Journal of Electronic Imaging

JElectronicImaging.org

Image quality wheel

Toni Virtanen
Mikko Nuutinen
Jukka Häkkinen

Toni Virtanen, Mikko Nuutinen, Jukka Häkkinen, "Image quality wheel," *J. Electron. Imaging* **28**(1), 013015 (2019), doi: 10.1117/1.JEI.28.1.013015.

Journal of Electronic Imaging 28(1), 013015 (Jan/Feb 2019)

Image quality wheel

Toni Virtanen,* Mikko Nuutinen, and Jukka Häkkinen
University of Helsinki, Faculty of Medicine, Department of Psychology and Logopedics, Helsinki, Finland

Abstract. We have collected a large dataset of subjective image quality "*nesses," such as sharpness or colorfulness. The dataset comes from seven studies and contains 39,415 quotations from 146 observers who have evaluated 62 scenes either in print images or on display. We analyzed the subjective evaluations and formed a hierarchical image quality attribute lexicon for *nesses, which is visualized as image quality wheel (IQ-Wheel). Similar wheel diagrams for attributes have become industry standards in other sensory experience fields such as flavor and fragrance sciences. The IQ-Wheel contains the frequency information of 68 attributes relating to image quality. Only 20% of the attributes were positive, which agrees with previous findings showing a preference for negative attributes in image quality evaluation. Our results also show that excluding physical attributes of paper gloss, observers then use similar terminology when evaluating images with printed images or images viewed on a display. IQ-Wheel can be used to guide the selection of scenes and distortions when designing subjective experimental setups and creating image databases. *© 2019 SPIE and IS&T* [DOI: 10.1117/1.JEI.28.1.013015]

Keywords: image quality; attributes; lexicon; subjective evaluation; diagram; *nesses.

Paper 180544 received Jun. 17, 2018; accepted for publication Dec. 27, 2018; published online Jan. 30, 2019.

1 Introduction

As image quality research can be seen as the subsection of the highly multidisciplinary science of quality of experience (QoE) consisting of the primary disciplines of vision science, color, computational and behavioral sciences, discrepancies in terminology, and variable definitions between fields can become a problem. For example, QoE and image quality are defined differently in various sources. In the Qualinet white paper, QoE is defined as follows: "Quality of Experience (QoE) is the degree of delight or annoyance of the user of an application or service. It results from the fulfillment of his or her expectations with respect to the utility and/or enjoyment of the application or service in the light of the user's personality and current state.[1]" Whereas the CPIQ phase 1 white paper states image quality to be the perceptually weighted combination of all visually significant attributes of an image when considered in its marketplace or application,[2] on the other hand Janssen and Blommaert considered the quality of an image to be the degree to which the image is both useful and natural,[3] while Engeldrum described image quality to be the integrated perception of the overall degree of excellence of an image,[4] and Keelan characterized image quality as the impression of its merit or excellence, as perceived by an observer neither associated with the act of photography nor closely involved with the subject matter depicted.[5] The variation in definitions not only reflects various time periods and application areas but also shows that context and research area affect the definition. A common feature of image quality definitions is the idea to conceptualize image quality as a combination of "*nesses" such as sharpness and colorfulness. The *nesses are weighted and summed to create the overall model of image quality.[6–8]

To facilitate communication and understanding between professionals in various fields of image quality as well as nonprofessionals alike an image quality wheel (IQ-Wheel) for the *nesses is presented. Reference wheels and terminology lexicons have a long tradition in the sensory evaluation fields such as taste sensory experience studies where they are used to facilitate communication between interested stakeholders.[9–14] Wheels can also be useful in teaching observers which attributes are related to each other. Understanding that attributes can be related in a wider perspective can prevent disagreement on the definition of a particular attribute in panel discussions helping consensus flow more naturally. It has also been noted that consumers tend to prefer "negative" terms over "positive" while judging image quality.[6] This leads to the fact that image quality is often judged solely on its weaknesses and not its strengths which might be problematic for product development processes. The proposed IQ-Wheel has both negative and positive image attributes in a hierarchical and condensed form helping observers to remember and consider all aspects of image quality in their judgments.

It is also common practice when benchmarking products or developing new solutions in image quality to test and validate their impact and various aspects with end-users. Various standards for image quality evaluation methods have been presented[15–18] but not much focus has been put on how the various aspects of image quality are communicated to the observers. Especially with nonprofessional observers, it is important that they, too, understand the terminology in the same way as imaging professionals. If the observers do not have clear understanding about the distortions they are asked to evaluate, they might focus on areas in the image where the distortion is masked out and thus underweight its effect on image quality. Studies have also shown that the terminology nonprofessionals use when evaluating image quality is not the same terminology what engineers and experts use.[19,20]

*Address all correspondence to Toni Virtanen, E-mail: toni.virtanen@helsinki.fi

1017-9909/2019/$25.00 © 2019 SPIE and IS&T

1.1 *State-of-the-Art*

Pedersen's seminal work can be considered the first attempt to create a standardized lexicon for color print quality. They surveyed attributes from the literature and condensed the results into six-dimensions of print image quality; color, lightness, contrast, sharpness, artifacts, and physical represented with a folded Venn ellipse diagrams.[21] Our study consists of both printed images and images presented on a display giving us the possibility to compare how the medium might affect the terminology of the observers. Keelan[5] approached the image quality attributes through measurability viewpoint and created hierarchical classification of image quality attributes by dividing them into personal, aesthetic, preferential, and artifactual attributes but also considering whether the attributes have objective tractability, first and third party rater correlation, and system dependence. Based on subjective interview experiments for printed photographs, Leisti et al.[22] classified attributes to have two levels: low level and high level. The most important low-level attributes for printed images are brightness of color, sharpness, graininess, brightness, color quality, gloss, contrast, and lightness. The high-level attributes, however, are used to funnel the importance of the low-level attributes and consist of realism, naturalness, clarity, depth, and aesthetic associations.

Contrary to the expert panel or literature review approach of designing terminology lexicons and wheels often used in the sensory evaluation fields,[9–14] we opted for an empirical approach based on the observers' free descriptions. Founding the image quality lexicon on empirical data gives us better understanding about the prevalence distribution between individual attributes. Prevalence can be considered as the visibility and impact of these attributes and how they influence the overall image quality experience.[8] Previous studies on image quality with free description exist but are based on fever set observers and stimuli. Such as 14 observers rating 8 scenes,[23] 29 observers rating 8 scenes,[24,25] 48 observers rating 4 scenes,[22] and 61 observers rating 17 scenes.[26] Excluding movement and sound, video quality has also significant overlap with image quality evaluations when using free descriptions. A study with 138 observers 22 devices and two different video clips[27] found that 12 attributes were specifically related to image quality of the video. The studies are adequate for their specific cases, but lack the needed coverage of both observers and scenes to represent the overall terminology of image quality in general.

In this study, we utilized the approach called interpretation-based quality (IBQ) for gathering the observers' free descriptions from visual stimuli.[24] In the IBQ approach, the subjects estimate the overall quality of each image and then described the most distinctive features of its image quality using free description. The IBQ was inspired by sensory profiling methods and other sensory modalities such as taste and touch,[28,29] and it was first conceived as a solution to gain more thorough knowledge of user-experience quality in high-quality magazine printing.[30] The methodology has been successfully tested with image quality evaluation,[26,31] print quality evaluation,[32] video quality evaluation,[27,33] stereoscopic quality evaluation,[34] and quality evaluation of 360 videos.[35]

The primary contributions of this paper are summarized below:

- We present an image quality lexicon founded on the free descriptions gathered from 146 observers rating 62 different scenes.
- We present IQ-Wheel, which is a visualization of the different aspects of image quality and their frequency distribution.
- We have gathered free descriptions from both printed images and images presented on a display to compare if there is an effect on the media the images are presented.

2 Methods

2.1 *Materials and Experimental Setup*

This paper consists of seven studies, see summary in Table 1. Studies 1 to 3 were conducted on printed photographs and studies 4 to 7 presented the photographs on display [see Figs. 1(a) and 1(b)]. The images were shot using three different imaging devices and raw signal was then manipulated using 60 different image signal processing (ISP) pipes. An ISP processes the raw data from the imaging sensor and controls, e.g., exposure and white balance algorithms. For example, if we have low-, mid- and high-end camera modules that all are processed with three different competing ISP's all striving to achieve a combination of manipulations that would produce the most pleasing outcome for the viewer in various scenes, we end up with nine different versions of the same scene. As previous studies have shown that image distortions can influence the way people view images.[36] All images had multiple overlapping manipulations that might even be counteracting each other, e.g., denoising versus sharpening, creating rich stimuli for collecting the free descriptions and create the IQ-Wheel. Unfortunately, the ISPs are the property of our industry partners or their subcontractors and our nondisclosure agreement prevents us from providing detailed information on what exact manipulation combinations they were using in their ISP's. Nevertheless, the ISP's included camera modules from high-end to low-end and as such the images represented the general variation that end-users would see in their daily life.

Observers were recruited through Helsinki University student's mailing lists and were not professionally involved in photography or imaging science. All the observers had normal or corrected to normal vision. Observers' near visual acuity,[37] near contrast vision (near F.A.C.T. Ginsburg, Stereo Optical Co.),[38] and color vision with Farnsworth D-15[39] were tested prior the experiment. After vision screening, the final number of observers in all the studies was 146 observers. The participants received a movie ticket as a reward for their time.

In the studies 1 to 3 (print), the room was covered with medium gray curtains and the room was illuminated with daylight simulating light tubes (Osram Biolux 36W/965). Illuminance levels were measured by Hagner digital luxmeter EC1. The images were developed in a professional developer. The image files were in sRGB photospace and the ICC profile of the printing company's printer was added to the image file to ensure the correct color management so that the test prints looked exactly like the pipelines would determine them to look like in the prints. Size A4 (210 × 297 mm) high-quality glossy printing paper was

Table 1 Breakdown of the studies.

Study	1	2	3	4	5	6	7	SUM
Observers	29	28	30	15	15	15	14	146
ISPs	6	8	6	13	9	9	9	60
Medium	Print	Print	Print	Display	Display	Display	Digital	
Contents	Art	Airplane	Alley	Beach	Baby	Boy	Bar	**62 Unique scenes**
	Three girls S1	Bike	Boarding	Breakfast S4	Breakfast S5	Grandma	Barefoot	
	Elk	Three Girls S2	Boats	Dinner S4	City	Hotel	Cars	
	Flower	Grill S2	Breakfast S3	Harbor	Dinner S5	Kiosk	Cathedral	
	Gentlemen	Kid	Cheers	Lounge S4	Town	Racing	Children	
	Girl	Monkey	Cottage	Winter	Lounge S5	Squirrel	Evening	
	Grill S1	Rapids	Ducks		Reindeer	Terrace	Hanami	
	Horse	Square S2	Fruits		Sea	Tulips	Restaurant	
	Lake	Street S2	Metro					
	Market	Sunset	Panda					
	Night	Swim S2	Pianist					
	Square S1	Windmill S2	Sausage					
	Street S1		Seaside					
	Swim S1		Snowboard					
	Windmill S1		Swimming pool					

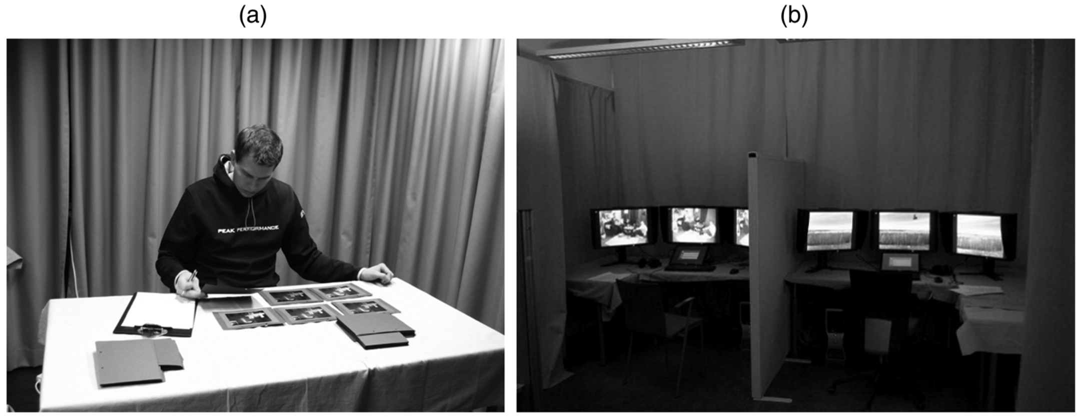

Fig. 1 Viewing conditions. (a) The print setup and (b) the display setup.

selected to make the smaller details more visible. Prints were presented on table covered with a medium gray tablecloth. Illuminance level in this area varied from 500 to 560 lx.

The print evaluation task followed a basic rank order methodology,[15,18] where the task was to rank the images in order of image quality and score them on a scale of 0 to 10 to get an interval scale for evaluations. After ranking the images, observers were instructed to "Write down free descriptions for each image of the reasons behind your judgment. You don't need to use whole sentences." We tried to keep the instructions as free as possible to prevent

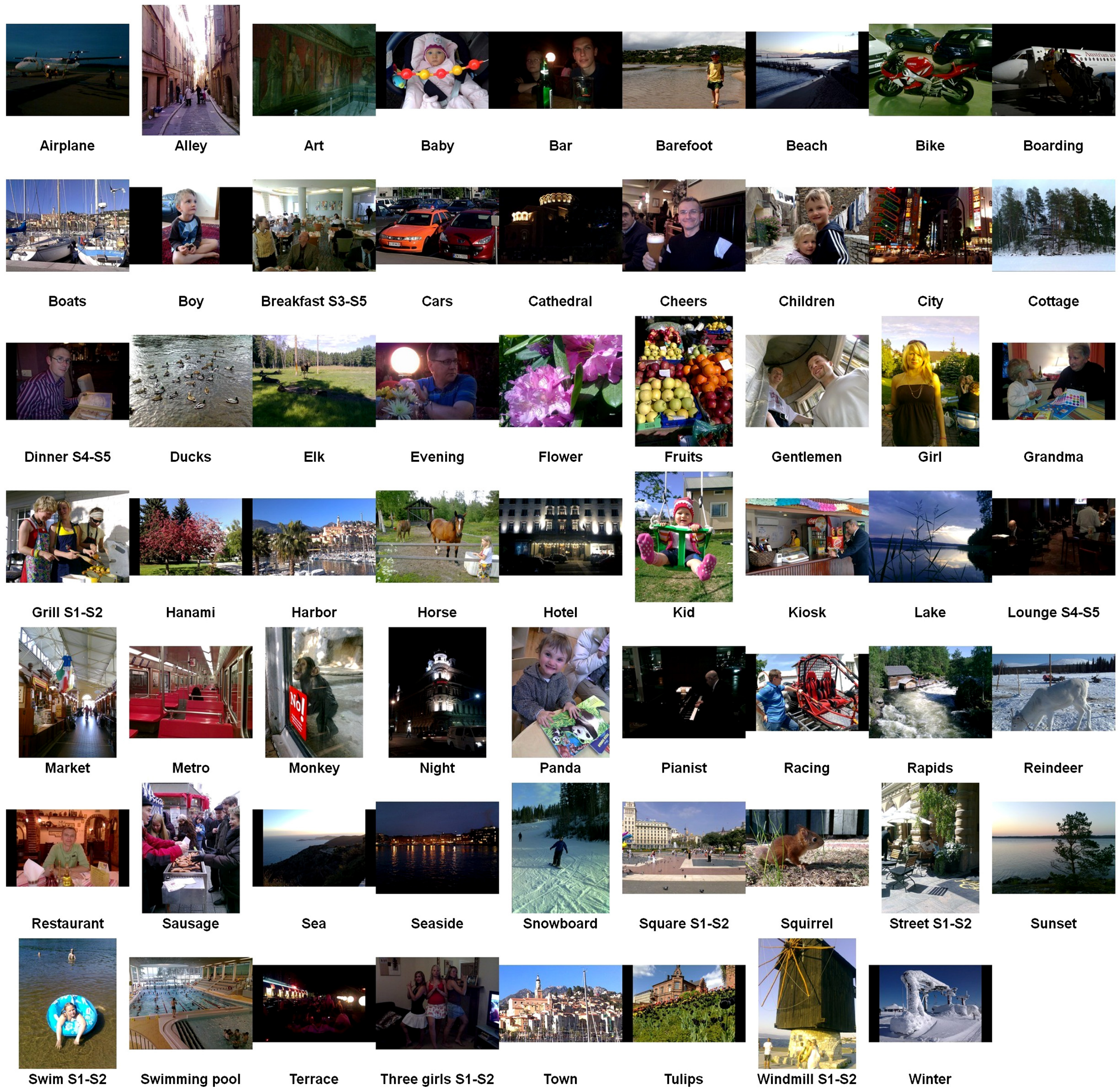

Fig. 2 The various scenes used in all the seven studies.

any leading questions, as it has been shown that it can have impact on the way people look at an image.[36] Images were randomized one scene at a time to reduce viewing order induced effects. Observers conducted one practice round to make sure they understood the instructions. The experiment lasted on average 1 h 34 min. The observers' evaluations were collected using pen and paper questionnaires.

In studies 4 to 7 (display), the room had been covered with medium gray curtains to diffuse the ambient illumination. Fluorescent lights (5800 K) were positioned behind the monitors and reflected from the back wall covered with gray curtain to create dim and uniform ambient illumination in the room. The light hitting the monitors measured below 20 lx. The observers' viewing distance (~80 cm) was controlled by a line hanging from the ceiling and they were instructed to keep their forehead steady next to the line. Because of the display size, images were scaled to a size of 1600×1200 pixels using the bicubic interpolation method. Three Eizo ColorEdge CG241W monitors with 1920×1200 pixel resolution were calibrated to sRGB using EyeOne Pro calibrator (X-rite co.). The target values were 80 cd/m^2, 6500 K, and gamma 2.2.[40]

The studies followed a modified softcopy version of the ISO 20462-2 Triplet comparison method,[41] where observers saw three images depicting the same scene on separate displays and had a fourth display where the rating scales were shown. Instead of just ranking the images from 1 to 3, each image got a rating from 0 to 10 to get an interval scale for evaluations. Giving the same score to two images in a triplet was prevented. After ranking the images, observers were instructed to write their free descriptions with exactly the same verbal instruction as in the print studies 1 to 3. The images were randomized so that each image was paired against each other at least once in all of the triplets. Observers conducted practice rounds until they indicated they were ready for the actual test. The experiment lasted on average 1 h 33 min. All experiments in the display photograph studies were conducted using the VQone MATLAB toolbox.[42]

Altogether 62 different scenes were collected for the studies. The scenes were intended to represent typical photographs that consumers might capture with their camera devices. Six images had animals, 10 images depicted architecture, 14 images had bright sunlight, 18 images were night or dark images, 4 flowers, 10 group pictures, 21 indoor images, 14 landscapes, 41 outdoor images, 26 images had people, 15 portraits, 3 images depicted snow, and 2 images were close-ups (Fig. 2).

2.2 *Analysis*

The observers' free descriptions, e.g., "very bright, but blurry image" were aggregated in a two-step process. First the grammatical nuances and different inflections, e.g., the terms bright, brighter, and brightest were all summed up manually under the term bright (Fig. 3). Second the remaining terms were cross-referenced for synonyms, e.g., bright, luminous, and radiant to form the final attribute bright (Fig. 3). Synonyms were identified using FinnWordNet version 2.0 lexical database for Finnish, a derivative of the Princeton WordNet. FinnWordNet contains words (nouns, verbs, adjectives, and adverbs) grouped by meaning into synonym groups representing concepts. These synonym groups are linked to each other with relations such as hyponymy and antonymy, creating a semantic network. As FinnWordNet has been created by having the words of the original English (Princeton) WordNet (version 3.0) translated into Finnish by professional translators,[44] we could use it to translate the final attributes into English. To the researchers' knowledge, lexical databases for words such as the FinnWordNet have not been used before in image quality studies, and previous studies have combined the synonyms manually.[26,27,31,33–35]

3 Results

From the free descriptions of 146 observers, we gathered 39,415 individual quotations. These quotations were then summarized in the first step into 2742 wider concepts by combining grammatical nuances and different inflections as described in Sec. 2.2. Finally in the second step, the remaining 2742 concepts were cross-referenced for synonyms using FinnWordNet making the final count of individual attributes of 68 that will create the empirical basis of the IQ-Wheel (Table 2).

The IQ-Wheel (Fig. 4) was inspired by the flavor reference wheels and terminology lexicons from the sensory experience fields.[9–14] Contrary to the flavor reference wheels

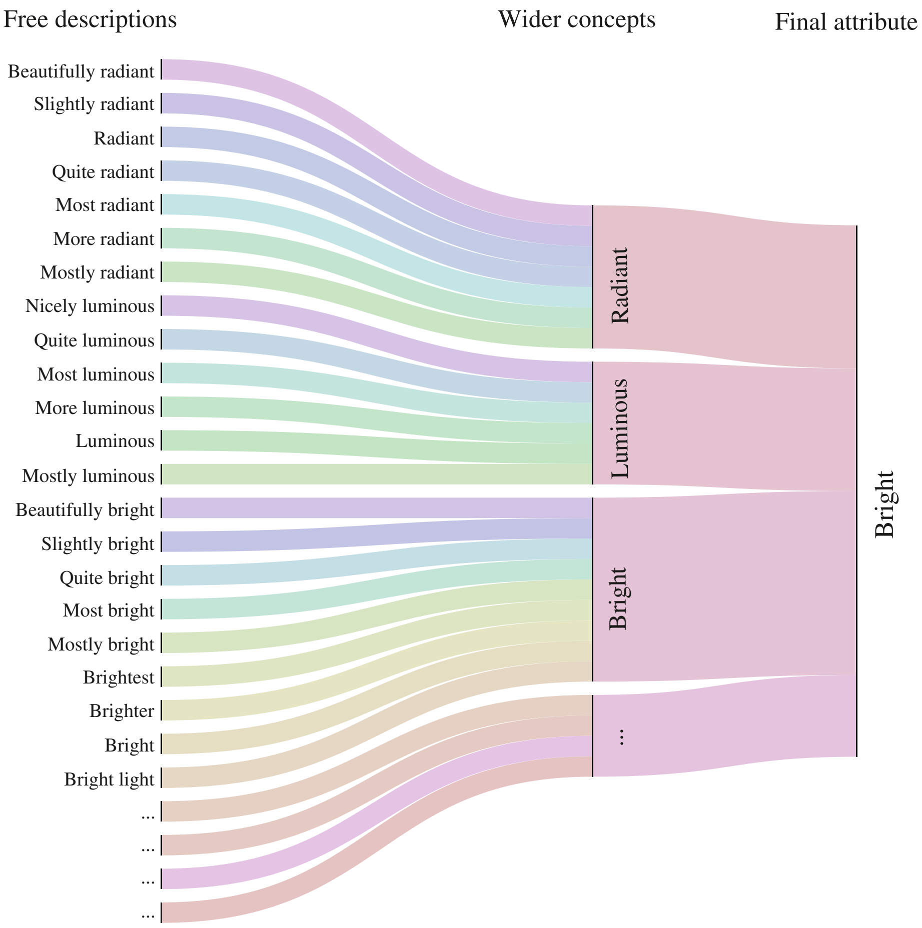

Fig. 3 Example of the condensing process for free descriptors. Figure created with Ref. 43.

Table 2 Attributes in the IQ-Wheel. Tier 1 is the inner circle, Tier 2 is the middle circle while attribute is the outer circle. Pos/Neg gives the valence to the attribute to image quality, wider concepts and free descriptions show how the attribute was condensed from the data (see Fig. 3).

Tier1	Tier2	Attribute	Pos/Neg	Wider concepts	Free descriptions
Artifacts	Geometric	Lens distortion	Neg	18	26
Artifacts	Geometric	Vignetting	Neg	16	22
Artifacts	Graininess	Blockiness	Neg	10	27
Artifacts	Graininess	Grainy	Neg	42	2578
Artifacts	Graininess	Noisy	Neg	11	56
Artifacts	Graininess	Pixelated	Neg	25	105
Colors	Brightness	Bright colors	Pos	18	299
Colors	Brightness	Dark colors	Neg	46	1284
Colors	Brightness	Too dark colors	Neg	53	956
Colors	Color cast	Blue	Neg	92	860
Colors	Color cast	Colorshift	Neg	20	47
Colors	Color cast	Green	Neg	64	481
Colors	Color cast	Orange	Neg	14	54
Colors	Color cast	Pink	Neg	8	24
Colors	Color cast	Purple	Neg	27	99
Colors	Color cast	Red	Neg	104	1648
Colors	Color cast	Sepian	Neg	37	161
Colors	Color cast	Turquoise	Neg	10	24
Colors	Color cast	Yellow	Neg	89	1494
Colors	Color temperature	Cold	Neg	37	227
Colors	Color temperature	Warm	Neg	55	570
Colors	Fidelity	Blurred colors	Neg	13	25
Colors	Fidelity	Clear colors	Pos	11	44
Colors	Fidelity	Uneven colors	Neg	22	147
Colors	Lack of color	Colorless	Neg	28	501
Colors	Lack of color	Faded colors	Neg	138	1912
Colors	Lack of color	Grey	Neg	51	756
Colors	Lack of color	Pale colors	Neg	37	152
Colors	Lack of color	Pastel colors	Neg	74	699
Colors	Much color	Loud colors	Neg	29	136
Colors	Much color	Saturated colors	Pos	43	401
Colors	Much color	Too colorful	Neg	18	76
Colors	Much color	Too saturated colors	Neg	36	204
Colors	Much color	Vivid colors	Pos	23	225
Colors	Valence	Bad colors	Neg	81	430

Table 2 (*Continued*).

Tier1	Tier2	Attribute	Pos/Neg	Wider concepts	Free descriptions
Colors	Valence	Good colors	Pos	147	2447
Contrast	Lack of contrast	Foggy overcast	Neg	28	255
Contrast	Lack of contrast	Not enough contrast	Neg	23	60
Contrast	Too much contrast	Too hard contrast	Neg	64	295
Contrast	Valence	Bad contrast	Neg	4	26
Contrast	Valence	Good contrast	Pos	32	128
Exposure	Overexposed	Bright	Pos	16	1110
Exposure	Overexposed	Overexposed	Neg	50	400
Exposure	Overexposed	Too bright	Neg	69	646
Exposure	Underexposed	Dark	Neg	12	499
Exposure	Underexposed	Dim	Neg	35	312
Exposure	Underexposed	Gloomy	Neg	20	118
Exposure	Underexposed	Shady	Neg	19	105
Exposure	Underexposed	Underexposed	Neg	24	132
Exposure	Valence	Bad exposure	Neg	29	111
Exposure	Valence	Good exposure	Pos	84	456
Naturalness	Natural	Natural	Pos	28	1882
Naturalness	Natural	Natural colors	Pos	32	720
Naturalness	Natural	Real	Pos	37	651
Naturalness	Natural	Real colors	Pos	24	299
Naturalness	Other	Aged photo	Neg	15	48
Naturalness	Other	Photoshopped	Neg	48	170
Naturalness	Unnatural	Unnatural	Neg	30	270
Naturalness	Unnatural	Unnatural colors	Neg	102	506
Naturalness	Unnatural	Unreal	Neg	36	206
Sharpness	Sharp	Clear	Pos	41	689
Sharpness	Sharp	Sharp	Pos	70	4283
Sharpness	Sharp	Too sharp	Neg	29	229
Sharpness	Unsharp	Blurry	Neg	45	1279
Sharpness	Unsharp	Motion blur	Neg	6	43
Sharpness	Unsharp	Soft	Neg	19	205
Sharpness	Unsharp	Unclear	Neg	51	735
Sharpness	Unsharp	Unsharp	Neg	73	3350
Total:				2742	39,415

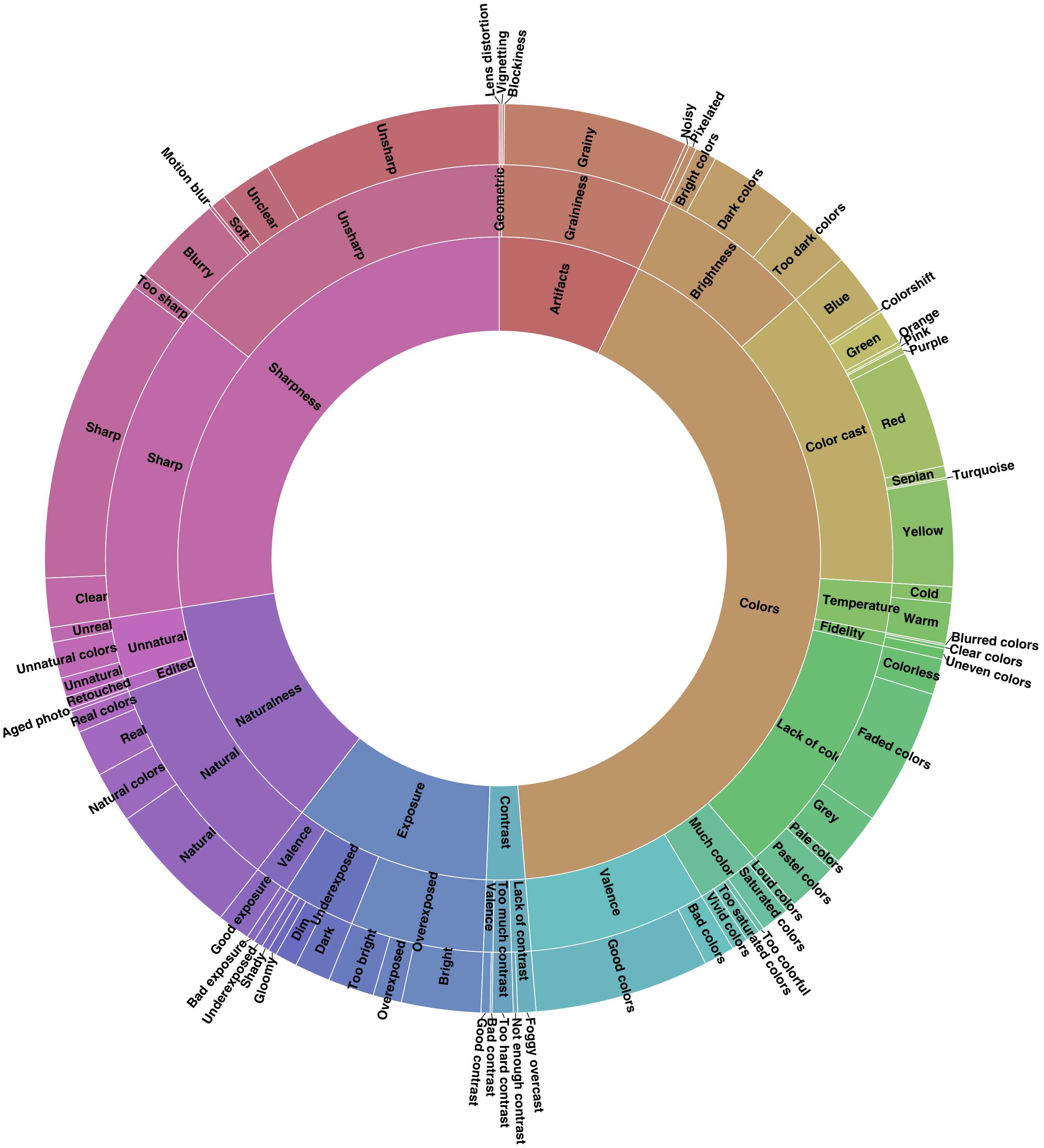

Fig. 4 IQ-Wheel. The area of the elements represents free description frequency of the attributes. The larger the area the more often it has been used by the observers. Figure created with Ref. 43.

that are mainly created with expert board meetings during conferences, the IQ-Wheel has an empirical background. With the empirical data from studies 1 to 7, it is possible to add frequency information to IQ-Wheel, where higher attribute frequency is represented by a larger area. Colors are is used to enhance readability, where each of the free descriptions is given their own hue that is translated inward to the central core categories of Tier 1.

3.1 *Attributes with Print and Display*

One of the purposes of this paper was to understand whether observers use different terminology when evaluating print images versus images presented on a display. Our results show that the frequency of use in the 68 attributes is highly correlated between printed images and images present on display. Pearson correlation coefficient was $r = 0.89$ and Spearman categorical correlation was $r = 0.90$ for the

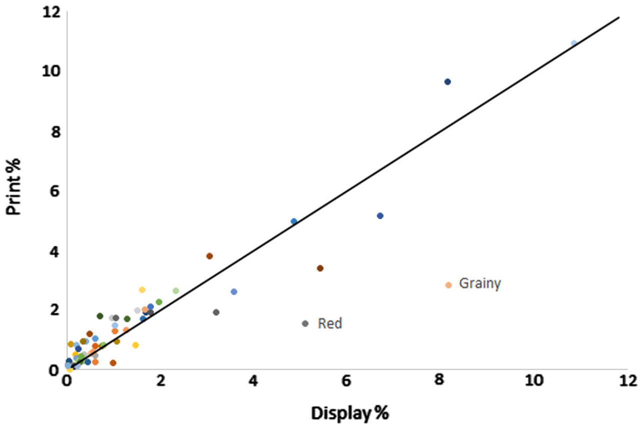

Fig. 5 Scatterplot of the relative prevalence of individual attributes. Vertical axis is the percentage distribution of individual attributes from the print studies 3 to 7 and horizontal axis is the percentage distribution of the same attributes from display studies 4 to 7.

whole range of attributes. To make sure that the good correlation is not just a bias because of the extreme values, we did another analysis with only attributes with under 3% frequency, excluding 10 attributes with the largest frequencies from the analysis. The Pearson correlation of $r = 0.86$ was still quite high with the attributes that had been mentioned less often. The main differences between individual attribute frequencies were with graininess and red color shift (Fig. 5).

We also compared the Pearson correlation within the attribute groups (the inner circle in the IQ-wheel) between display and print attribute frequencies. Where the artifacts group had $r = 0.99$; colors group $r = 0.87$, contrast group $r = 0.98$, exposure group $r = 0.69$, naturalness group $r = 0.89$, and sharpness group $r = 0.99$. Upon further inspection, the relative difference between display and print was calculated by (display % − print %) / print % for each attribute. Table 3 shows the 10 attributes that differed most between print and display stimuli.

Table 3 Ten largest relative differences between print and display attribute frequencies.

Attribute	Print frequency	Print %	Display frequency	Display %	Realtive difference
Gloomy	97	0.85	20	0.07	10.71
Motion blur	34	0.30	9	0.03	8.12
Turquoise	16	0.14	7	0.03	4.52
Blockiness	18	0.16	8	0.03	4.43
Vingeting	14	0.12	7	0.03	3.83
Pale colors	94	0.82	55	0.20	3.13
Clear colors	25	0.22	15	0.05	3.03
Bad exposure	60	0.53	49	0.18	1.96
Lens distortion	12	0.11	10	0.04	1.90
Unreal	109	0.96	92	0.33	1.86

4 Discussion

Our study presents image quality lexicon based on the attributes derived from the free descriptions of 146 participants in seven image quality studies. We visualized the most common quality attributes with the IQ-Wheel that can be used to facilitate communication and understanding between the professionals of the multidisciplinary fields of image quality. The lexicon could be used as the basis of creating an industry standard about image quality-related "*nesses."

This study supports previous findings suggesting that when observers are instructed to evaluate image quality, they tend to use more negative terminology than positive one.[6] In our study, only 20% of the attributes given by the observers can be considered positive. However, positive attributes often had larger frequency on average and all the positive attributes make out 35% of the total frequency of all free descriptions given. This can be interpreted that people use fewer words when commenting on the positive aspects of an image compared with negative aspects. In other words, there can be more ways observers perceive images can fail in quality than there are means to excel in quality. Another find from the data is that colors have a significant prevalence in observers' terminology. From all the free descriptions given, 45% were related to color. Our results also indicate that excluding the physical attributes such as paper gloss and others, observers mostly use the attributes in a similar way when evaluating printed images or images viewed on displays.

The IQ-Wheel is an efficient way of presenting the hierarchy, variation, and prevalence information of image quality in a single figure can also be used as a communication aid and education tool for observers helping them understand how different attributes might be related on a macro level. We also considered other ways to visualize the data, but ended up in the sunburst pie diagram for its ability to represent hierarchy in a condensed way. It is also the preferred diagram used in the sensory experience fields.[9–12,14]

4.1 *Practical Implications*

Having consensus on the common terminology can benefit the development and research of the whole field. In addition, industries can have multiple locations where they need to evaluate the same products and various business entities such as, marketing, manufacturing or research and development making explicit communication within the company crucial. Outsourcing the development and evaluation to varied software and component suppliers also necessitate clear communication across diverse audiences. However, as our data are based on the frequencies of attributes gathered from naïve observers, it might differ from what professionals in the field would have constructed. For example, the low amount of attributes related to contrast is surprising, but consistent with the previous results that professionals and naïve observers tend to use different terminology.[19,20] In the case of contrast, it might have translated into the more familiar terms relating to sharpness and brightness in the mouths of naïve observers.

The IQ-Wheel can be used as a tool for communication, and an aid in subjective experiments to make the observers better aware of the whole phenomenon and what is required of them. It can be added to the instructions section for the observers, used as a tool to select representative scales for subjective studies, or even implemented into the experimental setup by having the observers mark those attributes that represent the image under evaluation.

As the IQ-Wheel has a hierarchical structure, it can be partitioned based on specific needs. For some instances, just the six core attributes: artifacts, colors, contrast, exposure, naturalness, and sharpness in the center might be adequate. These six core attributes are almost exactly the same as in Pedersen's model,[21] only naturalness has replaced the class relating to physical paper properties. What is interesting is that Pedersen ended up with his attributes using a literature review, while our attributes are based on empirical data. For some other instances, all the attributes behind the colors fan might be necessary to gain the necessary nuanced view of the phenomenon under interest. The hierarchical model also allows us to better communicate where a certain attribute is situated in the hierarchy with the added information on the prevalence of that attribute using the frequency information provided.

The IQ-Wheel can also be used to guide the selection of scenes and distortions when creating image databases and support the development of a robust image quality assessment (IQA) algorithms. For example, IQA with a correlation of 0.85 against an image database does not tell us much about its general performance unless we would also know what aspects of image quality the database covers. IQ-Wheel could be one way to benchmark different image databases in evaluating what dimensions of image quality they cover. We acknowledge that the nondisclosure agreement preventing us from sharing information about the detailed manipulation combinations of the ISPs used in the experiments can limit the utilization of the IQ-Wheel in some cases. Nevertheless, the ISP's included camera modules from high-end to low-end with general variation that end-users would see in their daily life and therefore creates a representative set of images which the IQ-Wheel is based on.

Future work for the development of the IQ-Wheel would be to expand it to include motion-related attributes from videos and as imaging technologies develop also three-dimensional and virtual reality or augmented reality-induced artifacts. To further expand the utility of the IQ-Wheel. Yet another level of hierarchy could be added to it by taking the measurability viewpoint that Keelan[5] had presented and considered each attribute by whether it is personal, preferential, or artifactual and how it can be measured or tracked by technical means. Another addition would be to implement the high- and low-level categorization of the attributes suggested by Leisti et al.[22] in to the IQ-Wheel. Further studies are needed to better understand the interaction between attributes and numerical quality evaluations such as mean opinion scores (MOS). Some effort in that direction has already been made by Nyman et al.[23] who explored the relationship between observers' free negative and positive descriptions to the MOS for images. They found a paradigm shift in the observers' evaluation criterion where images with low MOS were rated using a different set of attribute space than the images with high MOS values.

Acknowledgments

This study was partly funded by User-Centered Information Technology (UCIT) doctoral school. The authors thank Nokia Corp. for providing the test material for the study. The authors have no competing interests to declare.

References

1. P. Le Callet, S. Möller, and A. Perkins, Eds., "Qualinet white paper on definitions of quality of experience," 2012, European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Lausanne, Switzerland, Version 1.1, (June 3 2012).
2. I3A, "CPIQ initiative phase 1 white paper: fundamentals and review of considered test methods" (2007).
3. T. J. W. M. Janssen and F. J. J. Blommaert, "Image quality semantics," *J. Imaging Sci. Technol.* **41**(5), 555–560 (1997).
4. P. G. Engeldrum, "A theory of image quality: the image quality circle," *J. Imaging Sci. Technol.* **48**(5), 446–456 (2004).
5. B. Keelan, *Handbook of Image Quality: Characterization and Prediction*, CRC Press, Boca Raton, Florida (2002).
6. S. Yendrikhovskij et al., "Enhancing colour image quality in television displays," *Imaging Sci. J.* **47**(4), 197–211 (1999).
7. P. G. Engeldrum, "Image quality modeling: where are we?" in *PICS*, pp. 251–255 (1999).
8. P. G. Engeldrum, "A short image quality model taxonomy," *J. Imaging Sci. Technol.* **48**(2), 160–165 (2004).
9. L. J. R. Lawless and G. V. Civille, "Developing lexicons: a review," *J. Sens. Stud.* **28**(4), 270–281 (2013).
10. M. C. Meilgaard, C. E. Dalgliesh, and J. F. Clapperton, "Beer flavour terminology," *J. Inst. Brew.* **85**(1), 38–42 (1979).
11. G. Richard, A. Oberholster, and F. I. Leigh, "A 'Mouth-feel Wheel': terminology for communicating the mouth-feel characteristics of red wine," *Aust. J. Grape Wine Res.* **6**(3), 203–207 (2000).
12. B. Chen et al., "Wineinformatics: applying data mining on wine sensory reviews processed by the computational wine wheel," in *IEEE Int. Conf. Data Min. Workshop ICDMW*, pp. 142–149 (2015).
13. M. Zarzo and D. T. Stanton, "Understanding the underlying dimensions in perfumers' odor perception space as a basis for developing meaningful odor maps," *Atten. Percept. Psychophys.* **71**(2), 225–247 (2009).
14. L. J. R. Lawless, A. Hottenstein, and J. Ellingsworth, "The McCormick spice wheel: a systematic and visual approach to sensory lexicon development," *J. Sens. Stud.* **27**(1), 37–47 (2012).
15. International Telecommunications Union, "Report R-REP-BT.1082-1: studies towards the unification of picture assessment methodology," (1990).
16. International Telecommunications Union, "ITU-T P.910: subjective video quality assessment methods for multimedia applications," (2008).
17. I.-R. R. BT, "500-13. Methodology for the subjective assessment of the quality of television pictures," in *Int. Telecommun. Union*, Geneva, Switzerland, pp. 53–56 (2002).
18. ISO, "20462-1:2005 photography—psychophysical experimental methods for estimating image quality—part 1: overview of psychophysical elements," (2005).
19. S. Bech et al., "Rapid perceptual image description (RaPID) method," *Proc. SPIE* **2657**, 317–328 (1996).
20. L. C. Cui, "Do experts and naive observers judge printing quality differently?" *Proc. SPIE* **5294**, 132–145 (2004).
21. M. Pedersen et al., "Attributes of image quality for color prints," *J. Electron. Imaging* **19**(1), 011016 (2010).
22. T. Leisti et al., "Subjective experience of image quality: attributes, definitions and decision making of subjective image quality," *Proc. SPIE* **7242**, 72420D (2009).
23. G. Nyman et al., "Evaluation of the visual performance of image processing pipes: Information value of subjective image attributes," *Proc. SPIE* **7529**, 752905 (2010).
24. J. Radun et al., "Explaining multivariate image quality: interpretation-based quality approach," in *Proc. Int. Congress Imaging Sci. (ICIS)*, pp. 119–121 (2006).
25. G. Nyman et al., "Measuring multivariate subjective image quality for still and video cameras and image processing system components," *Proc. SPIE* **6808**, 68080N (2008).
26. J. Radun et al., "Evaluating the multivariate visual quality performance of image-processing components," *ACM Trans. Appl. Percept.* **7**(3), 1–16 (2008).
27. T. Virtanen et al., "Forming valid scales for subjective video quality measurement based on a hybrid qualitative/quantitative methodology," *Proc. SPIE* **6808**, 68080M (2008).
28. P. Faye et al., "Perceptive free sorting and verbalization tasks with naive subjects: an alternative to descriptive mappings," *Food Qual. Preference* **15**(7–8), 781–791 (2004).
29. D. Picard et al., "Perceptual dimensions of tactile textures," *Acta Psychol.* **114**(2), 165–184 (2003).
30. G. Nyman, *Quality Experience Research: Trying to Understand the Modern Magazine Reader in Multimedia World*, PPA Professional Publishers Association, London (2002).
31. J. Radun et al., "Content and quality: interpretation-based estimation of image quality," *ACM Trans. Appl. Percept.* **4**(4), 1–15 (2008).
32. T. Leisti et al., "Process perspective on image quality evaluation," *Proc. SPIE* **6808**, 68080P (2008).
33. J. E. Radun et al., "Audiovisual quality estimation of mobile phone video cameras with interpretation-based quality approach," *Proc. SPIE* **6494**, 649403 (2007).

34. T. Shibata et al., "Evaluation of stereoscopic image quality for mobile devices using interpretation based quality methodology," *Proc. SPIE* **7237**, 72371E (2009).
35. O. Rummukainen et al., "Categorization of natural dynamic audiovisual scenes," *PLoS One* **9**(5), e95848 (2014).
36. J. Redi et al., "Interactions of visual attention and quality perception," *Proc. SPIE* **7865**, 78650S (2011).
37. Precision Vision, "Near visual acuity EDTRS," http://www.precision-vision.com/.
38. Stereo Optical Inc., "Near contrast vision F.A.C.T," http://www.stereooptical.com/.
39. Luneau ophtalmologie and VISIONIX, "Farnsworth D-15," www.luneau.fr.
40. X-Rite, "EyeOne Pro calibrator," www.x-rite.com.
41. International Organization for Standardization, "20462-2:2005 photography—psychophysical experimental methods for estimating image quality—part 2: triplet comparison method" (2005).
42. M. Nuutinen et al., "VQone MATLAB toolbox: a graphical experiment builder for image and video quality evaluations: VQone MATLAB toolbox," *Behav. Res. Methods* **48**(1), 138–150 (2016).
43. M. Mauri et al., "RAWGraphs: a visualisation platform to create open outputs," in *Proc. 12th Biannu. Conf. Ital. SIGCHI Chapter*, p. 28 (2017).
44. K. Linden and L. Carlson, "FinnWordNet– WordNet på finska via översättning," *LexicoNordica* **17**, 119–140 (2010).

Toni Virtanen received his MPsych degree from the Institute of Behavioural Sciences, University of Helsinki, in 2010. He is pursuing his PhD at the Department of Psychology and logopedics, Faculty of Medicine, University of Helsinki. His current research interests are in the areas of perceived experience of quality, attention, scene perception, human factors, and subjective assessment methods and analyses.

Mikko Nuutinen received his MSc (Tech.) and his LicSc. (Tech.) degrees from Helsinki University of Technology in 2004 and 2007, respectively, and his DSc (Tech.) degree from Aalto University, Helsinki, in 2012. His current research interests are in the areas of machine learning, advanced data analytics, computer vision, image and video signal processing, and subjective assessment methods and analyses.

Jukka Häkkinen received his PhD from the Institute of Behavioural Sciences, University of Helsinki. He worked as a principal scientist at Nokia Research Center and as an adjunct professor at the Department of Computer Science, Aalto University. Currently he is principal investigator at the Department of Psychology and logopedics, Faculty of Medicine, University of Helsinki, where he leads Visual Cognition Research Group. His interests include visual quality, attention, scene perception, and visual ergonomics of stereoscopic, head-mounted, and flexible displays.