRR*, abs/1608.07433, 2016.

[33] E. Zerman, V. Hulusic, G. Valenzise, R. K. Mantiuk, and F. Dufaux. The relation between MOS and pairwise comparisons and the importance of cross-content comparisons. 01 2018.

[34] M. Pérez-Ortiz and R. K. Mantiuk. A practical guide and software for analysing pairwise comparison experiments. 12 2017.

[35] R. S. Berns. Methods for characterizing CRT displays. *Displays*, 16(4):173 – 182, 1996. To Achieve WYSIWYG Colour.

[36] R. K. Mantiuk, K. Myszkowski, and H.-P. Seidel. *High Dynamic Range Imaging*, pages 1–42. American Cancer Society, 2015.

[37] H. R. Sheikh, M. F. Sabir, and A. C. Bovik. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15(11):3440–3451, Nov 2006.

Author Biography

Andrei Chubarau holds a bachelor's degree in computer and software engineering and is currently a second year PhD student at McGill University. His research interests include image quality assessment, visual perception, and computational vision.

Tara Akhavan is the founder and CTO of IRYStec Software Inc. She holds a bachelor's degree in computer engineering, a master's degree in artificial intelligence, and a Ph.D. in image processing and computer vision from the Vienna University of Technology in Austria. Akhavan is vice chair of marketing for the Society of Information Display.

Hyunjin Yoo is a senior research engineer and team lead at IRYStec Software Inc. She received an M.S. in information and communications and a Ph.D. in information and mechatronics from Gwangju Institute of Science and Tech-nology in Gwangju, South Korea.

Rafał K. Mantiuk is Reader (Associate Professor) at the Department of Computer Science and Technology, University of Cambridge (UK). He received PhD from the Max-Planck-Institute for Computer Science (Germany). His recent interests focus on computational displays, novel display technologies, rendering and imaging algorithms that adapt to human visual performance and viewing conditions in order to deliver the best images given limited resources.

James Clark received his BASc and PhD in Electrical Engineering from the University of British Columbia (1980,85). From 1985-1994 he was with the Division of Applied Sciences at Harvard University, and from 1994-96 was a visiting scientist at Nissan Cambridge Basic Research. Since then he has been with the department of Electrical and Computer Engineering at McGill University. His work has focused on computer vision and image processing. He is a senior member of IEEE.

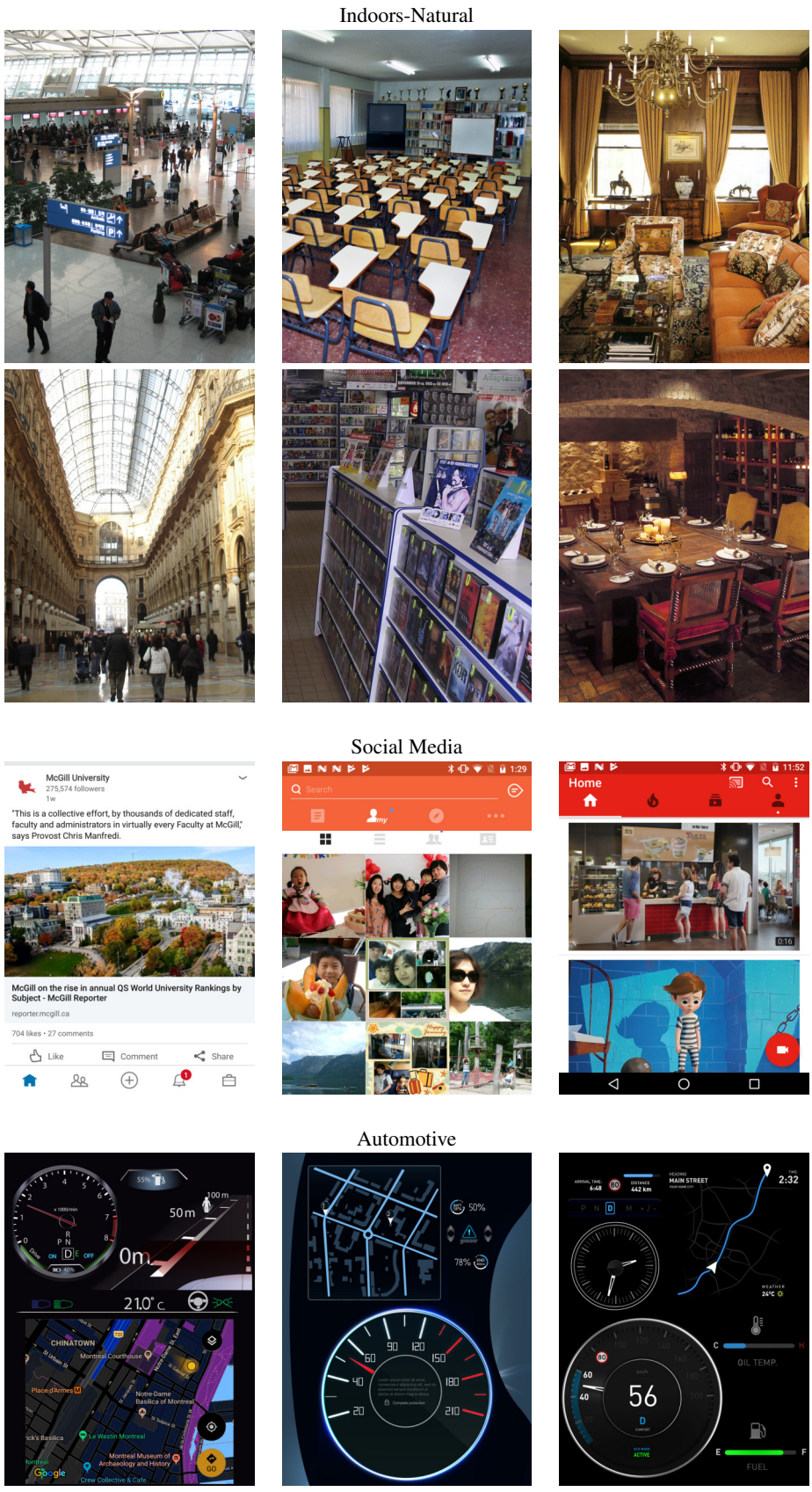

Figure 7. *The reference images that were used in our subjective study. The dataset contains 12 images split across three categories: Natural-Indoors, Social Media, and Automotive.*