Studies on the effect of MegaPixel sensor resolution on displayed image quality and relevant metrics

Sophie Triantaphillidou[1], Jan Smejkal[1], Edward W. S. Fry[1], and Chuang Hsin Hung [2]; [1]University of Westminster UK and [2]Huawei China

Abstract

This paper investigates camera phone image quality, namely the effect of sensor megapixel (MP) resolution on the perceived quality of images displayed at full size on high-quality desktop displays. For the purpose, we use images from simulated cameras with different sensor MP resolutions. We employ methods recommended in the IEEE 1858 Camera Phone Image Quality (CPIQ) standard, as well as other established psychophysical paradigms, to obtain subjective image quality ratings for systems with varying MP resolution from large numbers of observers. These are subsequently used to validate image quality metrics (IQMs) relating to sharpness and resolution, including those from the CPIQ standard. Further, we define acceptable levels of quality - when changing MP resolution - for mobile phone images in Subjective Quality Scale (SQS) units. Finally, we map SQS levels to categories obtained from star-rating experiments (commonly used to rate consumer experience). Our findings draw a relationship between the MP resolution of the camera sensor and the LCD device. The chosen metrics predict quality accurately, but only the metrics proposed by CPIQ return results in calibrated JNDs in quality. We close by discussing the appropriateness of star-rating experiments for the purpose of measuring subjective image quality and metric validation.

Introduction

According to the Camera Phone Image Quality (CPIQ) working group [1], "consumers most often use the [sensor megapixel] MP count as a way to evaluate the camera quality of their mobile devices, however, there are many other factors that influence perceived mobile camera image quality". The image quality attributes most affected by changes in MP resolution are visual resolution - concerned with the visibility of fine detail, and sharpness - concerned with visual definition of edges and texture [2]. Although it is well established that, increasing camera MP resolution contributes, generally, to sharper displayed images, it is well known to camera experts that optics, other sensor parameters (notably noise) and camera image signal processing (ISP, e.g. de-noising, sharpening) also affect considerably both resolution and sharpness. Evidently, the dynamic range, luminance and spatial characteristics of the display system presenting the digital image play a significant role in its perceived sharpness, along with the viewing conditions (observer acuity, viewing field, viewing distance). It is therefore useful to consider the camera-processing-display-observer system in the modeling of image's resolution and sharpness and to calculate relevant metrics in the spatial frequency domain at the plane of the observer's eye.

Given the effect of camera MP count on visual resolution and sharpness, and the camera-processing-display-observer imaging chain, the following questions are considered in this paper: i) what is the effect that increasing MP resolution has on the quality of pictures viewed on common LCD displays? ii) How many pixels are enough for the picture to be deemed by consumers of acceptable displayed quality for camera phone imaging? In this paper we investigated these questions by collecting psychophysical data from images originating from different resolution (simulated) sensors, using different psychophysical paradigms. We also validated relevant image quality metrics (IQMs).

There are several IQMs designed to predict perceived resolution and sharpness. Those employed in our work are MTF50 [3], Subjective Quality Factor (SQF) [4], Acutance [5], and the CPIQ Quality Loss (QL) [5]. The former is, strictly speaking, a camera performance metric that relates to sharpness, whereas the latter three account for the display's spatial properties, as well as the visual system's effects by implementing a model of the human Contrast Sensitivity Function (CSF) [6].

Validation of IQMs requires the collection of visual data from carefully designed and conducted psychophysical studies. The CPIQ IEEE 1858 standard [5] recommends the ISO 20462-3:2012 soft copy image quality ruler [7] for the purpose. It returns quality ratings in Subjective Quality Scale (SQS) units, separated by Just-Noticeable-Differences (JND) in quality. Often categorical scaling [8] is employed for rating the consumer experience (e.g. star rating experiments). Star-rating experiments are favoured by the consumer industries because they are uncomplicated and provide quick data collection. However, unless observers are given anchored points and precise instructions, and then data are analyzed using established psychophysical laws, category scales are not calibrated in equal intervals and thus correlations with metric results are not meaningful. In this work we implement both the image quality ruler and star-rating paradigms for deriving subjective scales and examine the relationship between their outcomes.

Threshold experiments [8] are commonly used to define limits of perceptibility of image artefacts/changes, or acceptability of quality (or its attributes). The psychometric function is implemented for the purpose [9]. We used threshold experiments to define limits of acceptable mobile camera phone image quality.

In summary, the objectives of the study are:

   i. to investigate the perceived quality of images from different MP sensors, when they are displayed at *full size* on very high quality desktop displays;

   ii. to validate relevant state-of-the-art relevant imaging performance and image quality metrics for the purpose;

   iii. to examine the relationship between MP camera resolution and display resolution;

   iv. to obtain acceptable limits of image quality in SQS units;

   v. to investigate and discuss the suitability of star rating experiments for obtaining meaningful image quality ratings.

The remaining sections of paper are structured as follows: the development of the test image stimuli is first presented; subjective

(i.e. psychophysical studies) and objective (i.e. metric calculation) methodologies follow; we continue by outlining our results and we close with conclusions.

Development of test images

Scene capture

A large number of test scenes (>50) were originally captured using the very high quality digital camera system: a PhaseOne IQ3 100 MP medium format digital back, mounted on an PhaseOne XF camera body, equipped with Schneider Kreuznach Blue Ring lenses (4 in total: 35mm f/3.5, 80mm f/3.5, 150mm f/3.5, 240mm f/4.5). The camera's sensor size is 53.7 x 40.4 mm, with digital resolution of 11608 by 8708 pixels (aspect ratio 4:3) and active pixel size of 4.6 by 4.6 microns. Scene contents and illuminations were based on the ISO 20462-3:2012 [7] recommendations and the IEEE P1858 CPIQ Standard validation study [10]. Scene lighting conditions were as follows:

- Daylight: > 1000 lux (full daylight, overcast, in shadow)
- Daylight & fluorescent mix: 250>x>500 lux
- Incandescent: 20>x>60 lux
- Moonlight & various artificial low-level color illuminations: ~5>x>~0.1 lux

Images were captured in RAW format at 16-bit per channel. In-camera sharpening and noise reduction were turned off. They were accessed using the Photoshop CC 2018 Camera RAW converter, operating also with noise removal and sharpening turned off, and saved as uncompressed RGB TIFF files, with linear luminance and sRGB chromaticities [11]. Following inspection and without any additional manipulation, images of 14 scenes were selected for further experimentation. These comprised of a mixture of landscapes, cityscapes, indoor environments, outdoor and indoor portraits, groups of people, and night city scenes. Appendix A includes thumbnails of the selected test scenes.

Simulation of different MP sensors

Original captures were decimated to produce simulated outputs from eight different camera sensors, varying only in MP resolution. Figure 1 describes the processing pipeline used to produce the outputs. The processes below were implemented in MATLAB^{TM}, as follows: Linear filtering was achieved in linear sRGB space using 31x31 finite impulse response (FIR) filters. The filters' frequency response was based on the SFR of the lens-camera systems (see Objective Evaluations section). The filters were designed so that the response at the Nyquist frequency of the target sensor was approximately 0.20, which matched the mean frequency responses of PhaseOne and lenses. The filter process prevented aliasing. The filters were designed so that the response at the Nyquist frequency of the target sensor was approximately 0.20, which matched the (average) combined responses of PhaseOne and lenses at the sensor's Nyquist frequency, and also prevented aliasing.

Pixel averaging in both orientations (2x2, 3x3, 4x4, 6x6, etc.) resulted to image sizes of 25, 11, 6, 3, 1.2, 0.53, 0.34 and 0.1 MP resolution. Including the original image of 100 MP, each scene was represented at 9 different MP resolutions. The total number of test images was 126.

It should be noted that, in real camera systems, varying the MP resolution on a constant senor size results not only to different pixel sizes, but often also different sensitivities, noise levels, etc., which in turn require different camera ISP tuning. Given that the focus of our research was on the perceived image quality as a function of digital image resolution, it was decided not to include noise and ISP in our simulations. This choice is not representative of real camera pipelines, but it did allow investigating displayed image quality solely as a function of changes in the single sensor parameter of interest.
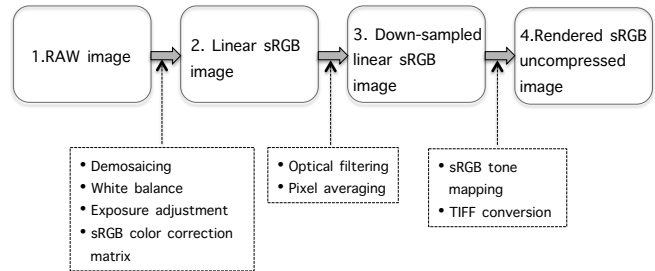

Figure 1: Simulation pipeline

Output for a reference display

Visual experiments were designed to test perceived image quality of different digital resolution images when they were displayed at full size - which is a very common scenario in mobile and consumer photography (i.e. not at a 1:1 camera-to-display pixel resolution, where only a crop of the image is shown, as in previous experiments relating to CPIQ metric validations [10]). Image interpolation was thus necessary to downsize or upsize images to the desired size for matching the display pixel resolution. Clearly, this very common operation, rendering the interpolation method an important component of the imaging chain.

Image resizing to the display resolution was achieved in MATLAB^{TM}, using *lanczos3* interpolation [12]. Since all interpolation algorithms introduce some distortion to the original image information, this choice was based on maximum frequency content preservation [13] (especially of mid and high frequencies which are important to sharpness and thus quality [2]), as well as cohesion of the algorithm with interpolations methods used in contemporary mobile phone cameras.

Figure 2 illustrates the measured Spatial Frequency Responses (SFR) of 3 common interpolation techniques, bicubic, bilinear and lanczos3, for images interpolated down from 100 MP to 1.2 MP; it indicates the clear benefits of using the lanczos3 algorithm. SFR of interpolating up from lower MP counts to the display resolution indicated similar trends, but the SFR differences between the algorithms were less noticeable.

Images originally decimated down to 1.2 MP resolution did not require any interpolation; they mapped 1:1 the display resolution. sRGB tone mapping was applied for output to a standard (reference) sRGB display. All test images were finally centrally cropped to 1089 x 1089 pixels. These pixel dimensions allowed for two images, if necessary, to fit at full size on one test display, side by side, while leaving boarders around them, which conformed with the recommended graphic user interface (GUI) of the image quality ruler paradigm [7].
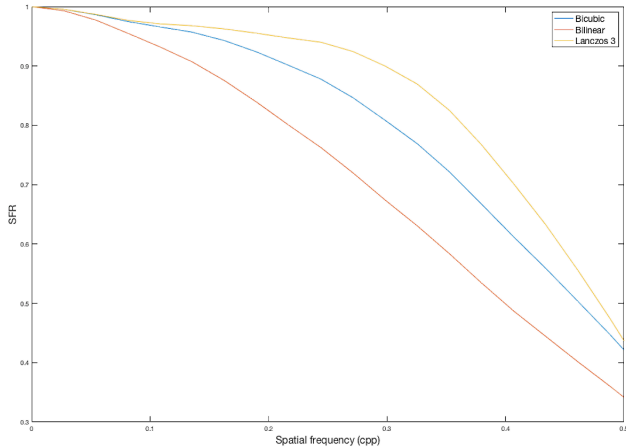
Figure 2: SFR of different interpolation algorithms: original image 100 MP interpolated down to 1.2 MP

Subjective evaluations

Psychophysical evaluations investigating the quality of the test images were carried out using three different experimental paradigms: the soft copy image quality ruler [7], categorical scaling [8, 9] and threshold experiments of acceptability in terms of quality [8, 9]. In all experimental paradigms the test images were displayed at full size.

Reference display, test viewing conditions and observers

The apparatus and setup were the same in all visual image evaluations. Two identical new 27-inch EIZO CG277 LCDs, with resolution 2560 x 1440 pixels, active area 596.7 x 335.6 mm, pixel size 0.233mm, aspect ratio 16:9 and maximum refresh rate 64Hz were set up in the same fashion in a dedicated visual laboratory. These high-quality, wide-gamut displays have a self-calibration sensor that delivers very accurate automated calibration, rendering them suitable for visual experiments. Although the native display white point is 300 cd/m^2, they were calibrated (daily) to sRGB color space with a white point luminance of 100 cd/m^2. Compliance with the sRGB transfer functions and chromaticities [11] was confirmed using a Konica Minolta CS200 chroma meter. Fluorescent tubes, covered by white diffuser screens to prevent direct reflections, provided 30-lux illuminance at CCT D60.

The psychophysical interfaces for all three experiments were designed in MATLAB^{TM}, in accordance with relevant recommendations [7]. They ran on Microsoft Windows®. Appendix B illustrates the 3 different experimental GUIs.

A minimum of 20 observers (and up to 33) took part in each experiment, mainly university students, male and female from several ethnical backgrounds. Their age ranged between 21 and 36 years old. To promote observer engagement with the experimental tasks, when sessions were concluded, we promised that upon completion, we would informed them of their individual results relatively to the mean ratings. We also offered purchasing vouchers to attract large numbers of observers.

Observers always sat at a viewing distance of 600mm from the display faceplate, maintained with the use of a chin rest. Images were subtended at 30 deg x 30 deg. The viewing distance complied with the 20462-3 recommendations (i.e. minimum of 2500 x display pixel pitch). Observer visual acuity was examined with the use of a calibrated for the distance Snellen visual acuity

chart (20/20 or corrected to 20/20) [14, 10]. Color deficiencies were not examined.

Observers were initially given written instructions. They were then instructed on how to run the test verbally. Depending on the experiment type, they were given a 5- (min) to 15- (max) minute supervised training before their ratings were recorded. Experimental sessions lasted between 20 (minimum) and 45 (maximum) minutes. Two sessions per day were allowed per observer, separated by a gap of at least one-hour.

Soft copy image quality ruler experiments

The ISO 20462-3:2012 soft copy image quality ruler [7] is the psychophysical method recommended by the ISO Image Quality Standards committee, and the CPIQ IEEE Standard 1858 for collecting image quality ratings [5, 7]. Results from the method are reported using the SQS, a numerical scale that, when it is anchored against physical standards, has a zero point and one unit corresponding to 1 JND in quality. When the SQS is not anchored, results are reported in SQS$_2$, a "floating" interval scale [8], with intervals equal to JNDs in quality. A validation of the ruler paradigm is provided in [15].

Development of standard reference stimuli

The standard provides standard reference stimuli (SRS) for different scene types, but also describes how experimenters can generate their own SRS sets. Each SRS set comprises of a series of digital images (ruler images), depicting the same scene, but varying in one single image quality attribute – sharpness. Each SRS corresponds to one SQS value, with SQS values ranging from SQS 0 (min quality) to SQS 31 (max quality). The MTF of each SRS conforms to the shape of a monochromatic MTF of an on-axis diffraction-limited lens (DLL) [7].

We generated our own reference stimuli sets, by accounting for the capture-processing system, the reference display characteristics and the viewing distance as described earlier, and by following procedures indicated in the standard and in [15]. These are summarized here:

- The imaging system MTF, $MTF(u)_{system}$, was obtained by Equ. 1; it was calculated at the plane of the observer in cycles per degree (cpd):

$$MTF(u)_{system} = MTF(u)_{camera} \times MTF(u)_{prefilter} \times MTF(u)_{decimation} \times MTF(u)_{display}$$
(Equ. 1)

- The $MTF(u)_{camera} \times MTF(u)_{prefilter} \times MTF(u)_{(decimation)}$ was measured by the e-SFR method in Imatest^{TM} software from a captured ISO test chart [16]. The chart was subjected to the same capture-processing pipeline as the "base" image, i.e. the image decimated to 1.2 MP, matching 1:1 the pixel resolution of our reference (EIZO CG277) display (see Objective evaluations section).

- The $MTF(u)_{display}$ was modelled using the frequency response of a square pixel (i.e. a *sinc(au)* function, a = reference display pixel pitch) [17], multiplied with an exponential decay function that forced a slight reduction in the response at mid-frequencies. The resulting function matched measured MTFs of the exact same type/trademark of displays [14, 18, 19].

- The calculated $MTF(u)_{system}$, in cpd, matched closely the "aim" MTF of the DLL corresponding to SQS$_2$ 31, but it was still slightly higher than the latter (demonstrating the

very high quality of our capture-processing-display system).

- A linear difference filter (a 31x31 kernel) was designed to compensate for the difference between $MTF(u)_{system}$ and the "aim" MTF. It was convolved with the "base" image (having MTF= $MTF(u)_{system}$) to produce the first SRS with SQS$_2$ value 31.
- A set of linear filters were subsequently designed and implemented to produce all remaining SRSs with corresponding SQS$_2$ values from 0 to 30.
- We conducted pilot studies, using the method of adjustments [8] and three expert observers to extend the SQS$_2$ to include higher quality values (32, 33 and 34). Based on mean results from observers and three "average scenes in content", we produced relevant filters and subsequently relevant SRSs. The relevant SQS$_2$ values are not considered very accurate in terms of JND separation; they were, nevertheless, proven useful for rating test images originating from the higher MP sensors.
- Following the processes above, we produced 14 sets of SRSs, with scene contents matching the test scenes described earlier (Appendix A).
- All SRSs were finally cropped to match the square dimensions (1089 x 1089 pixels) and exact contents of the test images.

Experimental procedure

In the soft copy image quality ruler studies, two images are presented simultaneously to the observer, a ruler image and a test image (Figure B.1). In our study, the ruler image always depicted the same scene as the test image. Our observers were asked to use the slider bar to adjust the sharpness of the ruler image until they judged it matched the quality (or "value") of the test image and then press NEXT. They were instructed to judge the quality of the entire image and try to ignore feelings relating to the individual scene contents. The GUI ran though all test images in a random order. In addition to the 126 test images, a ruler image from each scene set was added to the pool of the test images to determine how accurately the observers completed their task [10].

Thirty-three observers participated in the image quality ruler experiments. Results placed directly each test image on the SQS$_2$. Based on analysis of results and criteria stated in [15], 6 observers' ratings were excluded from the derived mean ratings.

Star-rating experiments (categorical scaling)

Star-rating tests are categorical scaling experiments, most often having 5 categories, identified by the number of stars. They are commonly used in industry because they are quick and thus allow the collection of large amounts of visual data. Also, they are universally understood, i.e. categories are not labelled so translation to different languages is not necessary. They have an odd number of categories and a mid-point, which is assumed to guide the observer maintaining the categories intervallic (separated by equal perceptual intervals) - a desirable attribute for any quality scale [17]. This assumption cannot be taken for granted [8], especially because the categories are not associated with labels that could guide the observers. So unless results from star-rating experiments are treated using Thurstone's psychophysical law of Categorical Judgements [8, 9], mean observer ratings deliver ordinal rather than intervallic quality scales.

We carried out star-rating experiments to define the levels of quality of our test images and to map star-rating categories to

JNDs in quality. During the tests, observers were shown one test image at a time presented in a random order at the centre of the display (Figure B2). They were asked to assign a number of starts reflecting their judgement on the image's quality, from 1 star (lowest quality) to 5 stars (highest quality), by clicking the relevant radio button and then press NEXT. We did not use anchor stimuli because in common consumer ratings these are not applicable. Nonetheless, all our observers were already acquainted with the range of test images' quality from having participated in the ruler experiment before they took the star-rating test.

Twenty observers participated in the experiment and, after relevant treatment to exclude poor observations [8], mean category ratings for each MP sensor resolution were calculated from all images and observers. These produced an ordinal scale of quality.

We then processed the mean data by implementing the Law of Categorical Judgments, Condition D model [8,9] to produce an intervallic quality scale. We surpassed the problem of having an "incomplete matrix" by applying the recommended solution in [8, p.133-134].

The treatment provided category scale boundaries (i.e. the perceptual points where there is a change in each star category) mapped on the Objective Quality metric (see Objective evaluations section); and SQS$_2$ values for each MP sensor resolution, placed on a perceptually linear scale.

Acceptability thresholds

After category assessments were determined, we used acceptability threshold experiments to define to the level of sensor MP resolution, and associated category, beyond which the displayed image quality was judged as "*acceptable for high quality mobile camera phone imagery*" by camera phone consumers. Further, to map acceptability thresholds to corresponding SQS$_2$ values. Acceptability was tested for images displayed on high quality desktop monitors (such as our reference display). The results provided the minimum quality level above which the image us judged as acceptable to observers.

During the experiment, one test image was presented at a time in the centre of the display in random order. Thirty observers, one at a time, were asked whether they judged the image of "acceptable" quality, yes (1) or (0). A psychometric curve was fitted to the mean proportion of yes responses, averaged from scenes and observers; points 0.5 and 0.75 of the psychometric function were used defined the relevant lower and upper levels of acceptability, respectively [8].

Objective evaluations

Spatial Frequency Response and SFR50

SFR evaluation [3] of the camera system and/or of the sensor simulation resolutions was necessary at different stages of the project: i) measurement of $MTF(u)_{camera}$ for the development of the spatial frequency filters, used for liner filtering in our simulations; ii) calculation of $MTF(u)_{system}$ for the development of the SRSs (ruler images); and iii) measurement of $MTF(u)_{camera} \times MTF(u)_{prefilter} \times MTF(u)_{(decimation)}$ for calculating the selected IQMs.

The Imatest^{TM} Enhanced ISO-12233:2014 e-SFR test chart, printed on high quality photographic paper [16] was used for the purpose. The specific version of the chart is suitable for full frame capture using up to 50MP cameras without the need for spatial frequency compensation. Doubling the camera-to-chart distance

allowed for a straight SFR camera evaluation of the 100 MP camera.

During capture, the chart was mounted according to Imatest[TM] recommendations [20] and ISO12233 [3] specifications. Four incandescent lamps with correlated color temperature (CCT) of 3700K were illuminated the target evenly, providing an average illumination of 640 lux on the plane of the target. Distances were set such that there were "no more than 140 sensor pixels per inch of target" [20], thus camera-to-chart distances varied with varying focal lengths and lenses.

We captured the target with all relevant focal length-aperture-ISO speed combinations. The test chart images were processed in the same fashion as the relevant captured test scenes. SFR measurements were carried out in Imatest[TM] Master 5.1 software. On- and off- camera axis SFRs were weighted according to [7] to produce a single SFR curve for any given system.

The SRF50 objective metric of imaging system performance was derived from these curves.

Relevant image quality metrics

IQMs relating to perceived sharpness and resolution were also calculated in Imatest[TM] Master 5.1 software; the implementations comply closely with CPIQ and ISO 12233 recommendations. All IQMs were calculated for our reference and viewing conditions; they employ the CSF (indirectly) via provision of observer distance, display pixel pitch and viewing field. The following metrics were derived directly from the relevant captured/processed test charts:

- *CPIQ Acutance* with respect to SFR [5], with values ranging from 0 (worst quality) to 1 (best quality).
- *Subjective Quality Factor* (*SQF*) [4], with values ranging from 0 (worst quality) to 100 (best quality).
- *CPIQ Quality Loss* (*QL*) [5]: based on CPIQ Acutance, it measures the difference in quality of a sample image from the maximum subjective quality value, in SQS_2 units (in JNDs). Values range from 0 (best quality) to – theoretically – the maximum subjective quality (worst quality), (see Subjective evaluations section).
- *Objective Quality:* calculated by subtracting *CPIQ QL* from SQS_2 max, set to the maximum subjective quality. Since Objective Quality it is based on *QL*, one unit is automatically calibrated to correspond to 1 JND in SQS_2. Values range from 0 (worst quality) to the maximum subjective quality (best quality).

Results

Subjective quality ratings in SQS_2

Figure 3 is a plot of observer responses - in SQS_2 values - for all observers, versus 9 levels of MP sensor resolution (from 0.34 to 100 MP). It includes mean responses from all observers/stimuli. Note the 1.2 MP point in the x-scale corresponds to 1:1 camera-to-display resolution (i.e. no interpolation before display).
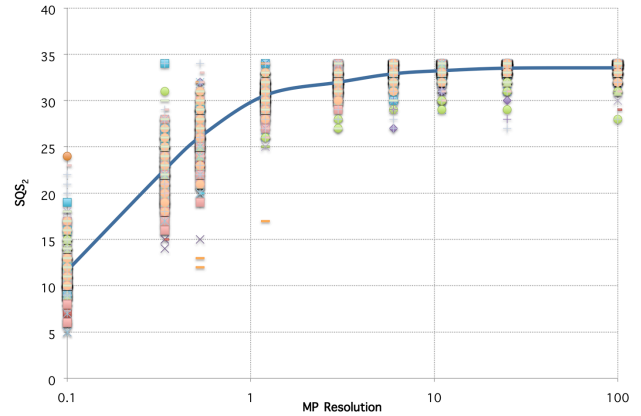

Figure 3. Mean observer SQS_2 versus MP sensor resolution.

Figure 4 is a plot of SQS_2 values per scene versus MP resolution, averaged over all observers, and including an average of all scenes (grand mean in SQS_2); a figure indicating test scene susceptibility [21].
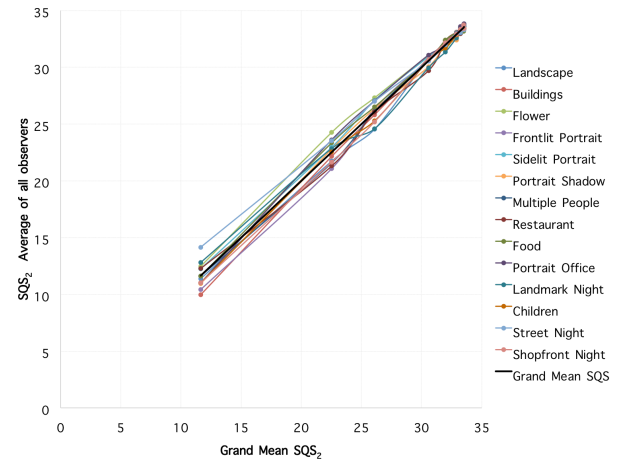

Figure 4. Test scene susceptibility to quality changes with changes in MP resolution.

Metric correlations

Figure 5, 6, 7, 8 and 9 illustrate, respectively, the MTF50, Acutance, SQF, QL and Objective Quality metric value correlations with SQS_2 values.
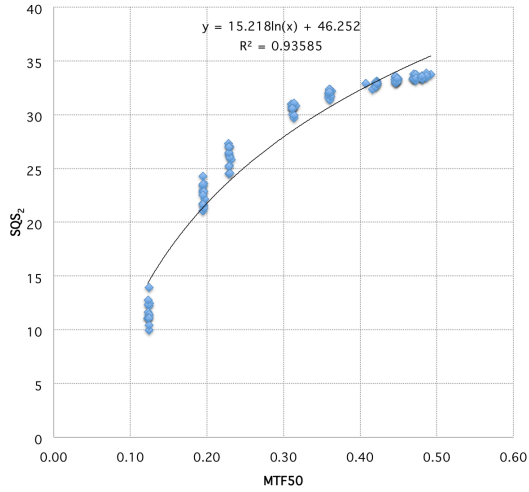
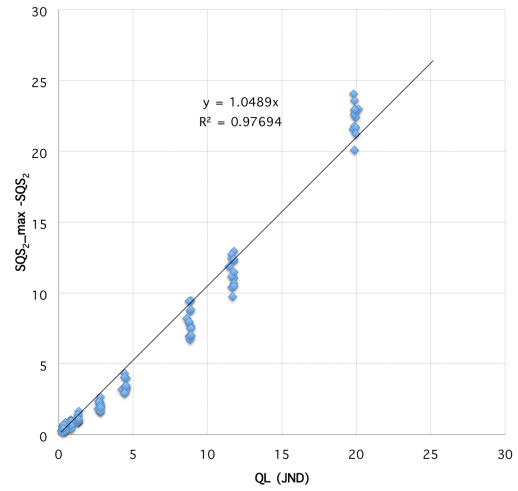Figure 5. MTF50 versus SQS_2 for all scenes.
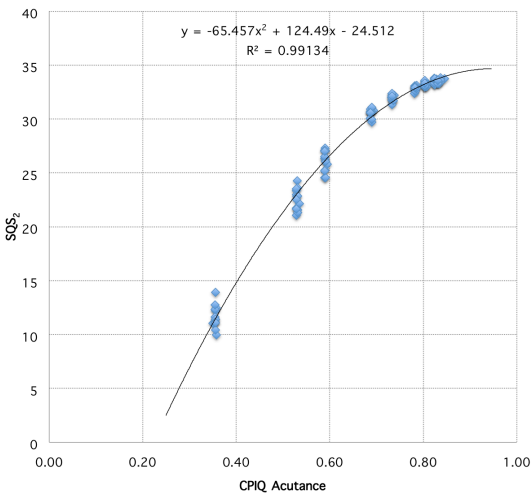

Figure 6. Acutance versus SQS_2 for all scenes.

Figure 7. SQF versus SQS2 for all scenes.

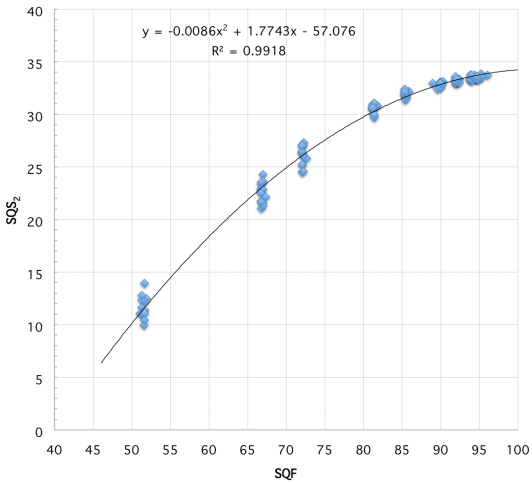Figure 8. QL versus SQS_2 for all scenes.

Figure 9. Objective Quality versus SQS_2 for all scenes.

Star ratings

Figure 10 plots the subjective interval scale obtained by implementing the Law of Categorical Judgement, Condition D [8,9] on the mean start rating data, versus Objective Quality metric values (closely matching SQS_2). Table 1 presents star rating categories in JNDs.

Table 1: Star-rating categories mapped to Objective Quality metric.

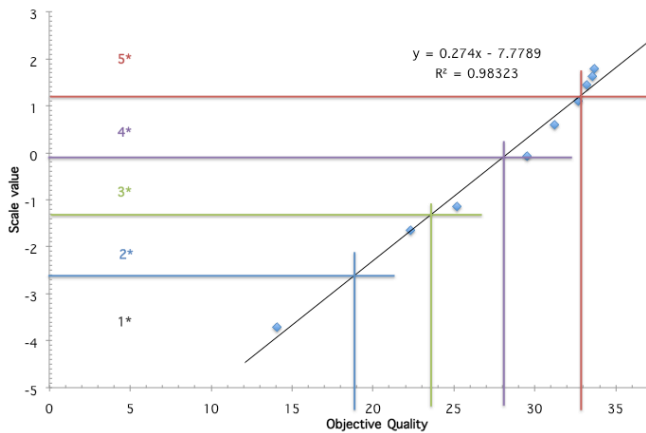| Stars	SQS_obj	JNDs between star categories
5	>33	
4	28 -33	5
3	23.5 -28	4.5
2	19 - 23.5	4.5
1	<19	

Figure 10: The star-rating scale and category boundaries values vs Objective Quality values.

Acceptable mobile phone camera quality

Figure 11 is the fitted psychometric curve as a function of Objective Quality in JNDs, with derived limits lower and upper limits of acceptable mobile phone quality. These correspond to metric values of 28 and 31 respectively, indicating a lower limit of acceptability of at least 4-star quality.
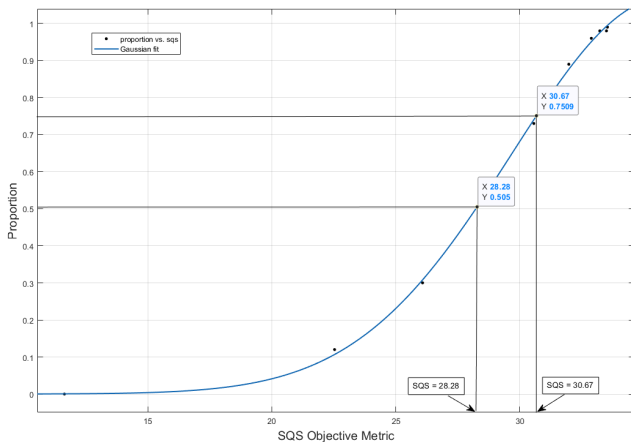

Figure 11. Psychometric function fitted to mean scene / observer responses from all 9 MP sensor images.

Conclusions

This paper investigated how the resolution of phone camera sensors affects the displayed image quality, when images are displayed at full size on quality LCDs. It is one of the few studies that consider the end-to-end capture-display-viewing conditions chain, and relate displayed quality ratings, in JNDs, to the resolution of the entire system. It further validated relevant IQMs using images from simulated sensors, including metrics proposed by CPIQ.

We conducted tests to identify quality scale values, categories and acceptability of mobile phone imagery, with test images originating from simulated sensors of different MP resolutions. We have collected a plethora of results and have carried out some analysis. Overall our analysis to date indicates:

- Subjective quality tests demonstrated that image quality differed by less than 2 JNDs between for images from

resolutions 100 MP and 6 MP, corresponding to 83:1 and 5:1 ratio of camera-to-display resolution. Between 5:1 and 2.5:1 ratios the quality decreased relatively slowly, but after 1:1 ratio it quickly deteriorated.

- Scene susceptibility (the variability in mean observer scores between scenes of the same MP resolution) increased as MP resolution thus and quality decreased - as expected [11, 21]. Well visible differences in quality (2<x<4.5 JNDs) are noticed at and below 1:1 ratio of camera-to-display resolution.

- MTF50, Acutance and SQF metrics were calculated directly from relevant charts and software, compliant with CPIQ recommendations. All three IQMs correlated well with image quality ratings, indicating that they are useful for predicting quality when changing sensor MP resolution. MTF50 does not take into account the viewing conditions. Acutance and SQF correlated excellently with subjective quality ($r^2 > 0.99$), but the relationship between the metric values and the subjective quality SQS_2 is not calibrated.

- Acutance correlated linearly with subjective quality, as expected, until high subjective qualities corresponding to the extended part of the SQS_2, where the relevant ruler images were over SQS_2 31 (and thus JNDs may have been inaccurate).

- CPIQ QL and the CPIQ Objective Quality metrics produced values directly proportional to SQS_2 values, separated by 1 JND in quality. These are pre-calibrated metrics and provided excellent correlation with subjective quality, $r^2 > 0.97$.

- Results from the star-rating experiments, when treated by implementing the relevant psychometric Law of Categorical Judgements, allowed an intervallic scale to be derived. Categories were mapped to the perceptually linear Objective Quality metric values, with the middle categories having 4.5<x<5 JNDs separation in quality. It is important to keep in mind that, category boundaries are only based on the given image set and MP resolutions. But the findings follow indicated research by [17, 21].

- The just-acceptable MP resolution limit was derived as 0.8MP, corresponding to a 0.67:1 ratio of sensor-to-display resolution and Objective Quality value 28. The acceptable resolution was found to be 1.34 MP and over, corresponding to a 1.2:1 ratio of sensor-to-display resolution ratio, and the relevant quality value equal to 30 (which is high). It should be noted that these results only apply to images displayed in full; "zooming-in" on an image would require it to be from a higher MP sensor to be acceptable.

- The four-star category indicated the limit of acceptable quality when observers - representing consumers of digital images - considered high quality mobile phone camera imagery. Both the last two conclusions indicate the ever-increasing expectation of camera-phone users.

References

[1] IEEE Standard Camera Phone Image Quality, https://standards.ieee.org/products-services/icap/programs/cpiq/index.html, November 19 2019

[2] E. Allen, S. Triantaphillidou, Eds 10th Edition, The Manual of Photography, Focal Press (Elsevier), 2011, Ch. 19, 24.

[3] ISO 12233:2017 Photography — Electronic still picture imaging — Resolution and spatial frequency responses, 2017.

[4] E. M. Granger and K. N. Cupery, "An optical merit function (SQF), which correlates with subjective image judgments", Photogr. Sci. Eng., vol. 16, no 3, pp. 221–230, 1972.

[5] Advisory Board of IEEE Standards Association Board of Governors, IEEE Standard 1858 - Standard for Camera Phone Image Quality, 2016.

[6] J. A. Movshon and L. Kiorpes, "Analysis of the development of spatial sensitivity in monkey and human infants", J. Opt. Soc. Am., vol. 5, no. 12, pp. 2166–2172, 1988.

[7] ISO 20462-3:2012 Photography — Psychophysical experimental methods for estimating image quality — *Part 3: Quality ruler method*, 2012.

[8] P. G. Engeldrum, Psychometric scaling: A toolkit for imaging systems development, Imcotek Press, 2000.

[9] F. C. Grum, R. Becherer and C. J. Bartleson, Optical radiation measurements, Volume 3, 1980.

[10] Jin, E. *et al.*, "Towards the Development of the IEEE P1858 CPIQ Standard – A validation study", Proc. IS&T Electronic Imaging: Image Quality and System Performance XIV, pp. 88-94, 2017.

[11] E. Allen, S. Triantaphillidou, Eds 10th Edition, The Manual of Photography, Focal Press (Elsevier), 2011, Ch. 23.

[12] C. A. Vargas Aguilera, LanczosFilter.m, https://uk.mathworks.com/matlabcentral/fileexchange/14041-lanczosfilter-m, accessed: November 19 2019.

[13] Gao, S. and Gruev, B., Bilinear and bicubic interpolation methods for division of focal plane polarimeters, Optics Express, vol. 19, no. 27, 2011.

[14] M. Shah, Creation and assessment of a Video Quality Ruler, Masters Thesis, Lund University, Sweden, 2016.

[15] E. W. Jin, and B. W. Keelan, "Slider-adjusted softcopy ruler for calibrated image quality assessment", J. of Electronic Imaging, vol. 19, no. 1, 2010.

[16] Imatest[TM], ISO 12233:2017 Edge SFR (eSFR) Photographic chart, color version, http://store.imatest.com/test-charts/iso-cpiq-standard-test-charts/photographic-iso-12233-2014-esfr-chart.html, accessed: November 2019.

[17] J. B. Phillips, and H. Eliasson, Camera Image Quality Benchmarking, Wiley-IS&T Series in Imaging Science and Technology, 2018.

[18] M. Persson, Subjective Image Quality Evaluation Using the Softcopy Quality Ruler Method, Masters Thesis, Lund University, Sweden, 2014.

[19] E. Allen, Image Quality Evaluation in Lossy Compressed Images, PhD Thesis, University of Westminster, UK, 2017.

[20] Imatest[TM], Test Lab Setup: Chart Mounting, http://www.imatest.com/solutions/Test-lab-setup/#ChartMounting, accessed: November 2019.

[21] B. W. Keelan, Handbook of Image Quality, Characterisation and Prediction, Marcel Dekker, NY, 2002.

Appendix A

Figure A.1 shows thumbnails of the test scenes, captured with the PhaseOne system.

outdoor landscape

outdoor buildings

burst child multiple people

side-lit portrait

side-lit portrait

front-lit portraits

portrait in office

portrait in shadow

restaurant

food close up

flower close-up

fruit shop

landmark at night

night city

Figure A.1: Test scenes

Appendix B

Figure B.1, B.2 and B.3 illustrate the interfaces for the soft copy ruler, the star-rating and the acceptability experiments, respectively.

Figure B.1: The soft copy ruler interface

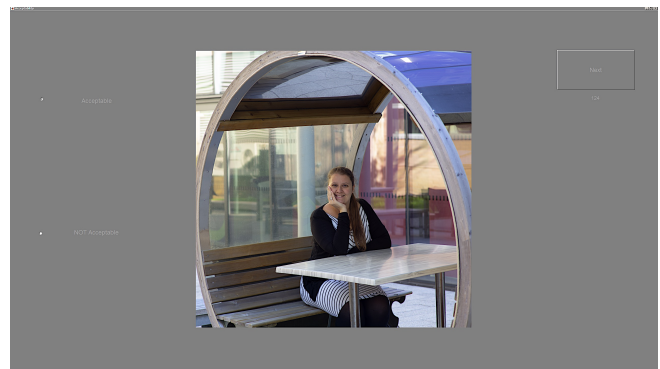
Figure B.2: The star-rating interface

Figure B.3: The acceptability interface

Author Biography

Sophie Triantaphillidou received her BS in Photography and Electronic Imaging Science (1992) and a PhD in Digital Imaging (2001), both from the University of Westminster, UK, where he is currently an Associate Professor and Director of Computational Vision and Imaging Technology group. Her research focuses on the quantification of imaging system performance and image quality. She is a Board Member of IS&T.