CID2013: A Database for Evaluating No-Reference Image Quality Assessment Algorithms

Toni Virtanen, Mikko Nuutinen, Mikko Vaahteranoksa, Pirkko Oittinen, and Jukka Häkkinen

Abstract—This paper presents a new database, CID2013, to address the issue of using no-reference (NR) image quality assessment algorithms on images with multiple distortions. Current NR algorithms struggle to handle images with many concurrent distortion types, such as real photographic images captured by different digital cameras. The database consists of six image sets; on average, 30 subjects have evaluated 12–14 devices depicting eight different scenes for a total of 79 different cameras, 480 images, and 188 subjects (67% female). The subjective evaluation method was a hybrid absolute category rating-pair comparison developed for the study and presented in this paper. This method utilizes a slideshow of all images within a scene to allow the test images to work as references to each other. In addition to mean opinion score value, the images are also rated using sharpness, graininess, lightness, and color saturation scales. The CID2013 database contains images used in the experiments with the full subjective data plus extensive background information from the subjects. The database is made freely available for the research community.

Index Terms—No-reference image quality assessment algorithms, subjective image quality evaluation, test image databases.

I. INTRODUCTION

IN THIS study, we present a new image database: CID2013-Camera Image Database. In contrast to previous image databases, this database uses retail cameras instead of introducing distortions via post-processing. Retail cameras contain images that can have enhancements and distortions that are multidimensional and more subtle in nature. The database consists of six image sets where on average 30 subjects have evaluated 12–14 devices depicting 8 different scenes. The subjective evaluation method was a hybrid ACR-Pair Comparison developed for the study and presented in this study.

T. Virtanen and J. Häkkinen are with the Institute of Behavioral Sciences, University of Helsinki, Helsinki 00100, Finland (e-mail: toni.virtanen@helsinki.fi; jukka.hakkinen@helsinki.fi).

M. Nuutinen was with the Department of Media Technology, Aalto University, Espoo 02150, Finland. He is now with the Institute of Behavioral Sciences, University of Helsinki, Helsinki 00100, Finland (e-mail: mikko.nuutinen@helsinki.fi).

M. Vaahteranoksa was with Nokia, Espoo 02610, Finland. He is now with Microsoft Corporation, Espoo 02150, Finland (e-mail: mikko.vaahteranoksa@microsoft.com).

P. Oittinen is with the Department of Media Technology, Aalto University, Espoo 02150, Finland (e-mail: pirkko.oittinen@aalto.fi).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

A. Image Databases in Quality Assessment

The research field of image quality is multidisciplinary, composed of the primary disciplines of vision science, color, computational and behavioral sciences. Amongst the top priorities of the research is the creation of a computational model capable of predicting the subjective visual quality of natural images and video. An established practice is to validate and test the performance of a new image quality algorithm with publicly available image databases. These databases include sets of images that have undergone some type of distortion and have subjective data about the effects from the distortions. Publicly available image databases include: LIVE, TID2008, TID2013, IVC, MICT, LIVE Multiply Distorted Image Quality Database, BID, and CSIQ [1]–[8]. Table I compares the CID2013 database against other available databases. Some of the figures for the other databases originate from an excellent comparison of image and video databases by Winkler [9], and the remaining figures were gathered from the articles, online pages and the files of the corresponding databases. An extensive list of various multimedia databases is also curated by Qualinet [10].

The LIVE database has become the de facto standard for validating algorithms [11]–[15]. Recently, the TID2008 image database has been used in parallel with LIVE [13], [15]–[17]. The new TID2013 is probably going to replace the TID2008 in this respect. The models or quality assessment algorithms used to predict the quality of natural images can be divided into three types based on the availability or usage of a reference image: full-reference (FR), reduced-reference (RR) and no-reference (NR). The term "reference image" in the domain of quality assessment algorithms refers to an image whose visual quality or information capacity is high compared to that of the distorted images. An FR algorithm requires a pixel-wise reference image. An RR algorithm requires some information from the original or reference image. An NR algorithm does not need a reference or original image. The values computed by NR measures are based only on the information that is available from the evaluated image.

As there are no reference images available in many real-world applications, NR algorithms have a high research potential. However, the performance of NR measures remains limited. Conventionally, NR measures are based on the assumption that a specific and known distortion type has distorted the image [14]–[17]. NR algorithms perform poorly if the distortion space of the test image is multi-dimensional. NR algorithms cannot handle images with many concurrent distortion types, such as real photographic images captured

TABLE I
COMPARISON OF THE CID2013 DATABASE WITH COMBARABLE PUBLIC DATABASES

Database	IVC(I)[4]	LIVE(I)[1]	MICT[5]	TID2008[2]	TID2013[3]	CSIQ[8]	BID[7]	LIVE(MDIQ[6]	CID2013
Year	2005	2006	2008	2009	2013	2009	2011	2012	2013
Images	195	1011	196	1725	3025	930	585	405	480
Rated	185	779	196	1700	3000	866	585	405	480
Contents	10	29	14	25	25	30	585	15	8
Distortion types	4	5	2	17	24	6	5 cat.	3	12–14
Distrotion levels	5	5-9	6	4	5	4-5	N/A	16	N/A
Cameras	N/A	KODAK CD	N/A	KODAK CD+1	KODAK CD+1	N/A	N/A	N/A	79
Simultaneous distortions	1	1	1	1	1-2	1	multiple	2	multiple
Avg. no. of ratings per image	15	23	16	33	9	5-7	11	18–19	31
Method[1]	DSIS	ACR/SS-HR	ACR/SS	PC	PC	Custom	ACR/SS	ACR/SS-HR	ACR-DR
Data	DMOS	DMOS	RAW	MOS+δ	MOS+δ	DMOS+δ	RAW	DMOS+δ	RAW
Scale	1-5	0-100	1-5	0-9	0-9	0-1	0-5	0-100	0-100
Viewing distance[2]	6 Hs	2-2.5 Hs	4 Hp	varying	varying	70 cm	N/A	4 Hs	~80 cm
Screen	CRT	21" CRT	17" CRT	19" LCD & CRT + online	laboratory & online	24" LCD	17" CRT	LCD	24" LCD
Image resolution	512 x 512	~768 x 512	768 x 512	512 x 384	512 x 384	512 x 512	varied	1280 x 720	1600 x 1200
Display Gamut	[24]	N/A	N/A	varying	varying	sRGB	N/A	[24]	sRGB
Laboratory illumination	[24]	N/A	Low	varying	varying	N/A	N/A	"normal indoor illumination"	5800K, 20 lx ambient
Format	BMP	BMP	BMP	BMP	BMP	PNG	JPG	BMP	JPG
Subjects	15	20-29	16	838		25	180	37	188
Expert / Naive	expert	naive	naive	naive	naive	N/A	naive	naive	naive
Vision testing	N/A	confirmed verbally	N/A	No	No	N/A	N/A	confirmed verbally	acuity, contrast sensitivity, color vision
Age	N/A	students	students	N/A	N/A	21–35	N/A	23 - 30	18–44
Female	N/A	minority	N/A	N/A	N/A	N/A	N/A	minority	67 %

N/A = Not available or not applicable
[1] DSIS = Double-Stimulus Impairment Scale, ACR = Absolute Category Rating, SS = Single Stimulus, HR = Hidden Reference, DR = Dynamic Reference PC = Pair Comparison, see [24,25] for more details
[2] Hp = Picture height, Hs = Screen height

by different digital cameras, which can be dependent or independent of each other. Furthermore, in such situations some distortion sources are known, but others are unknown.

However, according to our knowledge, BID is the only publicly available database that includes images with distortion sources similar to those arising from a real image capturing process by a consumer camera. The value of multidimensionality is, however, limited, because the images have been subjectively rated based only on the aspect of blurriness. The LIVE Multiply Distorted Image Quality Database [6]

includes images that have undergone two specific types of distortion. Also TID2013 [18] contains one distortion type that has more than one source: lossy compression of noisy images. LIVE [1], TID2008 [2], IVC [4], and MICT [19] image databases include only images that have undergone some specific type of distortion, such as JPEG or JPEG2000 compression, noise contamination or low-pass filtering.

B. Contributions of CID2013

As a more realistic image database with multidimensional distortions would be useful for NR image quality algorithm assessment, we have collected CID2013 database for this need. The contributions of CID2013 are as follows: providing real photographs with many concurrent distortion sources, using a large number of different cameras to capture the photographs, measuring subjective attributes across many scale values and publishing the full raw data. CID2013 includes images that have been captured using 79 different cameras or image signal processing pipelines. Every camera introduces different distortion combinations, depending on the camera-specific sensor type, optics and image processing aims. The cameras range from low quality to high quality, including low-, moderate- and high-quality mobile phone cameras; moderate-quality compact cameras; and low- to moderate-quality SLR cameras. The image contents were inspired by the "photospace" approach defined by I3A [20]. The photospace describes the picture-taking frequency as a function of the subject illumination level and the subject-to-camera distance and scene descriptions. In summary, CID2013 is useful to validate whether a quality algorithm functions with images containing many concurrent distortion types.

The subjective evaluations were conducted in a controlled environment with calibrated monitors. A new method, Dynamic Reference (ACR-DR), was developed for subjective testing; it uses slideshows of images as references for the subject to calibrate their evaluation. In addition to collecting mean opinion score (MOS) from each image, the subjects also evaluated the images using 4 attribute scales: sharpness, graininess, brightness and color saturation. The CID2013 database includes the complete raw data from the subjective experiments instead of only pre-calculated mean opinion scores from each image. This inclusion allows for further analyses by those who wish to use this database and gives them more opportunities to utilize the data to its full potential. The shared data also include the background information of the subjects, which consists of the following: use of glasses, gender, age, level of education, photography habits, owned photography gear (DSLR, DSC, Mobile, none), photography sharing habits, post-editing habits, and whether the subject develops the digital photographs they have taken in print format. This background information can be beneficial for creating subsets from the data. One example of such a subset would be to test whether the image quality algorithms predict photography enthusiast evaluations better than non-enthusiasts.

Using the CID2013 image database, we have evaluated several state-of-the-art NR quality assessment algorithms. The evaluation study showed that a database of real photographic images captured by a number of different cameras is an important tool for effective and ecologically valid testing of different NR quality assessment algorithms, as well as for the design of new algorithms. CID2013 has been made publicly available for research purposes and it is freely downloadable from www.helsinki.fi/psychology/groups/visualcognition/.

This manuscript is organized as follows. In Section II, we present the various images used in CID2013 more thoroughly and explain how they were produced. In Section III, we describe the processes involved in image signal processing that generate the multidimensional distortions between devices in addition to optical and sensor differences. In Section IV, we present the new Dynamic Reference methodology of the subjective experiments and explore the data. In Section V, we provide the results of the state-of-the-art NR measure evaluation study. Section VI concludes the study.

II. IMAGE CONTENTS

The images in CID2013 are intended to represent typical photographs that consumers might capture with their cameras. The photographed scenes were based partly on the photospace approach described by I3A [20]. According to Segur [21], the photospace statistically describes the picture-taking frequency as a function of the subject illumination level L and the subject-to-camera distance D: PSD(L, D). The PSD is defined as a probability distribution of: "the probability that an image is taken within a certain limit of subject illumination level and within a certain range of subject-camera distance" [20].

Segur [21] distinguished the photospaces of photographic utilization and photographic motivation. The photographic utilization space is a graph that describes where the camera users take photographs. The photographic motivation space is a graph that describes where the camera users would take photographs if possible. For example, compared with the range of a low-end compact camera, the operating range of a high-quality SLR camera is extensive; with the telephoto lens of an SLR, it is possible to photograph distant objects that could not be captured by low-end compact cameras. The scenes used in the CID2013 database represent the photospace of photographic utilization. The photospace was derived especially from low-end compact cameras or mobile phone cameras.

The probability distribution of the I3A photospace was separated into clusters based on the most typical subject-camera distance and subject illumination combinations. These clusters represent the most typical scenes captured by users. For this study, we have developed clusters with the following image features in mind:

- be difficult to capture for typical low-end consumer camera,
- be able to differentiate consumer cameras,
- reveal camera-specific problems and
- represent views that typical consumer camera users might capture with their cameras.

At this moment the CID2013 database includes six datasets (I-VI). Every dataset includes six different scenes derived from the I3A clusters defined above. Every scene was

TABLE II
The Luminance, Shooting Distances and Scene Descriptions for the Images of CID2013

Cluster	Subject luminance (lux)	Subject-camera distance (m)	Scene description	Example images	Image set	Motivation
1	2	0.5	Close-up in dark lighting conditions		I-VI	Bar and restaurant setting
2	100	1.5	Close-up in typical indoor lighting conditions		I-VI	Living room environment, indoor portrait
3	10	4.0	Small group in dim lighting conditions		I-VI	Living room environment, group picture
4	1000	1.5	Studio image		I-IV	Studio image generally used in image quality testing
5	> 3400	3.0	Small group in cloudy bright to sunny lighting conditions		I-V	Typical tourist image
6	> 3400	> 50	Landscape image in cloudy bright to sunny lighting conditions		I-VI	Landscape image
7	> 3400	3.0	Small group in cloudy bright to sunny lighting conditions (~3x optical or digital zoom)		VI	General zooming situation
8	> 3400 (outdoors) and < 1000 (indoors)	1.5	Close-up in high-dynamic range lighting conditions		V,VI	High dynamic range scene

captured by 12–14 different cameras. To maintain consistent terminology, we will refer to the various scenes portrayed in Table II as image clusters. The table shows the descriptions and example images for each cluster. Table II also acts as a reference to the available cluster combinations for each image set and explains the motivation behind the type of scene that each image cluster is trying to simulate. Notice that not all image clusters are present in every image set.

III. Image Signal Processing

The ISP pipeline's input is typically a 10- to 14-bit raw image from an imaging sensor after analog-to-digital (AD) conversion and the output of the pipe is an 8-bit jpeg image. Raw data includes all of the artifacts from the imaging sensor and the optics may be affected by several non-ideal analog signals and distortions. Such as photon noise, thermal noise, pixel defects, pixel saturation, optical aberrations and spatial under-sampling. ISP processes the raw data and also controls the three "A's" of the camera: Auto-focus, Auto-exposure and Automatic white balance algorithms. A failed exposure or a failed focus decreases the final image quality considerably because the lost data cannot be fully recovered.

ISP is divided into dedicated sequential blocks, and each block is tuned depending on the sensor and optic characteristics [22]. The preprocessing blocks of the ISP handle data in the Bayer format (R, Gr, Gb and B). According to [23] and [22]; typical operations are defective pixel correction, noise removal, black level adjustment and color correction. Defective pixel correction is the process of replacing too-high or too-low pixel values by interpolation between neighboring pixels, essentially producing an image blur. Black-level adjustment is performed to equalize black levels
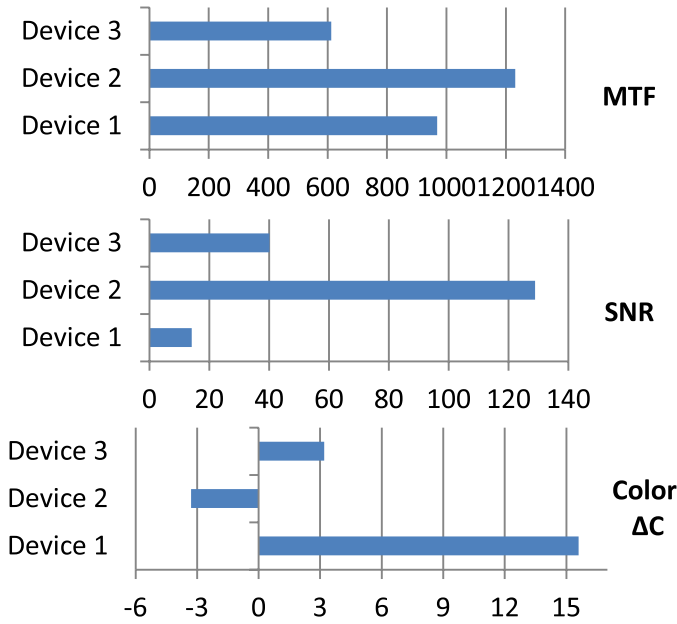
Fig. 1.   Test chart image analysis results.

and darken black areas. A black level that is set too high leads to the lightening of the dark areas and lower contrast. White balancing is performed to correct the color differences of illuminants. Global color errors such as a green overcast in the final image are often caused by insufficient white balancing. De-mosaicking is the process of interpolating missing color filter array-sampled pixel values. De-mosaicking causes aliasing. One example of spatial aliasing is the Moiré pattern. Aliasing can be avoided by using a low pass filter, but removing high frequencies causes the loss of fine details. Edge, contrast and color saturation enhancement operations are applied during post-processing to enhance the image before saving it to memory.

To give an example of the variation in CID2013, three devices were used to capture spatial frequency response (SFR), signal to noise ratio (SNR) and color test chart images. Images were analyzed using IQ-Analyzer (Image Engineering GmbH & Co. KG, Frechen, Germany) [24], and the results are shown on Fig. 1.

MTF and SNR for device 2 are at high levels but the color saturation is lower compared to devices 1 and 3. Variation in the results demonstrates a wide range of image quality measures and characteristics. Typically, a high MTF is the result of good optics, a good sensor and de-mosaicking. A high SNR is the result of a large pixel size and advanced noise filtering. Optimal color saturation is the combination of high quality color filters and color processing. Based on the results, the three devices and their pipelines differ considerably.

The three "A's" controls and other ISP algorithms are adaptive to shooting distance, content, illumination level and light spectrums. ISP tuning often compromises the image quality and latency, in addition to being limited by the raw data quality and the ISP algorithms.

Image databases such as LIVE [1], TID2008 [2], TID2013 [18], IVC [4], and MICT [19] are produced using

the same raw data but varying only the processing parameters, adding digital noise or blurring the image. As discussed above and shown in Fig. 1, many non-ideal signals already exist in the raw data and are corrected at the hardware level in the analog domain. Real DSC, SLR and mobile devices differ in sensor size, pixel count, CFA, optic, focus and ISP algorithms and their parameters. These differences introduce a highly multi-dimensional distortion space for the images. However, it is possible to name some general rules on what kinds of distortions are most prominent in different scene contents. Noise, sharpness and exposure variations seem to be the most common distortion types between devices in all scene contents. Low-light indoors scenes bring out more noise and exposure distortion with some auto white-balance issues. Brighter outdoor scenes, have less visible noise, various color differences and slight luminance shading.

IV. SUBJECTIVE EXPERIMENTS

A. Difficulties in Subjective Evaluations

Reliable subjective evaluations to represent the ground truth are one of the most important components for the evaluation and benchmarking of new image quality assessment algorithms [9]. However, subjective evaluation can include multiple sources of variation and error that must be controlled by strict test design. Controlling unwanted variation or noise is especially demanding when the test material has multidimensional degradations as with the CID2013 material.

One of the ways to improve the reliability of perceptual quality evaluations is to provide the observers with external anchors. This anchoring is similar to image quality assessment algorithms that can benefit when there is some reference information available. Thus, comparison tasks should be used preferentially over individual single-stimulus (SS) evaluations when the stimuli are multidimensional. Without any reference, the observers create their own standards for image quality that are unknown to the researcher and can cause unwanted variation.

The best choice would be to obtain direct comparisons between all combinations of stimulus pairs [25]. This pair-comparison method [26], [27] is unfortunately not always a viable option because of the limitations of subject fatigue [1], [26], [28]. As the number of testable stimuli increase, the number of comparisons increases exponentially. As a consequence, the Pair Comparison method is only viable when the number of images and scenes are small.

Another effective way to control unwanted variation is to use a benchmark. If direct comparison for each stimulus is not possible, a comparison against a reference stimulus whose properties are known is a viable option. One example of such a method is the Degradation Category Rating (DCR) [27], also known as Double-Stimulus Impairment Scale (DSIS) [26]. Using a reference gives the benefit of anchoring the evaluations against a known stimulus. However, even if a reference is selected, the evaluation of an image is rarely made in isolation from the other images in the test. In other words, the images in the test set with corresponding content will create an evaluative context for the other images in the experiment [29].
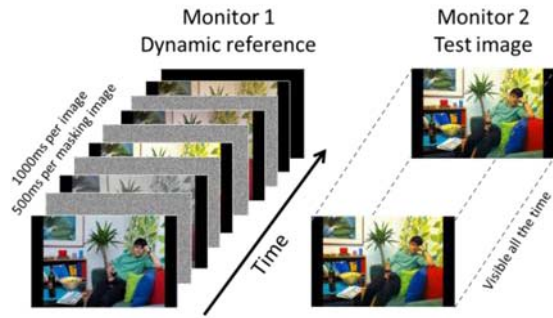
Picture 1.   Dynamic Reference presentation method.

B. Dynamic Reference Method

Our aim was to combine the benefits of direct comparison from the pair comparison method with the efficiency of single-stimulus methods by developing a Dynamic Reference method. The observer sees a slideshow of the test images with the corresponding content, i.e., the dynamic reference, prior to their evaluation (see Picture 1). As the observer views the other test images in the slide show, they form a general idea of the overall quality variation within the set of test images.

The test images in the dynamic reference slideshow are presented in random order for 1000 ms each, separated by a 500 ms white noise masking image with a center fixation point. The masking image effectively removes the illusion of movement between the images, preventing the attention of the subject from being diverted by differences in the image perspective by clearing the iconic memory buffer in the human visual system (HVS) [30]. During scene viewing, the eyes fixate three times each second on average by saccadic eye movements to bring the projection of a local scene region onto the area of the fovea, producing the highest acuity vision [31]. This assessment indicates that 1000 ms is long enough to see locally visible quality artifacts in the image but does not prolong the slideshow of images unnecessarily. A video of the Dynamic Reference method is provided at www.helsinki.fi/psychology/groups/visualcognition/.

After viewing the slideshow, observers had unrestricted time to examine the test image and give a general quality evaluation on a graphical scale of 0 to 100. Subjects did not know how many steps are in the slider to avoid the tendency to favor certain numbers, e.g., even tens and quartiles. They also view a record of their answers on a line graph on the right of the GUI (see Picture 2), helping them to remember their previous answers and encouraging them to use the whole scale. After giving an answer to the overall quality scale they were instructed to view the slideshow a second time and give specific evaluations of their preferences in four different scales (Picture 2). Sharpness and graininess scales also have a 0–100 range. Lightness, which represents the overall exposure of the image, is a bipolar scale ranging from −100 to 100. The value of 0 in the middle of the scale indicates perfect exposure while values below it are too dim and values above it are too bright. The saturation scale follows the same guideline with pale and loud colors at the extremes. These scales were collected for every image set (except that saturation is missing from image set VI due to a technical error). Only the far ends
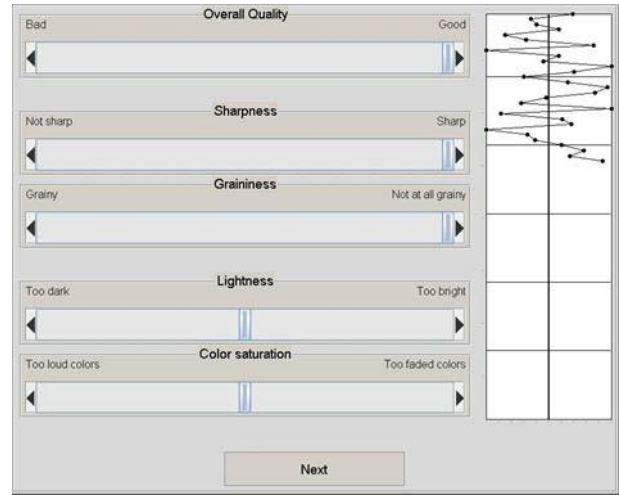
Picture 2.   GUI for the subjects.

TABLE III
OVERVIEW OF THE IMAGE SETS

Image set	Subjects	Cameras	Clusters	Scale anchoring	Duplicate	scales
I	30	14	1,2,3,4,5,6	Yes	0	Sharpness Graininess Lightness Saturation
II	32	14	1,2,3,4,5,6	Yes	1	Sharpness Graininess Lightness Saturation
III	31	13	1,2,3,4,5,6	Yes	0	Sharpness Graininess Lightness Saturation
IV	26	13	1,2,3,4,5,6	No	0	Sharpness Graininess Lightness Saturation
V	34	12	1,2,3,5,6,8	No	0	Sharpness Graininess Lightness Saturation
VI	35	14	1,2,3,6,7,8	No	0	Sharpness Graininess Lightness

of the scale were labeled to reduce variance, as studies have shown that subjects might have varying interpretations for the adjectives that are provided [32]. For example, the subjective distance between poor and bad are not necessarily the same as the subjective distance between good and excellent.

C. Image Sets

Image sets I-III differ from image sets IV-VI by the scale use in the MOS score. In image sets I-III, the observers were instructed to anchor their evaluations by giving a score of 100

Picture 3.   Illustration of the lab setup.

to the best image in the image cluster and a score of 0 to the worst image in the image cluster. The idea behind this was to have the observers create physical anchors for the scale in each cluster, allowing for combining the data between image sets. See Table III for a summary of differences between the image sets. However, observers soon reported that locking one image as the best and one as the worst for each image cluster was a very difficult task. This difficulty did not translate to the data, however, because it yielded reliable results. The tests for image set II had a secret repetition of a single image in all image clusters.

The within-subject correlation between individual evaluations of the same image were very high (r = 0.908, p < 0.01) and Cronbach's Alpha reliability coefficient also gave a very good measure of reliability (α = 0.951). This is a measure of internal consistency, gaining values between 0–1. A high alpha coefficient indicates that the there is a common underlying construct that is measured with substantial inter-relatedness of the items within the test [33]. However, the data indicated that there was variation between subjects as to which images were selected as the best and the worst. Unfortunately, this outcome prevented us from combining the separate image set data into one dataset, and the instruction was changed for image sets IV-VI to not force the observer to select the best and worst image within an image set cluster.

D. Test Environment and Setup

We used two Eizo ColorEdge CG241W 24" monitors, with a third smaller display underneath for presenting questions (see Picture 3). The data collection and image presentation software was created using MATLAB. The room was covered with medium gray curtains to diffuse the ambient illumination. Fluorescent lights (5800K) were positioned behind the monitors and reflected from the back wall covered with grey curtain to create dim and uniform ambient illumination in the room. In dim viewing conditions backlighting has been found to reduce eyestrain and visual fatigue [34]. The light hitting the monitors measured below 2 lx, and the ambient illumination from behind the monitors were 20 lx. The subject's viewing distance (approximately 80 cm, 2 1/2 picture heights) 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 resulting in a horizontal size of 30 degrees of visual angle and a resolution of 52 pixels per degree of visual angle. Monitors were calibrated to sRGB having target values of: 80 cd/m2, 6500K and gamma 2.2 using EyeOne Pro calibrator (X-rite co. Grand Rapids, MI, USA) [35]. A relatively low luminance on the monitors was selected as in dim viewing environment bright monitors would cause unnecessary eyestrain and subject fatigue. Gamma curves and chromaticity coordinates of R,G,B for all monitors are shared along with the database as well as provided at www.helsinki.fi/psychology/groups/visualcognition/.

E. Subjects and Data Preparation

Subjects (n = 30, 32, 31, 26, 34 and 34 for image sets I, II, III, IV, V and VI, respectively) were naïve in the sense that they did not study or work with image quality or in related fields. They were recruited through student mailing lists consisting mainly of humanities and behavioral science students. Background information was collected when the subjects reserved time for the test, which includes the following information: use of glasses, gender, age, level of education, photography habits, owned photography gear, photography sharing habits, post-editing habits, and whether the subject develops the digital photographs they have taken in print format. High proportion of the subjects were female (67%), the observers' vision was controlled for near visual acuity EDTRS (Precision Vision, La Salle, IL, USA) [36], near contrast vision F.A.C.T. (Stereo optical co. inc., Chicago, IL, USA) [37] and color vision Farnsworth D-15 (Luneau ophtalmologie, Chartres, France) [38] before participation. They received two movie tickets as a reward. On average, the experiment lasted 93 minutes. However, that time includes the visual testing, instructions and training for the observers. The observers were also able to have a break if they felt they needed one. Still the test duration was a bit over the total maximum experiment duration of 60 minutes stated by [28].

Six subjects participated in more than one image set study and are marked by an identical subject ID followed by the image set number; thus, the final number of individual subjects is 182. The background information for seven subjects is missing as a result of a recording error. Similar technical errors also caused missing data in a few random values from the scales: MOS (0.09%), Sharpness (0.51%), Graininess (0.34%), Lightness (0.38%) and Color Saturation (1.05%).

As the Dynamic Reference method is derived from ACR method and uses similar scale akin to it, same outlier removal methods are applicable to Dynamic Reference data as well. The Mean Opinion Score (MOS) was calculated from the overall quality evaluation scores and screened for outliers using recommendation in ITU-R BT.500-13 [26]. Thirteen cells of data were flagged as outlier and removed. The missing data and deleted outliers were excluded pairwise from further analyses. No outlier screening was done on the separate attribute scales.
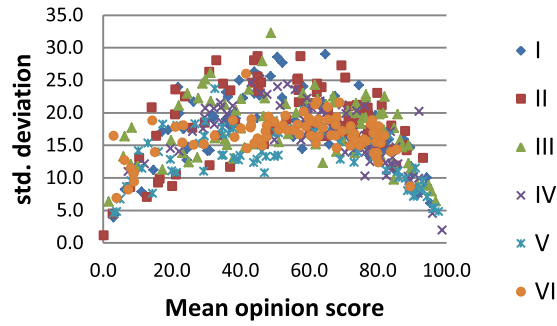
Fig 2.   Standard deviation as a function of MOS in each image set (I-VI).

TABLE IV
CORRELATIONS OF ATTRIBUTE SCALES AND MOS

Cluster Attribute	1	2	3	4	5	6	7	8	ALL
Sharpness	0.78	0.69	0.75	0.61	0.68	0.73	0.68	0.72	0.72
Graininess	0.79	0.71	0.78	0.54	0.60	0.62	0.69	0.74	0.71
Lightness neg	0.81	0.44	0.74	0.46	0.41	0.48	0.44	0.42	0.62
Lightness pos	-0.42	-0.15	-0.19	-0.34	-0.43	-0.32	-0.47	-0.16	-0.32
Saturation neg	0.42	0.49	0.46	0.38	0.30	0.49	a	0.42	0.43
Saturation pos	-0.32	-0.36	-0.37	-0.42	-0.43	-0.38	a	-0.2	-0.36

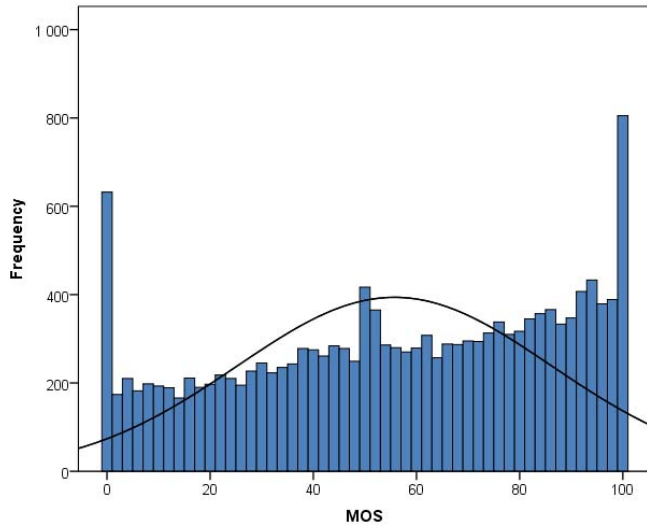a cluster 7 was only in Image Set VI where no saturation scale were present

Fig 3.   Histogram of all data. Notice that the far ends were locked in image sets I, II and III, making the peaks stand out. Otherwise the distribution is quite uniform with a slight tendency towards higher values.

F. Distribution and Variance of the Data

As observed in Fig. 2, the variation diminishes in all image sets when approaching the extreme values of the MOS scale. For image sets I-III, this was expected, as the extreme values of the scale only received one rating because of the different instructions to the observers (see Section IV). Still, the same effect is observed in image sets IV-VI. This inverted U-shape is very common in databases and subjective experiments of quality evaluation [39]. This is due to the clipping of the ratings at the far ends of the scale [39]. The higher variation in the middle of the MOS scale could also be interpreted to mean that it is easier to evaluate images that are either very bad or very good, resulting in reduced variation at the far ends [21].

The histogram in Fig. 3 shows that the data have a uniform distribution aside from the peaks at 0 and 100, which are due to the instructions for image sets I-III. A uniform distribution of the data is desired when using the dataset for testing image quality assessment algorithms. A uniform distribution indicates that the dataset contains equal frequencies of images with varying levels of quality, from low to high.

G. Attribute Scales

Table IV shows Pearson's correlation coefficients between the attribute scales and MOS for each Image Cluster.
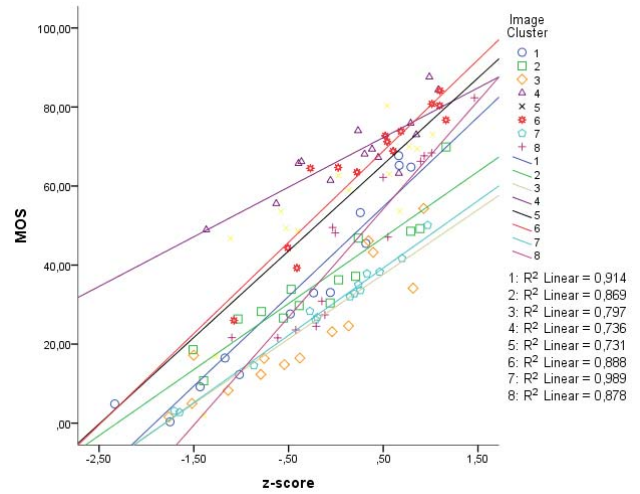

Fig 4.   Scatterplot and linear regressions from realignment study between MOS and z-scores.

The bi-polar scales, color saturation and lightness were divided into unipolar scales where negative sign relates to faded or dark image and positive image over saturated or bright image. The negative scale had values between −100 to 0 and the positive scale had values from 0 to 100. Sharpness and graininess seems to be the most important factors for the naïve subjects when they evaluate image quality in general. Image darkness (Lightness −) was determining factor for image quality in image clusters 1 and 3.

H. Alignment Study

As the Image Sets can be considered having a self-contained set of references the MOS values cannot be aggregated into one scale without a realigning study [1], [40]. To fit as many images to the scale realignment study while keeping the experiment duration within recommended duration a single stimulus ACR method was chosen. The study consisted of 34 (85% female) observers, with normal or corrected to normal vision, evaluating 112 images in randomized order using the same 0-100 scale. The Images were selected to roughly represent the overall scope of image quality variation from each Image Set and Cluster combination so that each Cluster had 14 images. The subjects had a training session by evaluating 24 images selected to represent the overall set of images in the scale alignment experiment. The viewing environment was in other respects the same as described in section *D*. The total

TABLE V

IMAGE QUALITY METRICS TESTED IN THIS STUDY

Metric	Description	Public implementation available at
BIQI [10]	Learning based approach. LIVE database has been used for training.	http://live.ece.utexas.edu/research/quality/BIQI_release.zip
BRISQUE [12]	Learning based approach. LIVE database has been used for training.	http://live.ece.utexas.edu/research/quality/BRISQUE_release.zip
NIQE [13]	Learning based approach. Natural images (Flickr images and Berkeley image segmentation database) have been used for training	http://live.ece.utexas.edu/research/quality/niqe_release.zip
BLIINDS-II [39]	Learning based approach. LIVE database has been used for training.	http://live.ece.utexas.edu/research/quality/BLIINDS2_release.zip
DESIQUE [40]	Learning based approach. LIVE database has been used for training.	http://vision.okstate.edu/yi/code/DESIQUE_release.rar
CPBD [11]	Distortion-specific sharpness metric.	http://ivulab.asu.edu/software
FISH and FISH_bb [41]	Distortion-specific sharpness metric.	http://vision.okstate.edu/phongvu/code/FISH.rar
S3 [42]	Distortion-specific sharpness metric.	http://vision.okstate.edu/s3/
LPC [44]	Distortion-specific sharpness metric.	https://ece.uwaterloo.ca/~rhassen/LPC-SI/
DIIVINE [53]	Learning based approach. LIVE database has been used for training.	http://live.ece.utexas.edu/research/quality/DIIVINE release.zip
Martziliano [45]	Distortion-specific sharpness metric.	http://ivulab.asu.edu/software
NJQA [43]	Distortion-specific JPEG compression artifact metric	http://vision.okstate.edu/njqa/

TABLE VI

LINEAR CORRELATION COEFFICIENTS AFTER NONLINEAR REGRESSION (MOS)

Cluster / Algorithm	1	2	3	4	5	6	7	8	ALL
FISH_bb	**0.79**	0.56	**0.84**	0.51	0.57	**0.72**	0.78	**0.72**	**0.62**
FISH	0.74	0.48	0.70	0.54	0.54	0.71	0.78	0.67	0.58
S3	0.73	0.47	0.75	0.42	0.44	0.63	0.73	0.63	0.55
BRISQUE	0.58	0.50	0.57	0.01	**0.58**	0.61	**0.84**	0.20	0.49
BLIINDS-II	0.66	0.44	0.58	0.45	0.48	0.45	0.57	0.46	0.47
BIQI	0.24	**0.59**	0.72	0.47	0.44	0.45	0.75	0.74	0.44
DESIQUE	0.50	0.34	0.60	0.21	0.33	0.40	0.65	0.59	0.39
LPC	0.70	0.56	0.69	0.24	0.08	0.25	**0.84**	0.34	0.38
NIQE	0.46	0.40	0.40	0.25	0.52	0.42	0.53	0.30	0.38
DIIVINE	-0.38	0.29	0.54	**0.57**	0.51	0.24	0.77	0.25	0.23
Martziliano	-0.32	0.34	0.06	0.45	0.49	0.61	0.64	0.18	0.23
CPBD	-0.43	0.22	0.09	0.39	0.51	0.53	0.55	0.21	0.17
NJQA	0.12	0.02	0.24	0.07	0.34	0.22	0.57	0.48	0.15

outlier screened Image Set specific MOS were for each corresponding Image Set I 0.83, II 0.83, III 0.85, IV 0.84, V 0.89, VI 0.89. If we consider the Image Set specific MOS values as ground truth, the realigned MOS values are only approximate of those values. In practice the realigned MOS values precision to model the Image Set specific MOS is equivalent to the best state of the art image quality assessment algorithms.

V. EVALUATION OF QUALITY ASSESSMENT ALGORITHMS

The image quality assessment algorithms selected and tested in this study are summarized in Table V. Because the photographs in the CID2013 database were captured by real cameras and reference images are missing, we omitted the FR and RR quality assessment algorithms. We tested only NR algorithms. The other requirement for selection was that an implementation of the algorithm should be publicly available on the Internet.

The tested algorithms follow different approaches for no-reference image quality assessment. NIQE [14], BRISQUE [13], BLIINDS-II [41], DESIQUE [42] and DIIVINE [43] are distortion-agnostic quality algorithms. CPBD [12], FISH [44], FISH_bb [44], S3 [45], NJQA [46], LPC [47] and Martziliano [48] are distortion-specific quality algorithms. Distortion-specific quality algorithms are designed for measuring a specific distortion type from an image. Distortion-agnostic quality algorithms try to measure the quality of an image without knowledge of distortion types. In principle, BIQI [11] is distortion-specific quality algorithms, but it is capable of measuring more than one specific distortion type.

Before evaluating the performance of an algorithm, it is common to apply a logistic transform to the predicted scores to bring the predicted (objective) and measured (subjective) values to the same scale and to account for the nonlinear relationships between values [1], [49]. We used a logistic

experiment duration was 34 minutes on average, including vision tests, instructions and training. Outlier screening was done using ITU-R BT.500-13 recommendation and no outliers were found. The training session data was used for reliability analysis, the average Cronbach's Alpha reliability coefficient was 0.930. The data from the Scale Realignment test with similar background information is also shared along with the CID2013 database.

To realign the CID2013 scale, MOS in CID2013 were transformed into Z scores as described in [1]. Averaged Z scores for each image were then compared against averaged Image Set specific MOS scores of 112 images that were part of the scale re-alignment study, see Fig. 4. Linear regression analysis was done for each Image cluster to obtain MOS values for the entire CID2013 database. The Pearson linear correlation between the realigned MOS score and the

TABLE VII
LINEAR CORRELATION COEFFICIENTS AFTER NONLINEAR
REGRESSION (REALIGNMENT MOS)

Cluster / Algorithm	1	2	3	4	5	6	7	8	ALL
FISH_bb	**0.69**	0.55	0.31	0.12	0.48	**0.68**	0.79	0.30	**0.49**
S3	0.65	0.43	0.27	0.05	0.44	0.61	0.73	0.19	0.48
FISH	0.67	0.45	0.25	0.17	0.41	0.61	0.79	0.13	0.46
BRISQUE	0.50	0.47	**0.36**	0.40	0.46	0.52	**0.84**	0.27	0.45
BLIINDS-II	0.64	0.45	0.12	**0.54**	**0.52**	0.18	0.57	0.57	0.40
DESIQUE	0.47	0.30	0.27	0.17	0.36	0.21	0.65	0.26	0.39
LPC	0.57	0.55	0.24	-0.09	0.20	0.35	0.85	0.40	0.34
BIQI	0.12	**0.58**	0.28	0.37	0.32	0.46	0.76	**0.66**	0.28
DIIVINE	-0.47	0.25	0.19	0.34	0.42	0.51	0.78	-0.18	0.26
Martziliano	-0.33	0.28	0.01	0.19	0.43	0.43	0.64	0.13	0.26
NIQE	0.44	0.36	0.20	0.11	0.47	0.38	0.53	0.28	0.22
CPBD	-0.43	0.21	-0.03	0.36	0.44	0.36	0.56	0.22	0.19
NJQA	0.10	0.11	0.06	0.25	0.22	0.44	0.57	0.26	0.18

TABLE VIII
RANK-ORDER CORRELATION COEFFICIENTS (REALIGNMENT MOS)

Cluster / Algorithm	1	2	3	4	5	6	7	8	ALL
FISH_bb	**0.71**	0.55	0.38	0.24	0.46	**0.72**	0.52	0.24	**0.46**
S3	0.63	0.44	0.33	0.20	0.43	0.62	0.56	0.16	0.45
FISH	0.66	0.45	0.28	0.29	0.37	0.61	0.54	0.16	0.41
BRISQUE	-0.48	-0.52	-0.38	-0.31	-0.38	-0.49	**-0.69**	-0.26	-0.38
BLIINDS-II	-0.58	-0.42	-0.15	**-0.51**	**-0.56**	-0.17	-0.46	-0.61	-0.33
BIQI	-0.13	**-0.58**	-0.33	-0.25	-0.29	-0.48	-0.36	**-0.66**	-0.33
LPC	0.60	0.51	**0.45**	-0.13	0.05	0.25	0.50	0.52	0.32
DESIQUE	-0.48	-0.27	-0.22	-0.19	-0.31	-0.20	-0.54	-0.24	-0.30
DIIVINE	0.47	-0.29	-0.23	-0.34	-0.39	-0.47	-0.60	0.10	-0.28
Martziliano	0.30	-0.29	0.08	-0.13	-0.27	-0.36	-0.24	0.13	-0.25
NIQE	-0.49	-0.36	-0.21	-0.05	-0.39	-0.32	-0.64	-0.32	-0.23
CPBD	-0.37	0.16	-0.06	0.35	0.34	0.28	0.31	-0.04	0.19
NJQA	-0.13	0.01	-0.07	-0.23	-0.20	-0.39	-0.23	0.00	-0.15

function with an added linear term:

$$f(x) = \beta_1 \left(\frac{1}{2} - \frac{1}{1 + \exp(\beta_2 (x - \beta_3))} \right) + \beta_4 \cdot x + \beta_5 \quad (1)$$

where $\beta_1, \beta_2, \beta_3, \beta_4$ and β_5 are the model parameters chosen to minimize the MSE between the predicted and the subjective scores.

Two performance measures were used to evaluate the tested quality assessment algorithms. The first measure was the linear correlation coefficient (LCC), which measures the prediction accuracy. The second measure was the Spearman rank order correlation (SROCC), which measures the relative monotonicity between the predictions and the subjective scores. LCC values were calculated after a nonlinear regression. The larger values of LCC and (absolute) SROCC denote that the objective and subjective scores correlate well (a better performance of the algorithm).

Tables VI-VIII show overall and cluster-specific performance for the tested quality assessment algorithms.
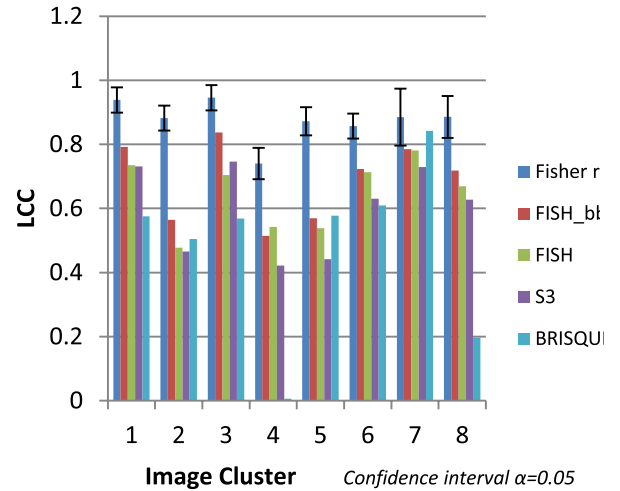

Fig 5. The four best performing algorithms against the Fisher z-transformed average of all the subjects' correlation to the overall mean of each cluster.

Table VI shows the Pearson's correlation against the Image set specific MOS for the tested quality assessment algorithms. Table VII and VIII show the LCC and SROCC between the realigned MOS and image quality assessment algorithms. Cluster-specific performance was calculated as an average over all datasets. Overall performance was calculated over all images from all datasets. Note that, before the logistic function parameters were estimated for the original MOS values, the metric values were normalized in image-set– and cluster-specific ways.

The boldface values indicate the best performer for a cluster. The quality assessment algorithms are sorted in the order of the values over all images. The results show that the performance of FISH_bb, FISH, S3 and BRISQUE are high compared to the other algorithms. The cluster-specific LCC values of FISH_bb were the highest except for clusters 2, 4, 5 and 7. In these cases, the performance of BIQI, FISH or BRISQUE algorithm was the best.

Having the full data from each individual observer we can calculate how much correlation there is between an individual evaluation and the overall mean. Figure 5 compares the four best performing algorithms even further. We calculated LCC for the subject ratings against the Image Cluster average (MOS), and averaged the correlations after Fisher z-transformation. The averaged Z values were then transformed back to r values according to model presented in [50] for easier comparison. This value can be considered representing the overall accuracy that the subjective data achieves for each Cluster. In other words, when the correlation of Fisher r is low there is less agreement between subjects so it is acceptable for an IQA to have lower correlation as well. What can also be derived from Fig. 5 is that when a IQA's correlation reaches the confidence interval of the Fisher r, it can be said to produce equally reliable estimate as with selecting a random observer from the data would on 95% of cases.

To determine which performance differences between the quality assessment algorithms are statistically significant, we performed a variance test. The test is the same as the one used in previous studies [1], [51]. The assumption is that

TABLE IX

STATISTICAL ANALYSIS OF ALGORITHM PERFORMANCE: A VALUE OF '1' IN CELL INDICATES THAT THE ROW (MEASURE) IS STATISTICALLY BETTER THAN THE COLUMN (MEASURE). A VALUE OF '0' INDICATES THAT THE ROW IS WORSE THAN THE COLUMN. A VALUE OF '-' INDICATES THAT THE ROW AND COLUMN ARE STATISTICALLY IDENTICAL

	FISH_bb	FISH	S3	BRISQUE	BLIINDS-II	BIQI	DESIQUE	LPC	NIQE	DIIVINE	Martziliano	CPBD	NJQA
FISH_bb	-	-	-	1	1	1	1	1	1	1	1	1	1
FISH	-	-	-	-	-	1	1	1	1	1	1	1	1
S3	-	-	-	-	-	-	1	1	1	1	1	1	1
BRISQUE	0	-	-	-	-	-	-	-	-	1	1	1	1
BLIINDS-II	0	-	-	-	-	-	-	-	-	1	1	1	1
BIQI	0	0	-	-	-	-	-	-	-	1	1	1	1
DESIQUE	0	0	0	-	-	-	-	-	-	-	-	-	-
LPC	0	0	0	-	-	-	-	-	-	-	-	-	-
NIQE	0	0	0	-	-	-	-	-	-	-	-	-	-
DIIVINE	0	0	0	0	0	0	-	-	-	-	-	-	-
Martziliano	0	0	0	0	0	0	-	-	-	-	-	-	-
CPBD	0	0	0	0	0	0	-	-	-	-	-	-	-
NJQA	0	0	0	0	0	0	-	-	-	-	-	-	-

the residuals (the difference between the subjective scores and the predicted scores) are normally distributed. We tested the normality using a kurtosis-based criterion, which labels the residuals as Gaussian if the kurtosis is between 2 and 4 [51]. The F-test was used to determine whether the variances of the residuals were identical, i.e., whether the two sample sets came from the same distribution. The null hypothesis was that the residuals of both measures are expressions from the same distribution and are statistically indistinguishable with 95% confidence.

According to the kurtosis-based criterion, the assumption of Gaussian residuals was met for all of the algorithms. The F-test results are shown in Table IX. A value of '1' in the table indicates that the row (algorithm) is statistically better than the column (algorithm); a value of '0' indicates that the row is worse than the column and a value of '-' indicates that the row and column are statistically identical. Table IX validates our observations from the performance measures: FISH_bb, FISH and S3 performed best.

The performance values of the best performer algorithms are low compared to the results of the earlier studies with previously published databases [11], [13]. For example the three best performing algorithms FISH_bb, FISH and S3 have a Pearson correlation with the LIVE database of 0.944, 0.904 and 0.943 respectively. Whereas correlation with CID2013 Image Set specific MOS scores were only 0.62, 0.58 and 0.55. This is an expected result because the quality assessment algorithms were developed for images with a single distortion source. The results clearly show that there is significant room for improvement. The algorithms were tested using default parameter values. The performance can increase if algorithms are trained with real photographic images with a multi-dimensional distortion space. In addition, some pooling strategy for the features of the best performing algorithms could increase the performance.

VI. DISCUSSION

This study presented a new method for subjective testing: the Dynamic Reference method (ACR-DR). It is an effort to mix the best of two of the most common subjective methods for image quality evaluation: the ACR and the Pair Comparison. We recognize that the method still needs some fine tuning and further work, especially in finding the optimal presentation duration for the slideshow. Some studies suggest that the duration of the masking image presentation can probably be shortened from 500ms as even 80ms presentation of the masking image could be sufficient to remove the illusion of movement [52]. Although, the longer 500ms duration allows the observer to have one confirming fixation on the test image between the reference slides if necessary. The cons of ACR-DR as compared to ACR and SS, are the longer test duration and the limitations for the maximum number of stimulus images within an experiment, before it becomes too long to watch every time. The pros of the method is that it give some point of reference to the subjective evaluations when there isn't a clear reference, like the original undistorted image, available. It also makes sure the observers have the same mental representation of the total quality distribution of the image set in the experiment.

We strongly advocate for releasing the complete data with the images. Although releasing only the MOS makes the utilization of the database more straightforward, we believe that it will be more beneficial to the imaging science community in the long run to have the complete data for analysis. It might make the existing image databases also more relevant and future proof as new ideas for analyses can be tested when the complete data is available.

Our aim was to create an image database that would represent the images that people take with common devices that would suit especially as a tool for no-reference image quality algorithm assessment. The images contain multiple

overlaying processes resulting in complex variation from image enhancements to degrading distortions, creating a real challenge to current and future image assessment algorithms. Real DSC, SLR and mobile devices differ in sensor size, pixel count, color filter array, optic, focus and ISP algorithms and their parameters. As discussed above, many non-ideal signals already exist in the raw data from the imaging sensor and are corrected at the hardware level in the analog domain. Real devices also produce distortions and possible image enhancements what ordinary people come across in their everyday life. The CID2013 database is considered to have somewhat different role in IQA algorithm development than those databases having only single distortion sources with well-defined levels like LIVE [1] and TID2013 [18]. It is meant to give a challenging real life test bed for the NR-IQA algorithm development, complementing the existing single distortion databases as a tool for image quality assessment algorithm development and research."

ACKNOWLEDGMENTS

REFERENCES

[1] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[2] N. Ponomarenko *et al.*, "TID2008—A database for evaluation of full-reference visual quality assessment metrics," *Adv. Modern Radioelectron.*, vol. 10, no. 4, pp. 30–45, 2009.

[3] N. Ponomarenko *et al.*, "Image database TID2013: Peculiarities, results and perspectives," *Signal Process. Image Commun.*, doi:10.1016/j.image.2014.10.009.

[4] P. Le Callet and F. Autrusseau. (2005). *Subjective Quality Assessment IRCCyN/IVC Database*. [Online]. Available: http://www2.irccyn.ec-nantes.fr/ivcdb/, accessed Jul. 25, 2013.

[5] Y. Horita. (2008). *MICT Image Quality Evaluation Database*. [Online]. Available: http://mict.eng.u-toyama.ac.jp/mictdb.html, accessed May 12, 2014.

[6] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik, "Objective quality assessment of multiply distorted images," in *Proc. 46th Asilomar Conf. Signals, Syst. Comput.*, 2012, pp. 1058–1697.

[7] A. Ciancio, A. L. N. T. da Costa, E. A. B. da Silva, A. Said, R. Samadani, and P. Obrador, "No-reference blur assessment of digital pictures based on multifeature classifiers," *IEEE Trans. Image Process.*, vol. 20, no. 1, pp. 64–75, Jan. 2011.

[8] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, pp. 11006–11021, 2010.

[9] S. Winkler, "Analysis of public image and video databases for quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 616–625, Oct. 2012.

[10] K. Fliegel, C Timmerer. (Mar. 2014). *Qualinet Databases (Version 2.5)*. [online]. Available: http://www.dbq.multimediatech.cz

[11] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 513–516, May 2010.

[12] N. D. Narvekar and L. J. Karam, "A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection," in *Proc. Int. Workshop Quality Multimedia Exper. (QoMEx)*, Jul. 2009, pp. 87–91.

[13] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[14] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[15] L. Zhang, D. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[16] J. Zhu and N. Wang, "Image quality assessment by visual gradient similarity," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 919–933, Mar. 2012.

[17] L. Capodiferro, G. Jacovitti, and E. D. Di Claudio, "Two-dimensional approach to full-reference image quality assessment based on positional structural information," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 505–516, Feb. 2012.

[18] N. Ponomarenko *et al.*, "A new color image database TID2013: Innovations and results," in *Proc. 15th Int. Conf. ACIVS*, 2013, pp. 402–413.

[19] Y. Horita, K. Shibata, and K. Yoshikazu. *MICT Image Quality Evaluation Database*. [Online]. Available: http://mict.eng.u-toyama.ac.jp/mictdb.html, accessed Jul. 25, 2013.

[20] *CPIQ Inititiave Phase 1 White Paper*. [Online]. Available: www.i3a.org, accessed May 16, 2008.

[21] R. Segur, "Using photographic space to improve the evaluation of consumer cameras," in *Proc. PICS Conf.*, 2000, pp. 221–224.

[22] R. Ramanath, W. E. Snyder, Y. Yoo, and M. S. Drew, "Color image processing pipeline," *IEEE Signal Process. Mag.*, vol. 22, no. 1, pp. 34–43, Jan. 2005.

[23] J. Zhou and J. Glotzbach, "Image pipeline tuning for digital cameras," in *Proc. IEEE Int. Symp. Consum. Electron. (ISCE)*, Jun. 2007, pp. 1–4.

[24] *IQ-Analyzer*. [Online]. Available: http://www.image-engineering.de/, accessed Nov. 4, 2013.

[25] R. K. Mantiuk, A. Tomaszewska, and R. Mantiuk, "Comparison of four subjective methods for image quality assessment," *Comput. Graph. Forum*, vol. 31, no. 8, pp. 2478–2491, 2012.

[26] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document Rec. ITU-R BT.500-13, Geneva, Switzerland, 2012.

[27] *Subjective Video Quality Assessment Methods for Multimedia Applications*, document Rec. ITU-T P.910, 1999.

[28] *Photography—Psychophysical Experimental Methods for Estimating Image Quality—Part 1: Overview of Psychophysical Elements*, document ISO 20462-1, 2005.

[29] P. P. L. Tinio, H. Leder, and M. Strasser, "Image quality and the aesthetic judgment of photographs: Contrast, sharpness, and grain teased apart and put together," *Psychol. Aesthetics, Creativity Arts*, vol. 5, no. 2, pp. 165–176, 2011.

[30] G. Sperling, "The information available in brief visual presentations," *Psychol. Monograph. General Appl.*, vol. 74, no. 11, pp. 1–29, 1960.

[31] A. Hollingworth and J. M. Henderson, "Accurate visual memory for previously attended objects in natural scenes," *J. Experim. Psychol., Human Perception Perform.*, vol. 28, no. 1, pp. 113–136, 2002.

[32] K. Teunissen, "The validity of CCIR quality indicators along a graphical scale," *SMPTE J.*, vol. 105, no. 3, pp. 144–149, 1996.

[33] J. Hair, R. Anderson, R. Tatham, and W. Black, *Multivariate Data Analysis*. Upper Saddle River, NJ, USA: Prentice-Hall, 1998.

[34] J. D. Bullough and Y. Akashi, "Impact of surrounding illumination on visual fatigue and eyestrain while viewing television," *J. Appl. Sci.*, vol. 6, no. 8, pp. 1664–1670, 2006.

[35] X-Rite. *EyeOne Pro Calibrator*. [Online]. Available: http://www.xrite.com/, accessed Nov. 6, 2013.

[36] Precision Vision. *Near Visual Acuity EDTRS*. [Online]. Available: http://precision-vision.com/, accessed Dec. 9, 2014.

[37] Stereo Optical Inc., *Near Contrast Vision FACT*. [Online]. Available: http://www.stereooptical.com/, accessed Nov. 6, 2013.

[38] Luneau Ophtalmologie and VISIONIX. *Farnsworth D-15*. [Online]. Available: http://www.visionix.com/row/, accessed Dec. 9, 2014.

[39] S. Winkler, "On the properties of subjective ratings in video quality experiments," in *Proc. Int. Workshop Quality Multimedia Exper. (QoMEx)*, Jul. 2009, pp. 139–144.

[40] A. M. van Dijk, J.-B. Martens, and A. B. Watson, "Quality asessment of coded images using numerical category scaling," *Proc. SPIE Advanced Image Video Commun. Storage Technol.*, vol. 2451, pp. 90–101, Feb. 1995.

[41] R. Ferzli and L. J. Karam, "A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB)," *IEEE Trans. Image Process.*, vol. 18, no. 4, pp. 717–728, Apr. 2009.

[42] Y. Zhang and D. M. Chandler, "No-reference image quality assessment based on log-derivative statistics of natural scenes," *J. Electron. Imag.*, vol. 22, no. 4, p. 043025, 2013.

[43] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011.

[44] P. V. Vu and D. M. Chandler, "A fast wavelet-based algorithm for global and local image sharpness estimation," *IEEE Signal Process. Lett.*, vol. 19, no. 7, pp. 423–426, Jul. 2012.

[45] C. T. Vu, T. D. Phan, and D. M. Chandler, "S_3: A spectral and spatial measure of local perceived sharpness in natural images," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 934–945, Sep. 2011.

[46] S. A. Golestaneh and D. M. Chandler, "No-reference quality assessment of JPEG images via a quality relevance map," *IEEE Signal Process. Lett.*, vol. 21, no. 2, pp. 155–158, Feb. 2014.

[47] R. Hassen, Z. Wang, and M. M. A. Salama, "Image sharpness assessment based on local phase coherence," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2798–2810, Jul. 2013.

[48] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "Perceptual blur and ringing metrics: Application to JPEG2000," *Signal Process., Image Commun.*, vol. 19, no. 2, pp. 163–172, 2004.

[49] L. Ma, W. Lin, C. Deng, and K. N. Ngan, "Image retargeting quality assessment: A study of subjective scores and objective metrics," *IEEE J. Sel. Topics Signal Process.*, vol. 6, no. 6, pp. 626–639, Oct. 2012.

[50] A. J. Faller, "An average correlation coefficient," *J. Appl. Meteorol.*, vol. 20, pp. 203–205, Feb. 1981.

[51] T. Eerola *et al.*, "Full reference printed image quality: Measurement framework and statistical evaluation," *J. Imag. Sci. Technol.*, vol. 54, no. 1, pp. 10201-1–10201-13, 2010.

[52] M. To, P. G. Lovell, T. Troscianko, and D. J. Tolhurst, "Summation of perceptual cues in natural visual scenes," *Proc. Roy. Soc. B, Biol. Sci.*, vol. 275, no. 1649, pp. 2299–2308, Oct. 2008.

[53] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011.
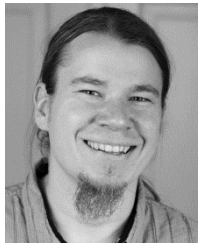
Toni Virtanen was born in Helsinki, Finland, in 1981. He received the M.Psych. degree from the University of Helsinki, Helsinki, in 2010. In 2011, he was admitted to Usability School, a joint program with Aalto University, Espoo, Finland, and the University of Helsinki. He received a funded position from the national User-Centered Information Technology (UCIT) doctoral school in 2012. He is currently pursuing the Ph.D. degree in psychology with the Visual Cognition Research Group, Institute of Behavioural Sciences, University of Helsinki.

Since 2005 and 2009, he has been a Project Researcher and Project Manager in a joint research initiative with Nokia Company, Espoo, Aalto University, and University of Helsinki, where he is involved in subjective image quality of mobile cameras and displays. This collaboration is continuing with Microsoft Company, Espoo. His current research interests include image quality, visual cognition, decision making, augmented sensory perception, visual ergonomics, and human–computer interaction.

Mikko Nuutinen received the M.Sc. (Tech.) and Lic.Sc. (Tech.) degrees from the Helsinki University of Technology, Espoo, Finland, in 2004 and 2007, respectively, and the D.Sc. (Tech.) degree from Aalto University, Espoo, in 2012. He is currently a Post-Doctoral Researcher with the Institute of Behavioral Sciences, University of Helsinki, Helsinki, Finland. His current research interests are in the areas of image and video quality assessment, image categorization algorithms, color image processing, camera performance measurements, and subjective image quality assessment methods and analysis.

Mikko Vaahteranoksa received the M.Sc. (Tech.) degree in electrical engineering from the Helsinki University of Technology, Espoo, Finland, in 2005. He is a Principle Engineer with Microsoft Company, Espoo. He is currently pursuing the master's degree with Aalto University, Espoo. He has worked in image quality metrics and camera image quality tuning at Nokia Company, Espoo, from 2004 to 2014, and has been at Microsoft Company, Espoo, since 2014.

Pirkko Oittinen is currently an Emerita Professor with the Department of Media Technology, Aalto University School of Science, Espoo, Finland. Her Visual Media Research Group has the mission of advancing visual technologies and raising the quality of visual information to create enhanced user experiences in different usage contexts. Her current research topics include still image, video and 3D image quality, content repurposing for mobile platforms, and media experience arising from human–media interaction.

Jukka Häkkinen is currently a Principal Investigator with the Institute of Behavioural Sciences, University of Helsinki, Helsinki, Finland, where he received the M.Psych. and Ph.D. degrees, in 1994 and 2007, respectively. From 1994 to 2000, he was a Research Scientist with the University of Helsinki, where he investigated the human stereoscopic vision. From 2000 to 2007, he was with the Nokia Research Center, Helsinki, where he was also a Senior Research Scientist, Research Manager, and Principal Scientist. In Nokia Research Center, his work focused on visual ergonomics of emerging display technologies, like head-mounted, stereoscopic, and flexible displays. He worked in image and video quality. He is also an Adjunct Professor with the Department of Media Technology, Aalto University School of Science, Espoo, Finland. He leads the Visual Cognition Research Group, Institute of Behavioural Sciences, University of Helsinki. His main research interests include image quality, visual attention, scene perception, material perception, visual ergonomics of head-mounted and flexible displays, and the ergonomics of stereoscopic touch displays.