Nuutinen, Mikko; Virtanen, Toni; Oittinen, Pirkko

Image feature subsets for predicting the quality of consumer camera images and identifying quality dimensions

Published in:
Journal of Electronic Imaging

DOI:
10.1117/1.JEI.23.6.061111

Published: 01/01/2014

Document Version
Publisher's PDF, also known as Version of record

Please cite the original version:
Nuutinen, M., Virtanen, T., & Oittinen, P. (2014). Image feature subsets for predicting the quality of consumer camera images and identifying quality dimensions. *Journal of Electronic Imaging*, *23*(6), 061111-1. https://doi.org/10.1117/1.JEI.23.6.061111

Journal of Electronic Imaging

SPIEDigitalLibrary.org/jei

Image feature subsets for predicting the quality of consumer camera images and identifying quality dimensions

Mikko Nuutinen
Toni Virtanen
Pirkko Oittinen

Journal of Electronic Imaging 23(6), 061111 (Nov/Dec 2014)

Image feature subsets for predicting the quality of consumer camera images and identifying quality dimensions

Mikko Nuutinen,[a,b,*] Toni Virtanen,[b] and Pirkko Oittinen[a]
[a]Aalto University, Department of Media Technology, P.O. Box 15500, Espoo FI-00076, Finland
[b]University of Helsinki, Institute of Behavioural Sciences, P.O. Box 9, Helsinki FI-00014, Finland

Abstract. Image-quality assessment measures are largely based on the assumption that an image is only distorted by one type of distortion at a time. These conventional measures perform poorly if an image includes more than one distortion. In consumer photography, captured images are subject to many sources of distortions and modifications. We searched for feature subsets that predict the quality of photographs captured by different consumer cameras. For this, we used the new CID2013 image database, which includes photographs captured by a large number of consumer cameras. Principal component analysis showed that the features classified consumer camera images in terms of sharpness and noise energy. The sharpness dimension included lightness, detail reproduction, and contrast. The support vector regression model with the found feature subset predicted human observations well compared to state-of-the-art measures. *© 2014 SPIE and IS&T* [DOI: 10.1117/1.JEI.23.6.061111]

Keywords: image quality; no-reference measure; consumer images; image feature; feature subset.

Paper 14159SS received Mar. 28, 2014; revised manuscript received Jul. 12, 2014; accepted for publication Aug. 26, 2014; published online Sep. 15, 2014.

1 Introduction

The quality of photographs depends on many interacting factors and distortion sources. Consumer-level cameras, the technology platform of this paper, are equipped with low-quality optics and image sensors; their shooting process causes motion blur and poor focus, and their low-sensitivity pixels increase noise level. In addition, the conditions that pictures are taken in, such as shooting distance and lighting conditions, affect the quality of raw images. After shooting, the raw images are processed by the image signal processing (ISP) pipe of the camera. The ISP includes operations such as color filter array demosaicking, automatic white balancing, color correction, noise filtering, tone reproduction, gamma correction, edge enhancement, color saturation enhancement, and image compression.[1,2] The parameters and the order of operations affect the resulting image. Some operations, such as demosaicking, white balancing, and noise filtering, seek to restore the image. Other components, such as edge enhancement and color saturation enhancement, aim to produce a pleasant image.

The quality of the photographs can be evaluated using both subjective and objective methods. In this study, the term subjective method denotes a test performed on test participants. The term objective method refers to an algorithm based on the computational process applied to the test images. The output of an algorithm is a value related to the image quality. In a subjective test, an observer rates test images based on the overall quality or quality attributes.

In the current research of objective image-quality assessment, image-quality measures are usually classified into full-reference (FR), reduced-reference (RR), and no-reference (NR) approaches depending on whether and how the reference image is used. The FR measures[3–5] cannot be applied in the case of consumer camera images because of the lack of pixel-wise reference images; that is, the original image or a reference image is not available. An RR measure[6] requires some information from the original or reference image. RR measures have been applied to consumer camera images, but the requirement for a calibrated reference camera makes them cumbersome.[7–9]

NR measures[10–17] do not need a reference or original image and are applicable to consumer camera images. However, NR measures tend to perform poorly in the case of images with multiple distortion sources. In fact, the fundamental problem from the perspective of this paper is that NR measures have been developed to characterize images with only one distortion type at a time as is commonly found in databases[18–22] or images with two distortions.[23] The types of distortions include JPEG or JPEG2000 compression, noise contamination, low-pass filtering, or fast-fading distortion.

Our experience suggests that a measure developed with training data of only one distortion type may respond to other distortions in an undesirable manner.[7,24] For example, a dedicated sharpness measure may interpret noise energy as image details. This behavior is not problematic if low-pass filtering is the only distortion source, but such a measure fails if the distortion space of the image is multidimensional.

A close research field with NR image quality is aesthetic evaluation. Aesthetic measures incorporate low-level image properties (brightness, contrast, color, edges, hue, etc.), image composition rules (rule of thirds, golden ratio, depth of field, and color harmonies), and content properties

*Address all correspondence to: Mikko Nuutinen, E-mail: mikko.nuutinen@helsinki.fi

(face detection and scene types).[25,26,27] Roughly speaking, aesthetic research aims at differentiating professional from nonprofessional photographs in a random stream and ranking images with respect to professional skill or viewer impression, whereas image-quality research strives toward sorting images in terms of technical performance or image processing. In this study, we focus on the latter approach. Image composition and content properties are out of the scope of the study.

Regardless of the measure (aesthetic and image quality), algorithms combine one or more image features to form feature vectors. The term feature refers to a piece of information computed from an image. The feature vectors are inputted into a regression model or fed into a learning model to predict human-provided ground truth of aesthetics or quality. The performance of a measure becomes as high as the features and training data allow. The hypothesis of this study is that a high-performance, reference-free quality measure for images suffering from multiple distortions uses a combination of interacting features as inputs.

The aim of this study is to find an efficient feature subset from a large feature set that could classify distortions and predict the visual quality of photographs captured by different consumer cameras. In the study, we applied several feature selection methods. Feature selection is a mature research topic and has been used for many applications.[28] However, to the best of our knowledge, this is the first study that attempts to find quality dimensions via feature selection for images captured by different types of cameras. We used images from the CID2013 image database,[29] which contains images captured by different consumer-level cameras under different shooting conditions as well as subjective evaluations of those images. The CID2013 image database provides distortion types for more realistic application scenarios than images used in the earlier studies. The results of this study are applicable to consumer photographs, such as quality ranking of images in photo-sharing sites or in image-retrieving systems. Conventional methods are applicable only to cases with a restricted set of distortions, such as image storage (JPEG compression distortion), image scaling (blurring distortion), or image transmission (white noise distortion).

The novelty of this study relates to finding the combinations of features that account for the overall quality of consumer-level photographs. Because the feature selection process provides a systematic way of restricting the number of features, we were able to find and to analyze the quality dimensions of consumer camera images. The dimensions were identified by principal component analysis and included a sharpness dimension and a noise energy dimension. The sharpness dimension consisted of lightness, detail reproduction, and contrast. The result strongly posits that the multidimensionality of the image quality should be taken into account in developing new measures. According to our knowledge, the quality dimensions have earlier been explored by subjective research[30–32] only. The results of Ref. 30 showed that the most important subjective image quality dimensions are contrast, naturalness, darkness, and sharpness. The authors concluded that the high-level attribute naturalness is a requirement for high-quality images, whereas quality can fail for other reasons in low-quality images. For example, a low-quality image can be dark and unsharp, as our results also indicated.

This paper is divided into four sections. Following this introductory section, Sec. 2 reviews the feature selection methods, reference measures, and image material used in this study. Section 3 presents the results of the study. We compare the performance of the found-feature subsets and state-of-the-art measures and explore the quality dimensions of sharpness and noise energy for consumer photographs. Section 4 concludes the study.

2 Methods

2.1 *Feature Set*

This study is based on the feature set S, which includes 270 features. The selection of these features was guided by their coverage of well-known image attributes. The measures derived from the features predict overall image quality, aesthetics, and attributes such as sharpness or blurriness, noise, colorfulness, and JPEG or JPEG2000 distortions. Moreover, the features f3 and f4 of the measure[33] have been developed to express the lightness of image, and features f113 and f117 of Ref. 34 to express the hue of image. Feature f221 of Ref. 35 is developed to measure contrast, just to name a few. Table 1 lists the features included in the feature set S, as well as the origins and the types of measures in which they were originally used.

We implemented features f1 to f131, f154 to f171, and f173 to f270 in this study by loosely following their respective references. The code for computing features f132 to f153 and f172 originated from their authors. In addition to the references listed in Table 1, we used some other toolboxes and functions. Implementations of features f8 to f95 use the PyrTool toolbox,[47] which computes steerable pyramid decomposition. Features f39 to f50 use the ssim_index code[48,49] for computing structural difference values between different scales and orientations. Features f96 and f97 apply the phasecong3 code[50] when phase congruency images are computed.

2.2 *Feature Selection Process*

We searched for the most efficient feature subsets from feature set S using different feature selection methods. Figure 1 shows the components of the selection process. Subset generation and subset evaluation are commonly used components of feature subset selection.[28,51,52] For this study, we also implemented a method-comparison component, which compares the best subsets obtained using different methods.

2.2.1 *Subset generation*

The most comprehensive strategy of candidate subset generation is a complete search. However, this strategy requires an exponentially large search space $O(N^2)$, where N is the number of features. Heuristic and random searches require less computational power. The complexity of the heuristic search is $O(N^2)$ or less, which can also be computationally demanding for a large feature space.

A heuristic search is often based on a hill-climbing approach, such as sequential forward selection (SFS) or sequential backward elimination (SBE).[28] SFS starts with an empty set and adds one feature at a time from the original set by maximizing the performance measure. SBE starts with the original group and eliminates one feature at a time by

Table 1 Original feature set S included 270 features, which were inspired and implemented from different references.

Features	Original measure type[a]	Reference
f1 to f7	Image quality	33
f8 to f95	Image quality	15
f96 to f99	Image quality	36
f100 to f109	Image quality	37
f110 to f131	Aesthetic	34
f132 to f150	Image quality	14
f151	Sharpness	11
f152	Sharpness	38
f153	JPEG distortion	39
f154	Sharpness	10
f158 to f162	Sharpness	40
f163 to f169	Sharpness	41
f170	Noise	42
f171	JPEG2000 distortion	12
f172	Sharpness	42
f173	Sharpness	43
f174 to f217	Sharpness / noise	44
f218 to f221	Sharpness	35
f222 to f257	Image quality	17
f258 to f261	Image quality	45
f261 to f270	Colorfulness	46

[a]In the original literature references, the features have been combined to form a measure or algorithm for predicting some image property, such as image quality, aesthetics, sharpness, noise, colors, or image compression distortion.

maximizing the performance measure. The performance measure is computed in the subset evaluation component.

In this study, we used the SFS method. We chose this method because the complete search was computationally too heavy due to the large size of the original feature set S. Our selection of SFS instead of the SBE method is borne out by the ratio of M and N, where M is the number of assumed relevant features and N is the total number of features. According to Liu and Yu,[28] if M is small, then the SFS strategy should be used, and if the number of irrelevant features (N minus M) is small, then the SBE strategy should be used. We assume that the size of M of the feature set S is small. The original feature set S includes many redundant features; thus, we prefer the SFS strategy.

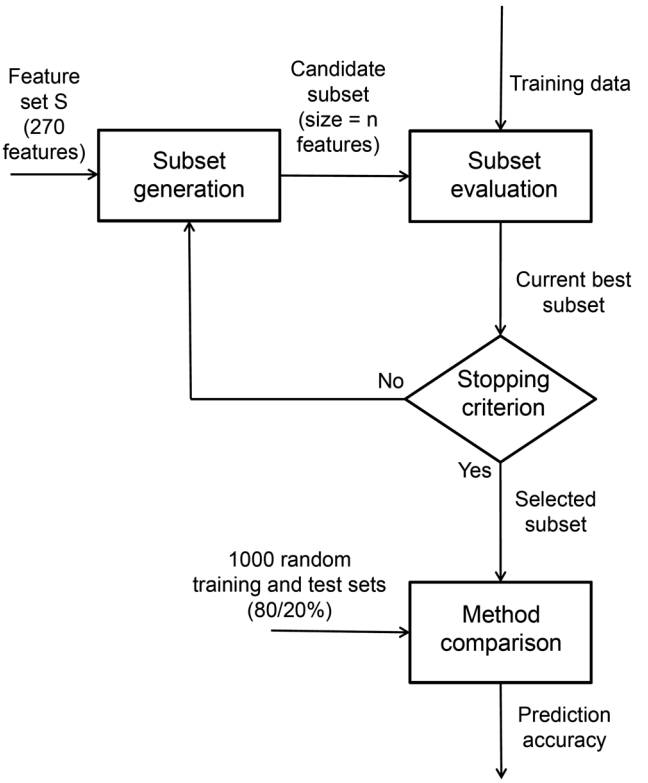

Fig. 1 The framework of feature selection used in this study.

2.2.2 *Subset evaluation: filter methods*

The subset generation component feeds candidate subsets to the subset evaluation. The component selects the most efficient feature subset from the group of candidate subsets. The evaluation methods can be divided into three types: filter, wrapper, and hybrid.[28] For this study, we implemented strategies based on all three of these approaches.

Filter methods can be further divided into two types: direct and subset ranking. Direct filter methods measure how well any given feature in a subset classifies or ranks the subjective evaluation data. In practice, direct filter methods remove irrelevant features, while subset ranking methods filter both irrelevant and redundant features.

In this study, we used the correlation-based feature selection (CFS)[51] method as our subset ranking filter method and the Pearson linear correlation (LCC), symmetrical uncertainty (SU), and relief[51] as our direct filter methods. Table 2 lists the filter methods used in this study with their associated abbreviations.

The CFS filter method evaluates candidate subsets by Eq. (1):

$$\mathrm{CFS}_{\mathrm{value}} = \frac{k r_{\mathrm{cf}}}{\sqrt{k + k(k-1) r_{\mathrm{ff}}}}, \tag{1}$$

where k is the size of the candidate feature set, r_{cf} is the mean correlation between perceived quality and individual feature values, and r_{ff} is the average feature-feature intercorrelation. The numerator of Eq. (1) expresses the performance of a subset in its predictions of perceived quality. The denominator indicates the redundancy among the features arising from their mutual correlation.

Table 2 Filter methods for feature subset evaluation used in this study.

Subset evaluation	Method type	Abbreviation
CFS with LCC	Subset ranking	CFS(LCC)
CFS with SU	Subset ranking	CFS(SU)
CFS with relief	Subset ranking	CFS(relief)
LCC	Direct filter	direct_LCC
SU	Direct filter	direct_SU
Relief	Direct filter	direct_relief

Note: CFS, correlation-based feature selection; LCC, Pearson linear correlation; SU, symmetrical uncertainty.

We applied three versions of CFS: CFS(LCC), CFS(SU), and CFS(relief). The correlation values of r_{cf} and r_{ff} in Eq. (1) were determined by LCC, SU, or relief measures. LCC was calculated by Eq. (2):

$$\mathrm{LCC} = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n(\sum x_i^2) - (\sum x_i)^2}\sqrt{n(\sum y_i^2) - (\sum y_i)^2}}, \quad (2)$$

where n is the number of compared pairs and x_i and y_i are the samples of X and Y, respectively. SU was computed by Eq. (3):

$$\mathrm{SU} = 2 \cdot \left[\frac{\mathrm{IG}(Y|X)}{H(Y) + H(X)}\right], \quad (3)$$

where

$$\mathrm{IG}(Y|X) = H(Y) - H(Y|X), \quad (4)$$

$$H(Y|X) = -\sum_{x\in X} p(x) \sum_{y\in Y} p(y|x)\log_2[p(y|x)], \quad (5)$$

$$H(Y) = -\sum_{y\in Y} p(y)\log_2[p(y)], \quad (6)$$

where $H(Y)$ is the entropy of Y and $H(Y|X)$ is the entropy of Y after observing X. The amount by which the entropy of Y decreases reflects additional information about Y provided by X; this amount is known as the information gain (IG). According to this measure, feature Y is regarded as more correlated to feature X than to feature Z if $\mathrm{IG}(X|Y) > \mathrm{IG}(Z|Y)$. SU values range from 0 to 1; a value of 1 indicates that knowing the values of either feature completely predicts the values of the other; a value of 0 indicates that X and Y are independent.

The relief metric[51] estimates the weight of features according to how well their values distinguish samples near to each other. The values of this metric range from −1 to 1. Relief is calculated by Eq. (7):

$$\mathrm{Relief}_X = \frac{\mathrm{Gini}' \cdot \sum_{x\in X} p(x)^2}{[1 - \sum_{y\in Y} p(y)^2]\sum_{y\in Y} p(y)^2}, \quad (7)$$

$$\mathrm{Gini}' = \left\{\sum_{y\in Y} p(y)[1 - p(y)]\right\} - \sum_{x\in X}\left\{\frac{p(x)^2}{\sum_{x\in X} p(x)^2}\sum_{y\in Y} p(y|x)[1 - p(y|x)]\right\}. \quad (8)$$

In addition to CFS, we used three direct filter methods: direct_LCC, direct_SU, and direct_relief with the correlate measures of LCC, SU, and relief described in Eqs. (2), (3), and (7).

2.2.3 *Subset evaluation: wrapper and hybrid methods*

Wrapper methods seek efficient feature subsets for a measure by using the same learning model that the measure uses. Hybrid methods combine the steps of the filter and wrapper methods. First, a hybrid method selects k features from the original feature set using a filter method and then applies a wrapper method to select a subset with l features ($l \leq k$).

In this study, when applying wrappers, SFS fed candidate feature subsets (subset generation component) to a learning model (subset evaluation component). The learning models used were the support vector regression (wrapper_SVR) and linear regression (wrapper_REG) models. We utilized the libSVM package[53] in order to implement the SVR. The kernel used for regression was the radial basis function kernel. The weighting factors of the linear regressions were trained by the regress function in MATLAB®.

The subset evaluation component used a fivefold cross-training method for parameter training and validation. We used 80% of the data for model training and 20% for performance testing. This random 80/20 division was performed 1000 times for each of the candidate subsets. The performance measure was the average LCC between the output of the trained model and perceived data.

The hybrid methods used in this study (hybrid_LCC, hybrid_SU, hybrid_relief, and hybrid_IG) were based on the measures of Eqs. (2), (3), (4), and (7), and SVR. The hybrid methods used filter methods to filter 20 features from the feature set S and used SVR (as a wrapper) to select subsets of 1 to 20 features. Table 3 lists the wrapper and hybrid methods used in this study for feature subset searching.

2.2.4 *Method comparison*

The method comparison component compares the performance of the feature subsets found by the various feature subset selection methods listed in Tables 2 and 3.

The performance metric of this component is the mean prediction accuracy (linear correlation between predicted and subjective values) as a function of subset size. The subset sizes are limited to 1 to 20 features, as larger subset sizes do not increase the prediction performance and smaller subset sizes reduce the risk of overfitting. The performance values of the methods are computed using 1000 randomly selected

Table 3 Wrapper and hybrid methods for feature subset evaluation used in this study.

Subset evaluation	Method type	Abbreviation
SVR	Wrapper	wrapper_SVR
REG	Wrapper	wrapper_REG
LCC and SVR	Hybrid	hybrid_LCC
SU and SVR	Hybrid	hybrid_SU
Relief and SVR	Hybrid	hybrid_relief
IG and SVR	Hybrid	hybrid_IG

Note: SVR, support vector regression; REG, linear regression; IG, information gain.

training and testing image data (80/20%) as used with the wrappers in the subset evaluation component.

2.3 *Subjective Data for Performance Study*

In this study, we used the CID2013 image database,[29] which is publicly available for research purposes and is freely downloadable. CID2013 includes six image sets (I to VI), each captured by 12 to 14 different consumer cameras at a given time of the year. The quality levels of the cameras range from low to high; the cameras comprise low-, moderate-, and high-quality mobile-phone cameras, moderate-quality compact cameras, and low- to moderate-quality SLR cameras.

The images contained in the sets were captured at different times of the year from the same eight scenes. The scenes represent environments in which consumers typically shoot photos, ranking from dark to bright indoors to bright outdoor conditions. The types of scenes chosen for the database were partly based on the photospace approach described in Ref. 54. Figure 2 shows sample images of the different scenes. In total, CID2013 includes 474 images, called test images, captured by 79 different cameras, i.e., on average, six images per camera.

The subjective experiments were performed in a dark room with controlled lighting directed toward a wall behind the displays to avoid flare. The lighting produced ambient illumination of 20 lux. The setup included two colorimetrically calibrated 24 in. 1920×1200 displays (Eizo Color Edge CG210) for, respectively, displaying a test image at a time and its reference images, with a third smaller display underneath for presenting questions. The subject's 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, the images were scaled to a size of 1600×1200 pixels using the bicubic interpolation method.

The number of observers was 30, 32, 31, 26, 34, and 34 for image sets I, II, III, IV, V, and VI, respectively. All observers were náive in terms of evaluating image quality and had normal or corrected-to-normal vision. Of the subjects, 67% were female. The observers' vision were controlled for near vision acuity EDTRS (Precision Vision, La Salle, Illinois), near contrast vision F.A.C.T. (Stereo Optical Co. Inc., Chicago, Illinois), and color vision Farnsworth D-15 (Luneau Ophtalmologie, Chartres, France) before participation. They received two movie tickets as a reward. On average, the experiment lasted 93 min. 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.

Test images were presented in random order, one scene at a time for each observer using the dynamic reference absolute category rating (DR-ACR) method. The DR-ACR method creates reference image series from the test images. Before evaluating a test image of a given scene on one display, all of the test images of the scenes in question are shown to the observer as a slide show for reference on the other display. This process is repeated before the observer evaluates each test image. In other respects, the DR-ACR method very much resembles a basic ACR method,[55] except that the observers see a slideshow of all the other images in the test depicting the same scene before every evaluation. By seeing the other images in the test setup as reference, the observers are more aware of the total variation of quality represented within a single image set. This improves their evaluation as they do not need to avoid using the far ends of the scale in case there would be better or worse images later on in the experiment.

The observers first rated the overall quality of each image in a test series and then evaluated its quality attributes of sharpness, graininess, brightness, and color saturation. The sharpness scale ranged from very blurry to very sharp (0 to 100), the graininess scale ranged from very grainy to not grainy at all (0 to 100), the brightness scale ranged from too dim to too bright (−100 to 100), and the color

Fig. 2 The images used for feature searching and validation were captured from different scenes by 12 to 14 different cameras.

saturation scale ranged from too pale to too saturated (−100 to 100).

For this study, we divided the image sets of CID2013 into two parts: image sets I to III were used for exploring efficient feature subsets (Sec. 3.1) and image sets IV to VI for comparing the performance of the best feature subsets and the state-of-the-art measures (Sec. 3.2). In total, we used 240 images captured by 40 cameras for feature subset selection and 234 images captured by other 39 different cameras for performance analysis. We grouped the images in the two parts because the subjective evaluation method differed between image sets I to III and IV to VI. When image sets I to III were evaluated, the observers fixed the best and the worst image to the ends of the scale, which was not required with image sets IV to VI. More details and the data analysis of CID2013 can be found in Ref. 29.

In addition to image sets IV to VI of the CID2013 image database, we explored the performance of the best feature subset found in this study with the images of the LIVE multiply distorted image quality database (MDIQD)[23] (Sec. 3.4).

2.4 *Reference Measures*

In Sec. 3.2, we compare the performance of the two most efficient feature subsets found in this study to the performance of state-of-the-art NR measures of image quality. These measures include NIQE,[56] BRISQUE,[17] BLIINDS-II,[16] JNBM,[38] CPBD,[11] Wang et al.,[39] Sheikh et al.,[13] Marziliano et al.,[10] BIQI,[14] and BIQAA.[44] We required that implementations of these algorithms are publicly available on the Internet. These algorithms follow different approaches: JNBM, CPBD, Wang et al., Sheikh et al., and Marziliano et al. are distortion-specific measures. BIQAA, BIQI, NIQE, BRISQUE, and BLIINDS-II are distortion-agnostic measures, i.e., measures designed to measure the quality of an image without knowledge of the distortion type. However, these measures had been trained using databases of test images with limited sets of distortions.

The Marziliano et al., JNBM, and CBBD measures predict image sharpness. The measure proposed by Marziliano et al. finds edges and calculates edge widths as pixels. JNBM divides images into blocks. If the edge widths, calculated by the Marziliano et al., inside blocks are higher than the just-noticeable-blur (JNB) threshold, the probability that the image is not sharp increases. The CPBD measure is based on the JNB concept, but it computes the percentage of edges at which blur cannot be detected.

The measure proposed by Wang et al. computes the quality of JPEG images by multiplying measures of blockiness and activity. The measure proposed by Sheikh et al. is based on joint distributions of the sub-bands of wavelet coefficients to measure how JPEG2000 compression changes those distributions.

BIQAA uses the Renyi entropy measure along various orientations to determine anisotropy. The assumption in this method is that blur or noise introduces a substantial change in the directional information of the scene. Hence, anisotropy decreases as additional blurring or noise is added to the image.

The BIQI, BLIINDS-II, BRISQUE, and NIQE measures use learning models, as our approach in this paper does, with feature values computed from distorted images. The goodness of this kind of measures depends on the features and images used for training. BIQI is based on a framework with two phases; it fits the wavelet coefficients of the distorted image to the generalized Gaussian distribution (GGD) model and derives the features from the parameters of GGD. In the first phase, the probabilities for the different predetermined distortions are calculated using a support vector machine. These probabilities are used as weighting factors for the second phase, in which the values for the specific distortion set are calculated using SVR.

BLIINDS-II fits discrete cosine transform coefficients of the distorted image to the GGD model and derives features from the parameters of the GGD model. BRISQUE fits locally normalized luminance values to the GGD and asymmetric generalized Gaussian distribution (AGGD) models; feature values are derived from the parameters of the models. NIQE also uses the features derived from the GGD and AGGD models. First, these features are computed for distortion-free natural image patches (model training) and fit to obtain the multivariate Gaussian (MVG) model density. NIQE computes the quality of the distorted image as the distance between the parameters of the MVG model for the natural image and those for the distorted image.

3 Results

3.1 *Performance of Feature Subsets Found by Different Selection Methods*

Figures 3 and 4 show the performance of the feature subsets found by different selection methods as a function of the subset size: Fig. 3 shows the high-performance subsets and Fig. 4 shows the low-performance subsets. The performance values are determined using the LCC values between the subjective overall quality and values predicted by the feature subsets. The best-performing subsets have LCC values of 0.6 to 0.8. The performance saturated at subset sizes of 10 to 15 features. According to Fig. 3, the wrapper_SVR, wrapper_REG, and CFS(LCC) methods found the most efficient feature subsets for the data used in this study.

Table 4 shows the LCC values for the feature subsets found by various methods. The selected feature subset sizes in the table maximized the LCC value of the individual method. The best method was wrapper_SVR, the second

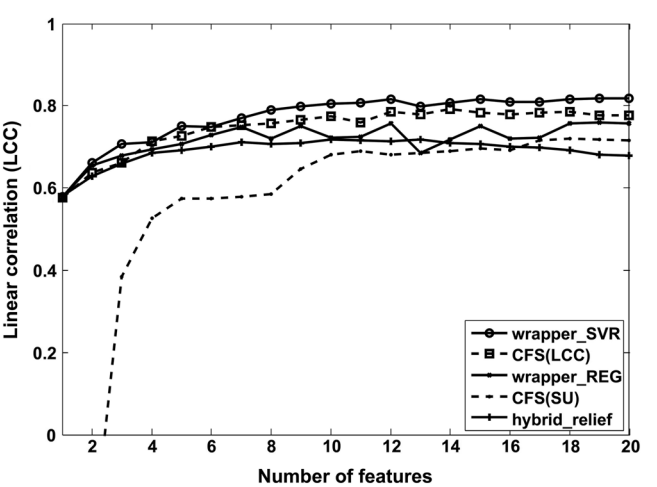

Fig. 3 Performance of feature subset selection methods as a function of the number of features (high-performance methods).

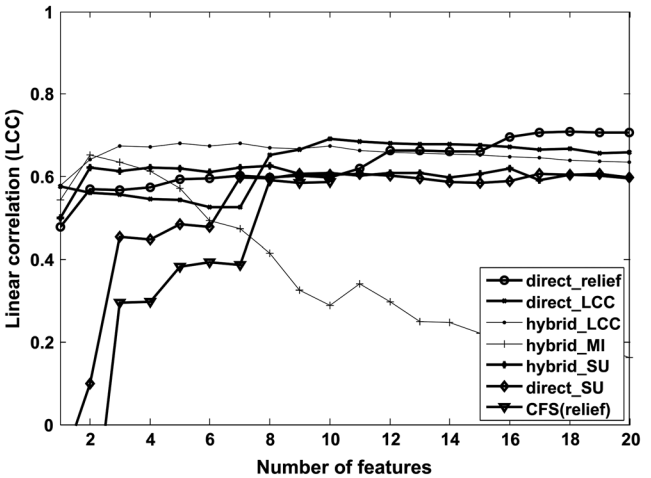

Fig. 4 Performance of feature subset selection methods as a function of the number of features (low-performance methods).

best was CFS(LCC), and the third best was wrapper_REG. The wrapper_SVR method used 20 features. Note that the size of the best feature subset found by the CFS(LCC) method was small (14 features) compared to the wrapper methods.

According to the student's t tests,[57] the average LCC values of 1000 (80/20% train and test) samples for the wrapper_SVR, wrapper_REG, and CFS(LCC) methods differed statistically from each other: wrapper_SVR versus CFS (LCC) ($\mathrm{df} = 1998$, $p < 0.0001$), wrapper_SVR versus wrapper_REG ($\mathrm{df} = 1998$, $p < 0.0001$), and CFS(LCC) versus wrapper_REG ($\mathrm{df} = 1998$), $p < 0.0001$).

The wrapper_SVR method found the feature subset with the highest prediction accuracy for image sets I to III. The size of the subset, however, should be taken into account. As the subset size increases, so does the risk of data overfitting. To compare the measures in this paper with the state-of-the-art measures presented in Sec. 3.2, we selected two feature subsets: wrapper_SVR and CFS(LCC). The feature subset found by the wrapper_SVR was selected because its LCC value was the highest, while the subset found by the CFS (LCC) was selected because its LCC value was the second highest and we wanted to favor small subset sizes.

Table 4 The performance of feature subsets found by different selection strategies (LCC values between the best feature subsets and subjective overall quality).

Strategy	LCC	Number of features
wrapper_SVR	0.821	20
CFS(LCC)	0.793	14
wrapper_REG	0.759	19
CFS(SU)	0.719	17
hybrid_relief	0.718	10
direct_relief	0.709	17
direct_LCC	0.691	10
hybrid_LCC	0.681	5
hybrid_MI	0.652	2
hybrid_SU	0.626	8
direct_SU	0.607	9
CFS(relief)	0.591	8

3.2 *Performance of the Efficient Feature Subsets and Reference Measures*

In this section, we compare the performance of the most efficient feature subsets and the reference measures using image sets IV to VI from the CID2013 image database.

The feature subsets were fed to the SVR model, which was trained using a content-based sixfold cross-training method. First, we divided the images into six groups. The images in group 1 are close-up photos in dark lighting conditions with illuminance values of 2 lux. The images in groups 2 and 3 are close-up photos in typical dim indoor lighting conditions with illuminance levels of 100 and 10 lux, respectively. The images in group 4 had illuminance levels equivalent to high indoor lighting levels. The images in groups 5 and 6 are images in typical cloudy to sunny outdoors lighting conditions taken of small groups of people (group 5) or landscapes (group 6). The parameters of SVR were estimated using data from five groups; these data were used as the training data. The sixth group was used as the testing data. Testing was performed six times, and all of the groups acted as the testing data once through these tests.

Figure 5 shows example images from the different groups. These images were grouped according to illuminance levels. The illuminance level greatly impacts the distortion types present in images captured by consumer-level cameras. The division of six groups and the usage of these groups for the cross-training of the methods were justified by the different distortion types in the different groups.

Note that before the SVR parameters were estimated, the feature values were normalized in image-set and group-specific ways, and the values of the reference measures were also normalized. We normalized these feature values and reference measures because the subjective values of the image groups and sets in the CID2013 database were evaluated independently of each other. The observers evaluated one image group from one image set at a time. Thus, we expect that the same quality ratings for images from different groups are not equivalent to each other because the quality scales of the various groups differ. For example, a quality value of 10 for group 3 in image set IV is not the same as a quality value of 10 for any of the other groups.

Before evaluating the performance of an algorithm, it is common to apply a logistic transform to the predicted objective scores to bring the predicted and measured (subjective) scores on the same scale and to account for the typically nonlinear relationship between the two scores. The trained SVR model performs this transform for the feature subsets. For the reference measures, we used a logistic function[18] with an added linear term

$$f(x) = \beta_1 \left\{ \frac{1}{2} - \frac{1}{1 + \exp[\beta_2 (x - \beta_3)]} \right\} + \beta_4 \cdot x + \beta_5, \qquad (9)$$

Fig. 5 Example images of six different groups used for feature subset performance comparison study.

where β_1, β_2, β_3, β_4, and β_5 are the model parameters chosen to minimize the root-mean-square error (RMSE) between the reference measures and the subjective quality values.

Tables 5 to 7 show the LCC, Spearman rank-ordered correlation, and RMSE values for the feature subsets, respectively, found by the wrapper_SVR and CFS(LCC) methods and the reference measures. The methods are sorted by their overall performance values. These results suggest that the performance of the CFS(LCC) and wrapper_SVR methods are higher than those of the reference measures. The overall LCC of the CFS(LCC) method is 0.76, while the best reference measure is BRISQUE, with an LCC value of 0.62. The group-specific performance of the wrapper_SVR method was the highest except for groups 1 and 5. In these cases, the performance of either the CFS(LCC) or Marziliano et al. measure was best. Figure 6 compares the subjective and predicted quality for the CFS(LCC) and wrapper_SVR measures.

To determine which differences between the feature subsets found by the wrapper_SVR and CFS(LCC) methods and the reference measures were statistically significant, we used a variance test. This test is the same as that used in Ref. 18. The assumption of this test is that the residuals (the differences between the subjective scores and the predicted scores) are normally distributed. We tested the normality using a kurtosis-based criterion, according to which the residuals are Gaussian if the kurtosis is between 2 and 4.[18] The assumption of the Gaussian distribution was met for all the methods.

We used an F test to test whether the variances of the residuals were identical, i.e., whether two sample sets came from the same distribution. The null hypothesis is that the residuals of both measures are expressions from the same distribution and are statistically indistinguishable with 95% confidence. According to this variance test, there is a significant difference between the methods based on the feature subsets found in this study and all tested reference measures. The difference between wrapper_SVR and CFS(LCC) is not significant with 95% confidence.

Next, we try to determine how the accuracy of the best-performing measure compares to the accuracy of random human observers—in other words, whether time-consuming

Table 5 LCC values between the subjective overall quality evaluations and algorithmic predictions (feature subsets and reference measures). The best performers are bolded.

Measure	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Overall
CFS(LCC)	**0.796**	0.817	0.862	0.762	0.658	0.796	**0.757**
wrapper_SVR	0.718	**0.859**	**0.874**	**0.807**	0.553	**0.844**	0.733
BRISQUE	0.618	0.644	0.643	0.383	0.760	0.692	0.615
BIQI	0.358	0.753	0.619	0.751	0.559	0.695	0.540
JNBM	−0.300	0.543	0.695	0.600	0.544	0.530	0.430
BLIINDS-II	0.505	0.552	0.485	0.454	0.400	0.297	0.381
Martziliano et al.	−0.342	0.487	0.356	0.586	**0.810**	0.509	0.368
NIQE	0.485	0.287	0.214	0.386	0.489	0.510	0.341
BIQAA	−0.130	0.343	−0.025	0.601	0.581	0.412	0.300
CPBD	−0.413	0.395	0.409	0.486	0.692	0.397	0.260
Sheikh et al.	0.286	0.309	0.026	0.163	−0.206	0.421	0.175
Wang et al.	0.104	−0.329	0.401	0.055	−0.277	0.339	0.048

Table 6 Spearman rank-ordered correlation (SROCC) values between the subjective overall quality evaluations and algorithmic predictions (feature subsets and reference measures). The best performers are bolded.

Measure	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Overall
CFS(LCC)	**0.792**	**0.828**	**0.830**	0.651	0.666	0.769	**0.753**
wrapper_SVR	0.719	0.820	0.812	**0.706**	0.595	**0.803**	0.746
BRISQUE	0.545	0.675	0.672	0.224	0.690	0.712	0.583
BIQI	0.191	0.774	0.656	0.678	0.485	0.585	0.524
JNBM	−0.177	0.527	0.677	0.476	0.400	0.377	0.418
BLIINDS-II	0.581	0.483	0.473	0.404	0.461	0.273	0.373
NIQE	0.480	0.294	0.238	0.458	0.475	0.536	0.358
BIQAA	−0.044	0.432	−0.022	0.637	0.441	0.404	0.316
Martziliano et al.	−0.311	0.454	0.340	0.380	**0.719**	0.358	0.308
CPBD	−0.347	0.389	0.354	0.389	0.613	0.175	0.207
Sheikh et al.	0.314	0.315	−0.084	−0.077	−0.129	0.314	0.178
Wang et al.	0.113	−0.2821	0.278	0.081	−0.172	0.355	0.065

subjective studies can be replaced by the measures found in this study.

We compared the accuracy of our measure and n random human observers by computing RMSE values. Figure 7 shows the RMSE values for the subjective data as a function of the number of observers for image sets IV to VI. These RMSE values were calculated by comparing the mean value of n observers with the mean values of all of the observers. For example, if $n = 3$, the mean value of three selected observers was compared with the mean of all observers.

Table 7 Root-mean-square error values between the subjective overall quality evaluations and algorithmic predictions (feature subsets and reference measures). The best performers are bolded.

Measure	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Overall
CFS(LCC)	**17.73**	11.05	**15.00**	14.11	**18.47**	13.23	**14.93**
wrapper_SVR	20.45	**9.35**	15.60	**14.07**	19.87	**12.20**	15.26
BRISQUE	21.58	14.31	21.61	19.79	15.78	15.46	18.09
BIQI	26.48	12.78	22.79	14.98	19.30	18.31	19.11
JNBM	31.11	15.31	21.84	16.53	19.64	18.28	20.45
Martziliano et al.	31.25	16.21	25.62	17.76	16.93	18.74	21.08
BIQAA	28.94	17.23	27.19	18.52	19.88	19.94	21.95
CPBD	30.91	17.05	26.33	18.91	19.59	20.19	22.16
NIQE	24.06	19.84	30.67	19.97	21.19	20.21	22.66
Sheikh et al.	26.95	17.59	27.08	19.75	24.53	20.82	22.79
Wang et al.	27.63	18.83	27.29	20.23	23.39	21.79	23.19
BLIINDS-II	25.35	22.18	21.03	32.74	30.91	28.71	26.82

Table 8 The descriptions of features found by CFS(LCC) strategy.

Feature	Description	Reference
F1 (f8)	Variance parameter of the GGD model for wavelet coefficients after DNT (steerable pyramids decomposition, first scale, orientation 0 deg)	15
F2 (f135)	Variance parameter of the GGD model for wavelet coefficients (Daubechies 9/7 decomposition, first scale, vertical orientation)	14
F3 (f136)	Variance parameter of the GGD model for wavelet coefficients (Daubechies 9/7 decomposition, second scale, vertical orientation)	14
F4 (f31)	Shape parameter of the GGD model for wavelet coefficients after DNT (steerable pyramids decomposition, second scale, orientation 150 deg)	15
F5 (f148)	Shape parameter of the GGD model for wavelet coefficients (Daubechies 9/7 decomposition, second scale, diagonal orientation)	14
F6 (f223)	Variance parameter of the GGD model for locally normalized luminance values	17
F7 (f57)	Parameter of the spatial correlation value function computed between central pixel and pixels from chess board distance (parameter a_1, steerable pyramids decomposition, first scale, orientation 30 deg)	15
F8 (f58)	Parameter of spatial correlation value function computed between central pixel and pixels from chess board distance (parameter a_2, steerable pyramids decomposition, first scale, orientation 30 deg)	15
F9 (f54)	Parameter of spatial correlation value function computed between central pixel and pixels from chess board distance (parameter a_3, steerable pyramids decomposition, first scale, orientation 0 deg)	15
F10 (f130)	Magnitude of wavelet coefficients in saturation channel between image center and foreground (Daubechies 9/7 decomposition, third scale, average of the diagonal, vertical and horizontal orientations)	34
F11 (f116)	Average intensity value of image center area	34
F12 (f260)	The width of the middle 98% mass of the gray level histogram	45
F13 (f158)	Comparison between kurtosis values computed from edge areas of image and its low-pass filtered version (Gaussian low-pass filter: size 100 pixels; standard deviation 1 pixel)	40
F14 (f168)	Shape of histogram of gradient profiles from edge areas	41

Note: GGD, generalized Gaussian distribution; DNT, divisive normalization transform.

We randomly selected different observer combinations from the group containing all observers, and the subjective RMSE was the average value computed from all combinations.

The RMSE values of the feature subset found by CFS (LCC) method for image sets IV, V, and VI were 16.48, 14.99, and 13.33, respectively, as shown in Fig. 7. It should be noticed that these values are image set specific. The RMSE values in Table 7 are averages over image sets IV, V, and VI. Figure 7 shows that the accuracy of the CFS (LCC) method equals the accuracy of a single random observer ($n = 1$). Figure 7 shows that standard deviations between subjects stabilize the actual value with 10 to 15 observers. This means that the goal in developing objective measure would be to approximately reach an RMSE value of 4.

3.3 *Characteristics of the Feature Subset Found by CFS(LCC) Method*

In this section, we explore the efficient feature set more closely. According to our knowledge, this is the first study that attempts to link quality dimensions of photographs and features selected using feature selection methods. Table 8 shows descriptions of features F1 to F14 of the subset selected by the CFS(LCC) method and how they correspond to the feature codes of Table 1. Table 9 shows descriptions of the features of the subset selected by the wrapper_SVR.

Features F1 to F5 calculate the image properties in the wavelet domain. Features F2, F3, and F5 (Ref. 14) decompose images into three scales and three orientations using the Daubechies 9/7 wavelet basis. Features F1 and F4 (Ref. 15) decompose images into two scales and six orientations using the steerable pyramid decomposition. The wavelet coefficients are parameterized using the generalized Gaussian density (GGD) model. The GGD is

$$f_X(x;\mu,\sigma^2) = ae^{-[b|x-\mu|]^\gamma}, \tag{10}$$

where μ, σ^2, and γ are the mean, variance, and shape parameter of the distribution, and

$$a = \frac{b\gamma}{2\Gamma(1/\gamma)}, \tag{11}$$

$$b = \frac{1}{\sigma}\sqrt{\frac{\Gamma(3/\gamma)}{\Gamma(1/\gamma)}}, \tag{12}$$

where $\Gamma(\cdot)$ is the gamma function.

Features F1 and F4 incorporate the divisive normalization transform (DNT) before GGD. Features F1 to F3 are the variance (σ^2) parameters of GGD and features F4 and F5 are the shape (γ) parameters.

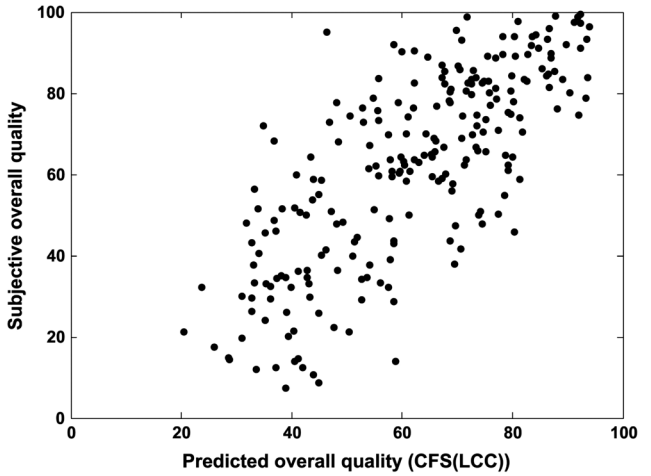

Fig. 6 Subjective overall quality as a function of predicted quality: the feature subset selected by the correlation-based feature selection (Pearson linear correlation) CFS(LCC) strategy (a) and the feature subset selected by the wrapper_SVR strategy (b).

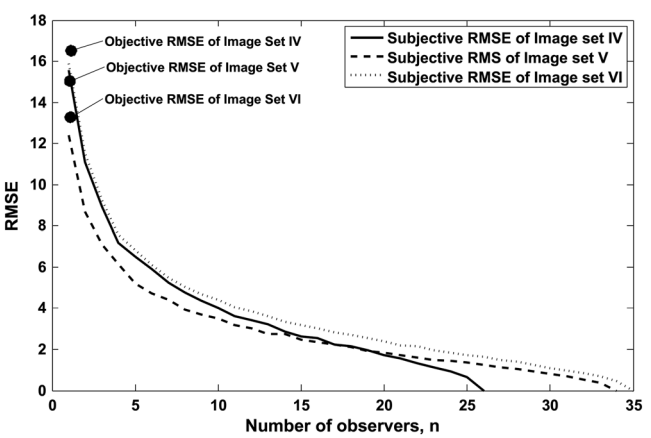

Fig. 7 Subjective root-mean-square error as a function of the number of observers for image sets IV to VI.

Feature F6 (Ref. 17) functions in the spatial domain. Image luminance values, $L(i,j)$, are locally normalized by

$$\widehat{L}(i,j) = \frac{L(i,j) - \mu(i,j)}{\sigma(i,j) + 1}, \tag{13}$$

where (i,j) are spatial indices, and

$$\mu(i,j) = \sum_{k=-3}^{3} \sum_{l=-3}^{3} w_{k,l} L_{k,l}(i,j), \tag{14}$$

$$\sigma(i,j) = \sqrt{\sum_{k=-3}^{3} \sum_{l=-3}^{3} w_{k,l} [L_{k,l}(i,j) - \mu(i,j)]^2}, \tag{15}$$

where w is a two-dimensional circularly symmetric Gaussian weighting function. The normalized luminance values are fitted to the GGD model and feature F6 is the variance (σ^2) parameter of GGD.

Features F7 to F9 (Ref. 15) capture the spatial correlation statistics. The image is decomposed into six orientations using the steerable pyramid decomposition. After the DNT, the correlation coefficients $p(\text{dist})$ are estimated between the central pixel and the pixels from the chess board distances as a function of distance (dist). Once $p(\text{dist})$ is obtained, a curve is parameterized by fitting it with polynomial function $f(x) = a_3x^3 + a_2x^2 + a_1x + a_0$. Feature F7 is the parameter a_1 and feature F8 is the parameter a_2 for the decomposition orientation of 30 deg. Feature F9 is the parameter a_3 for the orientation of 0 deg.

Feature F10 (Ref. 34) decomposes hue, saturation, value image (I) into three scales and three orientations using the Daubechies 9/7 wavelet basis. The average of wavelet coefficients, $w = (w_d + w_h + w_v)/3$, of the third wavelet scale for saturation channel I_S, is divided into 16 blocks $\{B_1, \ldots B_{16}\}$ and the value of F10 is obtained by

$$F10 = \frac{\sum_{i,j \in B_6 \cup B_7 \cup B_{10} \cup B_{11}} w(i,j)}{\sum_{b=1}^{16} \sum_{i,j \in B_b} w(i,j)}. \tag{16}$$

Feature F11 (Ref. 34) is obtained by

$$F11 = \frac{9}{HW} \sum_{i=H/3}^{2H/3} \sum_{j=W/3}^{2W/3} I_V(i,j), \tag{17}$$

where H and W are the height and width of the image.

Feature F12 (Ref. 45) computes histogram $H(i) = H(r) + H(g) + H(b)$, where r, g, and b are the red, green, and blue channels of the image. Feature 12 is the width of the middle 98% mass of the histogram H.

Feature F13 (Ref. 40) detects the edges of the image using the Sobel operator. The edge pixels are set as the central block. Each block is divided into four overlapping sub-blocks capturing directional information. The sub-blocks having the largest variance are chosen to characterize the edge areas. The statistic computed on the sub-block $L = \{l_i; 1 \le i \le M\}$, where l_i denotes the gray level is

Table 9 The descriptions of features found by wrapper_SVR strategy.

Feature	Description	Reference
G1 (f8)	Variance parameter of the GGD model for wavelet coefficients after DNT (steerable pyramids decomposition, first scale, orientation 0 deg)	15
G2 (f135)	Variance parameter of the GGD model for wavelet coefficients (Daubechies 9/7 decomposition, first scale, vertical orientation)	14
G3 (f144)	Shape parameter of the GGD model for wavelet coefficients (Daubechies 9/7 decomposition, first scale, horizontal orientation)	14
G4 (f31)	Shape parameter of the GGD model for wavelet coefficients after DNT (steerable pyramids decomposition, second scale, orientation 150 deg)	15
G5 (f124)	Sum of vertical, horizontal, and diagonal wavelet coefficients in value channel (Daubechies 9/7 decomposition, second scale, average of the diagonal)	34
G6 (f68)	Parameter of spatial correlation value function computed between central pixel and pixels from chess board distance (parameter a_2, steerable pyramids decomposition, first scale, orientation 90 deg)	15
G7 (f54)	Parameter of spatial correlation value function computed between central pixel and pixels from chess board distance (parameter a_3, steerable pyramids decomposition, first scale, orientation 0 deg)	15
G8 (f83)	Structural correlation [structural similarity (SSIM) measure] between orientations of wavelet coefficients (steerable pyramid decomposition, second scale, orientations 0 and 90 deg)	15
G9 (f93)	Structural correlation (SSIM measure) between orientations of wavelet coefficients (steerable pyramid decomposition, second scale, orientations 90 and 120 deg)	15
G10 (f116)	Average intensity value of image center area	34
G11 (f4)	It is assumed that value 128 corresponds to a well-exposed image. If mean value of pixels >128, then overexposure value is (255 – mean)/128.	33
G12 (f6)	The number of saturated pixels in 1/3 top image area	33
G13 (f220)	The image is filtered by the human visual system response,[58] and its variance is computed and divided by the mean luminance value of the image	35
G14 (f260)	The width of the middle 98% mass of the gray level histogram.	45
G15 (f264)	Standard deviation of chroma values in the CIELAB space	46
G16 (f118)	Sum of vertical, horizontal, and diagonal wavelet coefficients in hue channel (Daubechies 9/7 decomposition, second scale)	34
G17 (f122)	Sum of vertical, horizontal, and diagonal wavelet coefficients in saturation channel (Daubechies 9/7 decomposition, third scale)	34
G18 (f161)	Comparison between kurtosis values computed from edge areas of image and its low-pass filtered version (Gaussian low-pass filter: size 100 pixels; standard deviation 4 pixels)	40
G19 (f162)	Comparison between kurtosis values computed from edge areas of image and its low-pass filtered version (Gaussian low-pass filter: size 100 pixels; standard deviation 5 pixels)	40
G20 (f168)	Shape of histogram of gradient profiles from edge areas	41

$$\mathrm{kurt} = \frac{1}{M}\sum_{i=1}^{M}(l_i - \mu)^4, \tag{18}$$

where μ is the mean intensity value of the area. Feature F13 is obtained by

$$F13 = \frac{\sum_{j=1}^{N}\mathrm{kurt}_j^{\mathrm{test}} - \sum_{j=1}^{N}\mathrm{kurt}_j^{\mathrm{blur}}}{\sum_{j=1}^{N}\mathrm{kurt}_j^{\mathrm{test}} + 0.01}, \tag{19}$$

where N is the number of edge blocks in the image, and $\mathrm{kurt}^{\mathrm{test}}$ and $\mathrm{kurt}^{\mathrm{blur}}$ are indexed statistics kurt in the test image and its low-pass filtered version.

Feature F14 (Ref. 41) calculates the gradient histograms of the image edge areas. The horizontal and vertical gradient images, g_x and g_y, are computed by the Sobel operator. If $g_x > g_y$, the gradients are calculated along the horizontal direction; otherwise, they are calculated along the vertical direction. The gradient profile of an edge is calculated by taking the standard deviation of the gradients for the edge area. The gradient profile histogram of an image is divided into 100 bins. Feature F14 is obtained by

$$F14 = \sigma_{\min} \frac{(\sigma_{\max} - \sigma_{\min})}{100} b, \tag{20}$$

where $\sigma_{\min}$ and $\sigma_{\max}$ are the minimum and maximum standard deviations of the gradient profile histogram, and

$$b = \frac{\sum_{i \le T} i \cdot h_i}{\sum_{i \le T} h_i}, \tag{21}$$

where h_i is bin value and T is

$$T = \arg_T \max \left[\left(\sum_{i \le T} h_i / \sum_{i \le 100} h_i \right) < 0.03 \right]. \tag{22}$$

The hypothesis of this study was that the features of the efficient feature subset interact with and complement each other. To analyze the relationships of features F1 to F14 and the image quality, we used attribute data from the CID2013 image database. We computed LCC values between the features and the subjectively evaluated attributes of sharpness, graininess, lightness, and saturation. Principal component analysis (PCA) was used for dimension reduction to explore whether or not the features clustered as dimensions, which are expressions of different characteristics of the images.

Figure 8 shows the LCC values between the features F1 to F14 and the attributes. Features F1, F2, F3, and F6 correlate especially well with sharpness. Features F7, F8, and F9 correlate with graininess and sharpness. Feature F11 correlates strongly with lightness.

The LCC values between the features and the attribute of saturation were low. This result does not mean that saturation variations do not occur in the images; it can also mean that saturation was not as important as the other attributes in the perception of overall quality. The features of the most efficient feature subsets contributed more to the other factors; subjective evaluation data indicate the same result. Table 10 shows a cross-tabulation of LCC values for the subjective attribute data. The LCC values for the attributes of overall quality, graininess, and sharpness are high compared to the attributes of lightness or saturation. In addition, the LCC values of lightness are higher than those of saturation. The scales of lightness and saturation originally ranged from −100 (too dark/pale) to 100 (too bright/colorful). Note that before the LCC calculations, these scale values were recalculated as distances from neutral (value = 0).

Figure 9 shows the two first principal component scores of the images from image sets I to III and the principal component coefficients for each feature. The principal components were found with the princomp function in MATLAB® (R2012a). The scores of the images are the coordinates of the original feature data in the new coordinate system defined by the principal components. Each of the 14 features is represented in this plot by a vector; the direction and length of this vector indicate how the feature contributes to the two principal components in the plot. It is evident that the features point to two main directions, which are interpreted as two dimensions: features F7, F8, F9, and F12 point in one dimension (DIM2), and the other features point in the other dimension (DIM1).

From DIM1, we can identify two groups of strong vectors. Features F1, F2, and F11 form one group, while features F3, F5, and F6 form the other. According to the LCC values, features F1, F2, and F3 aligned with the DIM1 measure sharpness. Features F1, F2, and F3 make the wavelet decomposition. For decomposition purposes, feature F1 uses the steerable pyramid technique, while features F2 and F3 use

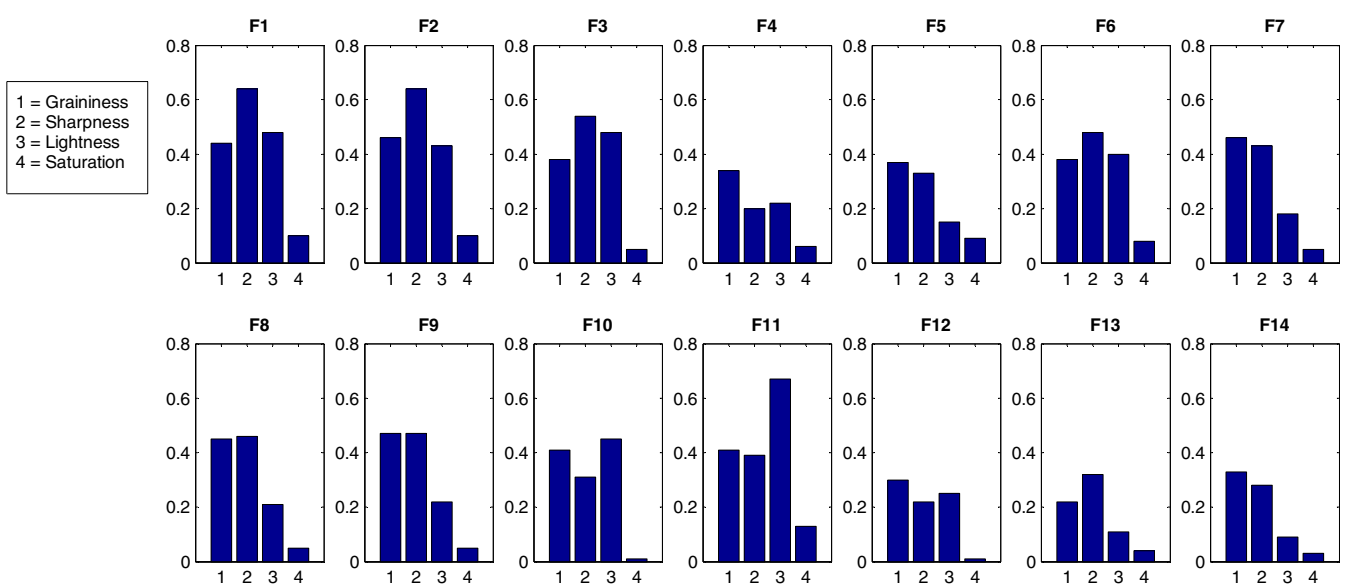

Fig. 8 LCC values for the attributes of graininess (1), sharpness (2), lightness (3), and saturation (4) for features found by the CFS(LCC) method.

Table 10 Cross-tabulated LCC values of subjectively evaluated image quality attributes.

	Overall quality	Graininess	Sharpness	Lightness	Saturation
Overall quality	1.00	—	—	—	—
Graininess	0.80	1.00	—	—	—
Sharpness	0.86	0.85	1.00	—	—
Lightness	−0.61	−0.57	−0.57	1.00	—
Saturation	−0.40	−0.35	−0.36	0.09	1.00

the Daubechies 9/7 wavelet basis. In addition, feature F1 applies DNT. The sub-band coefficients are parameterized using GGD. The GGD model has parameters of variance (σ^2) and shape (γ). One difference between features F1 and F2 and feature F3 is the scale of the decomposition; features F1 and F2 use the first scale and feature F3 uses the second scale. A low variance value in the first scale can mean detail loss (unsharp image) or an image without information (dark image). A high variance in the first scale means that the image has high-frequency energy, which may imply that the image contains many small details or noise energy. The second scale of decomposition handles mid-frequency energy; a high variance in the second scale relates to image contrast. High image contrast relates more to strong edges in the image than to small details.

The functional principles of features F1, F2, and F3 and that of feature F6 are the same: they compute the variance of intensity distribution. However, feature F6 functions in the spatial domain and features F1, F2, and F3 function in the wavelet domain. The PCA shown in Fig. 9 suggests that feature F6 and feature F3 measure similar properties (related to image contrast) from the image.

Based on the LCC values of Fig. 8, when in the direction DIM1, feature F11 measures a different image property than the other features; it computes the average intensity for the area near the center of the image. Feature F11 correlates strongly with the lightness attribute, while the other features pointing in the same direction correlate strongly with sharpness. Based on the computation process (average intensity) of feature F11, the high correlation between feature and the lightness attribute is expected.

Features F7, F8, F9, and F12 function in the DIM2 direction. Features F7, F8, and F9 compute spatial correlation values after wavelet decomposition and DNT. The values are computed for the center pixel and the pixels from the chess board distances as a function of distance. The computed values are fitted to a third-degree polynomial. Feature F7 is the second parameter of that polynomial, feature F8 is the third, and feature F9 is the fourth. The calculated spatial correlation values are higher if an image consists of smooth areas and its neighboring pixels correlate with each other. The spatial correlation is low if an image contains random-intensity variation, such as noise.

Feature F12 computes the gray-level histogram and measures the width of the middle 98% gray-level mass; a high value can mean a noisy image.

We deduce that DIM2 measures image information from the perspective of the uncertainty in predicting the value of a pixel in the image. If an image consists of random pixels without spatial correlation, DIM2 is low. If image lacks details, DIM2 is higher.

Figure 10 shows example images selected from the regular spatial locations of the principal component plane. First, we calculated the polar coordinates for the images on the plane. We then formed eight constant-size segments on the plane. The image from image group 3 with the longest radius vector was selected for each segment. The eight selected images (1 to 8) are indicated in Fig. 9 and shown in Fig. 10. It is evident that the example images suffer from different types of distortions. Image 1 has a low DIM1 value because it is dark; it has no detail energy or

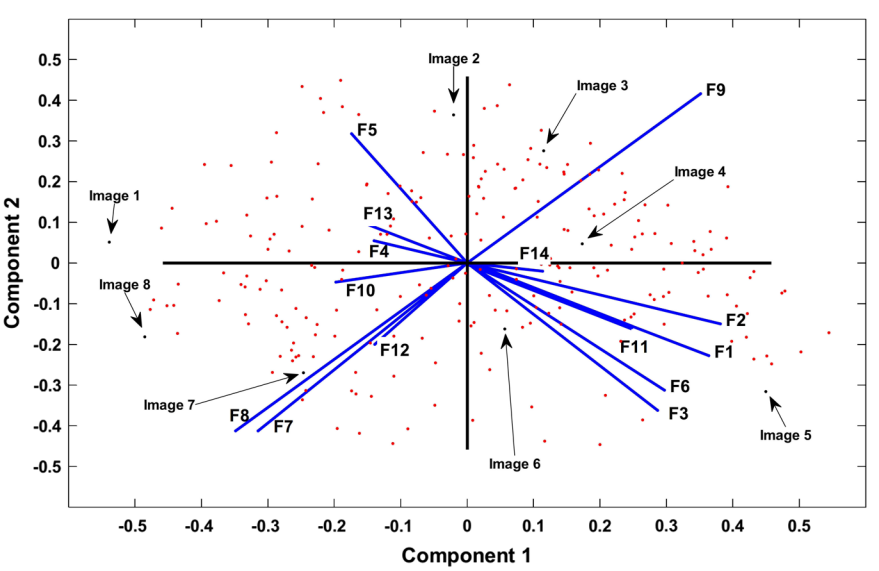

Fig. 9 Principal component scores of images (image sets I to III). Each of the features F1 to F14 is represented in this plot by a vector; the direction and length of the vector indicate how the feature contributes to components 1 and 2. Arrows point to the images shown in Fig. 10.

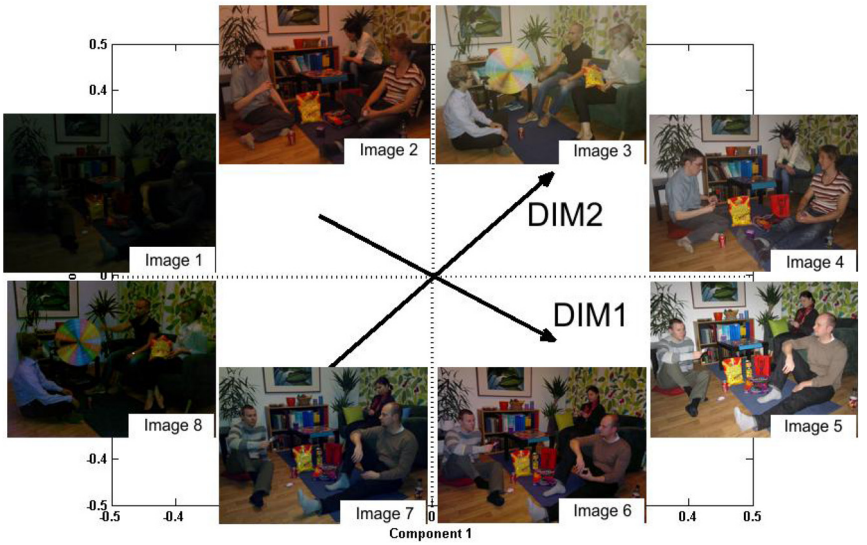

Fig. 10 Images (group 3) representing given locations of the principal component plane.

any information, which would increase the value of features directed in the direction of DIM1. Example images 3 and 7 are in the mid-range of DIM1. These images differ from each other: for example, image 3 has a high value in DIM2, while image 7 has a low value. Image 3 has some detail energy, but it is not sharp. Image 7 includes details, but it is noisy. Image 7 has more information, as its pixel values are more random than those of image 3. Image 5 has a high value in DIM1; it is sharp, bright, and noise-free.

Example images in Fig. 10 explain why feature F11 was projected in the same direction (DIM1) as the features F1, F2, F3, and F6, which measure variance values associated with details and sharpness of image. Feature F11 measures brightness (darkness) of image and a low value of DIM1 indicates dark images, such as images without details. A high value of DIM1 indicates bright images with details.

3.4 *Performance Using the LIVE MDIQD*

The performance of features F1 to F14 found by the CFS (LCC) method was evaluated and validated with the images of the LIVE MDIQD.[23] In this evaluation, we used the LIVE MDIQD images which simulate the camera image acquisition process. The 15 original images were first blurred by simulating narrow depth of field or other defocus and then corrupted by white noise by simulating sensor noise. Totally, the image set includes 225 images with subjective evaluation data.

Figure 11 shows the LCC values between features F1 to F14 and the LIVE MDIQD images. By comparing the LCC values and the strong features (long vectors) of dimensions DIM1 and DIM2 shown in Figs. 9 and 10, it can be noticed that DIM2 characterizes the LIVE MDIQD images more than DIM1. The LCC values of the strong features F7, F8, and F9 from the direction of DIM2 are high and the LCC values of the strong features F1, F2, and F6 from the direction of DIM1 are low, respectively.

Features F7, F8, and F9 of DIM2 measure image information from the perspective of the uncertainty in predicting the value of a pixel in the image. If an image consists of random pixels without spatial correlation (as white noise component in the LIVE MDIQD), DIM2 is low. If image lacks details (as blur component in the LIVE MDIQD), DIM2 is higher.

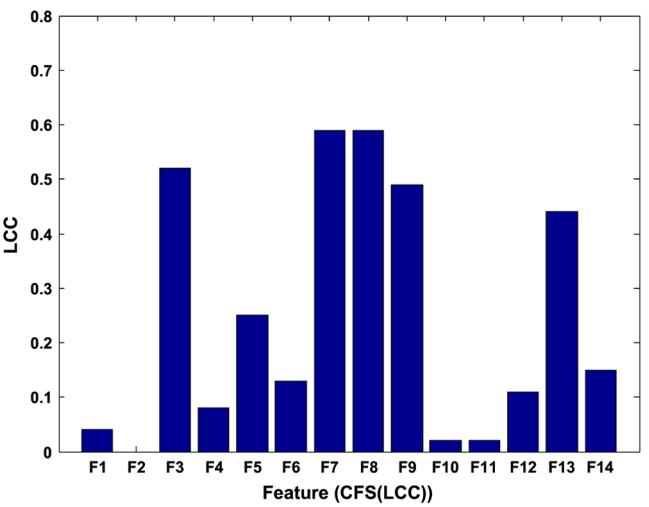

Fig. 11 Absolute LCC values for the features F1 to F14 found by CFS (LCC) method and images of the LIVE multiply distorted image quality database (study 2).

Figure 11 shows that the strong features F3 and F5 from DIM1 characterize, to some degree, the LIVE MDIQD images. A difference between the correlated strong features (F3 and F5) and the low-correlated strong features (F1, F2, and F6) from the direction of DIM1 is the scale of decomposition. Features F1, F2, and F6 were computed from the first scale of decomposition. The white noise component of the LIVE MDIQD images randomizes the values of the first scale coefficients. Features F3 and F5 were computed from the second scale of decomposition. The second scale of decomposition is robust to high-frequency white noise. Features F3 and F5 characterized the contrast (as blur component in the LIVE MDIQD images).

Comparing the CID2013 and the LIVE MDIQD we can reason that the quality space of the LIVE MDIQD is simpler.

Table 11 LCC values of the features found by CFS(LCC) strategy and BRISQUE NR quality measure for images from the LIVE multiply distorted image quality database (study 2).

	LCC	SROCC	Training set
CFS(LCC)-1	**0.576**	**0.350**	CID2013
BRISQUE-1	0.378	0.214	LIVE
CFS(LCC)-2	0.815	0.711	LIVE MDIQD
BRISQUE-2	**0.923**	**0.893**	LIVE MDIQD

Note: Bold values indicate best performance values.

The CID2013 images are corrupted by real different cameras and include at least one more dimension (characterized by features F1, F2, and F6), covering the scale of dark and noise-free images to sharp, bright, and detailed images.

We also tested the performance of the found features to predict the subjective evaluation values of LIVE MDIQD images. The SVR model was trained by two separate sets: the image sets I to VI [CFS(LCC)-1] and the LIVE MDIQD itself [CFS(LCC)-2]. With the CFS(LCC)-2, the performance was computed using 1000 randomly selected training and testing data (80/20%). We also report the performance of the state-of-the-art NR image quality assessment algorithm (BRISQUE) as provided in Ref. 23. The BRISQUE-1 was trained by LIVE standard database[18] and the BRISQUE-2 was trained by the LIVE MDIQD itself.

The results are shown in Table 11. The performance of the CFS(LCC)-1 using CID2013 for training of SVR parameters was higher than the performance of BRISQUE-1. The features found by the CFS(LCC) method and trained with the CID2013 database were more efficient than the state-of-the-art NR metric trained with the standard LIVE image database for predicting the quality of simulated camera acquisition process.

4 Conclusions

Studying consumer camera images has been a neglected domain in the field of image-quality research, as most computational measures are only useful for images suffering from a single type of distortion. This study systematically compared feature subset selection methods to find feature combinations that measured the image properties best linked to the subjectively assessed overall quality of images with multiple distortions. We used PCA and correlation analysis to find the underlying dimensions of overall quality perception. The analysis found two main dimensions: one associated with image contrast, detail reproduction, and lightness, and the other with the effect of noise energy. These results proved the hypothesis of the study: measuring the quality of real photographs requires interacting and compensating features.

According to the results, the two underlying dimensions of overall quality perception were not related to color. No doubt, color is an important feature in perceptual image quality. One reason why the best performing feature subset did not contain color-related features may be that color error is rarely the dominant distortion in the images of CID2013 although it often appears in images that suffer from blur and noise. Another reason may be that human color perception of natural images is a complex process and the feature subset simply did not include good enough features. In the future, more efficient features for predicting color error in the natural images should be developed.

This study was possible because we had access to a database of real photographs captured by a large number of cameras with multidimensional subjective evaluation data (CID2013). The other publicly available image databases include images that have undergone some specific type of distortion or two specific distortions, such as the LIVE MDIQD, and have been evaluated only for overall quality but not for quality attributes. In contrast, CID2013 includes data on the quality attributes of graininess, sharpness, color saturation, and lightness.

This study expanded the traditional scope of research by comparing the performance of objective measures and human subjects. We found that the performance of the optimal subset was comparable to the accuracy of one random human assessor. This result suggests that the optimum feature subset can be used prior to manual selection in applications that filter large image sets, such as image retrieval and editorial software systems.

References

1. W.-C. Kao et al., "Design considerations of color image processing pipeline for digital cameras," *IEEE Trans. Consumer Electron.* **52**(4), 1144–1152 (2006).
2. S. Bianco et al., "Color correction pipeline optimization for digital cameras," *J. Electron. Imaging* **22**(2), 023014 (2013).
3. L. Zhang, L. X. Mou, and D. Zhang, "FSIM: a feature similarity index for image quality assessment," *IEEE Trans. Image Process.* **20**(8), 2378–2386 (2011).
4. H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.* **15**(2), 430–444 (2006).
5. Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. of the Asilomar Conf. on Signals, Systems, and Computers*, pp. 1398–1402, IEEE, Pacific Grove, CA (2004).
6. Z. Wang et al., "Quality-aware images," *IEEE Trans. Image Process.* **15**(6), 1680–1689 (2006).
7. M. Nuutinen et al., "A framework for measuring sharpness in natural images captured by digital cameras based on reference image and local areas," *EURASIP J. Image Video Process.* **2012**(8) (2012).
8. M. Nuutinen and P. Oittinen, "Measurement of color differences of digital cameras from natural images," in *Int. Symp. on Image and Signal Processing and Analysis*, pp. 224–229, IEEE, Dubrovnik, Croatia (2011).
9. M. Nuutinen et al., "A reduced-reference method for characterizing color noise in natural images captured by digital cameras," in *Proc. of Color and Imaging Conf.*, pp. 80–85, IS&T and SID, San Antonio, TX (2010).
10. P. Marziliano et al., "Perceptual blur and ringing metrics: application to JPEG2000," *Signal Process. Image Commun.* **19**(2), 163–172 (2004).
11. N. D. Narvekar and L. J. Karam, "A no-reference image blur metric based on the cumulative probability of blur detection (CPBD)," *IEEE Trans. Image Process.* **20**(9), 2678–2683 (2011).
12. J. Zhang and T. M. Le, "A new no-reference quality metric for JPEG2000 images," *IEEE Trans. Consumer Electron.* **56**(2), 743–750 (2010).
13. H. R. Sheikh, A. C. Bovik, and L. Cormack, "No-reference quality assessment using natural scene statistics: JPEG2000," *IEEE Trans. Image Process.* **14**(11), 1918–1927 (2005).
14. A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Process. Lett.* **17**(5), 513–516 (2010).
15. A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: from natural scene statistics to perceptual quality," *IEEE Trans. Image Process.* **20**(12), 3350–3364 (2011).
16. M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: a natural scene statistics approach in the DCT domain," *IEEE Trans. Image Process.* **21**(8), 3339–3352 (2012).

17. A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.* **21**(12), 4695–4708 (2012).
18. 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.* **15**(11), 3440–3451 (2006).
19. C. J. N. Ponomarenko et al., "TID2008—a database for evaluating of full-reference visual quality assessment metrics," *Adv. Mod. Radioelectron.* **10**, 30–45 (2009).
20. P. Le Callet and F. Autrusseau, "Subjective quality assessment IRCCyn/IVC database," 2005, http://www.irccyn.ec-nantes.fr/ivcdb/ (23 September 2013).
21. D. M. Chandler and S. S. Hemami, "Supplement: performance on the A57 database (preliminary results)," 2006, http://foulard.ece.cornell.edu/dmc27/vsnr/vsnr.html (23 September 2013).
22. Y. Horita, "MICT image quality evaluation database," 2010, http://mict.eng.u-toyama.ac.jp/mictdb.html (23 September 2013).
23. D. Jayaraman et al., "Objective quality assessment of multiply distorted images," in *Proc. of the Asilomar Conf. on Signals, Systems and Computers*, pp. 1058–1697, IEEE, Pacific Grove, CA (2012).
24. R. Halonen, T. Leisti, and P. Oittinen, "The influence of image content and paper grade on quality attributes computed from printed natural images," in *Proc. NIP*, pp. 459–462, IS&T and ISJ, Pittsburgh, PA (2008).
25. S. Bhattacharya, R. Sukthankar, and M. Shah, "A framework for photo-quality assessment and enhancement based on visual aesthetics," in *Proc. ACM Multimedia*, pp. 271–280, ACM, Firenze, Italy (2010).
26. S. Dhar, V. Ordonez, and T. L. Berg, "High level describable attributes for predicting aesthetics and interestingness," in *Proc. CVPR*, pp. 1657–1664, IEEE, Colorado Springs, CO (2011).
27. A. K. Moorthy, P. Obrador, and N. Oliver, "Towards computational models of visual aesthetic appeal of consumer videos," in *Proc. ECCV*, Vol. 6315, pp. 1–14, Springer, Crete, Greece (2010).
28. H. Liu and L. Yu, "Towards integrating feature selection algorithms for classification and clustering," *IEEE Trans. Knowl. Data Eng.* **17**(4), 491–502 (2005).
29. T. Virtanen et al., "CID2013—a real photographic image database," *IEEE Trans. Image Process.*, in press (2014).
30. J. Radun et al., "Evaluating the multivariate visual quality performance of image-processing components," *ACM Trans. Appl. Percept.* **7**(3), 16 (2010).
31. R. Gong et al., "Investigation of perceptual attributes for mobile display image quality," *Opt. Eng.* **52**(8), 083104 (2013).
32. M. Pedersen et al., "Attributes of image quality for color prints," *J. Electron. Imaging* **19**(1), 011016 (2010).
33. V. Ojansivu et al., "Degradation based blind image quality evaluation," *Lec. Notes Comput. Sci.* **6688**, 306–316 (2011).
34. R. Datta et al., "Studying aesthetics in photographic images using a computational approach," in *Proc. ECCV*, Vol. 3, pp. 288–301, Springer, Graz, Austria (2006).
35. A. Ciancio et al., "No-reference blur assessment of digital pictures based on multifeature classifiers," *IEEE Trans. Image Process.* **20**(1), 64–75 (2011).
36. C. Li, A. C. Bovik, and X. Wu, "Blind image quality assessment using a general regression neural network," *IEEE Trans. Neural Netw.* **22**(5), 793–799 (2011).
37. M. A. Saad, A. C. Bovik, and C. Charrier, "A DCT statistics-based blind image quality index," *IEEE Signal Process. Lett.* **17**(6), 583–586 (2010).
38. 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.* **18**(4), 717–728 (2009).
39. Z. Wang, H. R. Sheikh, and A. C. Bovik, "No-reference perceptual quality assessment of JPEG compressed images," in *Proc. IEEE Int. Conf. on Image Processing*, pp. 477–480, IEEE, Rochester, New York (2002).
40. C. Li et al., "No-reference blur index using blur comparisons," *Electron. Lett.* **47**(17), 962–963 (2011).
41. L. Liang et al., "A no-reference perceptual blur metric using histogram of gradient profile sharpness," in *Proc. of 16th IEEE Int. Conf. on Image Processing*, pp. 4369–4372, IEEE, Cairo, Egypt (2009).
42. A. V. Murthy and L. J. Karam, "MATLAB based framework for image and video quality evaluation," in *Proc. QoMEX*, pp. 242–247, IEEE, Trondheim, Norway (2006).
43. J. Caviedes and S. Gurbuz, "No-reference sharpness metric based on local edge kurtosis," in *Proc. of IEEE Int. Conf. on Image Processing*, pp. 53–56, IEEE, Rochester, New York (2002).
44. S. Gabarda and G. Cristóbal, "Blind image quality assessment through anisotropy," *J. Opt. Soc. Am. A* **24**(12), B42–B51 (2007).
45. Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *Proc. of IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pp. 419–426, IEEE, New York, NY (2006).
46. D. Hasler and S. Susstrunk, "Measuring colorfulness in natural images," *Proc. SPIE* **5007**, 87–95 (2003).
47. E. P. Simoncelli, "MATLAB PyrTools toolbox," 2001, http://www.cns.nyu.edu/~lcv/software.html (14 June 2012).
48. Z. Wang et al., "Image quality assessment: from error measurement to structural similarity," *IEEE Trans. Image Process.* **13**(4), 600–612 (2004).
49. Z. Wang, "The SSIM index for image quality assessment, 2004, https://ece.uwaterloo.ca/~z70wang/research/ssim/index.html (14 June 2012).
50. P. D. Konvasi, "MATLAB and octave functions for computer vision and image processing," 1996, http://www.csse.uwa.edu.au/~pk/research/matlabfns/PhaseCongruency/phasecong3.m (14 June 2012).
51. M. A. Hall and L. A. Smith, "Feature selection for machine learning: comparing a correlation-based filter approach to the wrapper," in *Proc. of 12th Int. Florida Artificial Intelligence Research Society Conf.*, pp. 235–239, EAAI, Orlando, Florida (1999).
52. L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *J. Mach. Learn. Res.* **5**, 1205–1224 (2004).
53. C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," 2011, http://www.csie.ntu.edu.tw /~cjlin/libsvm/ (17 August 2012).
54. I3A, "CPIQ initiative phase 1 white paper—fundamentals and review of considered test methods," 2007, http://www.ivl.disco.unimib.it/Teaching/EI%20specilistica%20PDF%206xP%202011/cpiq_white_paper.pdf (28 August 2014).
55. ITU-R Recommendation BT.500-11, *Methodology for the subjective assessment of the quality of television pictures*, International Telecommunication Union, Geneva, Switzerland (2002).
56. A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.* **20**(3), 209–212 (2013).
57. R. R. Wilcox, *Basic Statistics: Understanding Conventional Methods and Modern Insights*, Oxford University Press, New York, NY (2009).
58. J. L. Mannos and D. J. Sakrison, "The effect of a visual fidelity criterion on the encoding of images," *IEEE Trans. Inf. Theory* **20**(4), 525–536 (1974).

Mikko Nuutinen received his MSc (Tech) and LicSc (Tech) degrees from the Helsinki University of Technology in 2004 and 2007, respectively, and his DSc (Tech) degree from Aalto University in Helsinki in 2012. He is a postdoctoral researcher at the University of Helsinki in the Institute of Behavioral Sciences. His current research interests are in the areas of objective image quality assessment, color image processing, camera performance measurements, and subjective image quality assessment methods and analysis.

Toni Virtanen has been mainly involved in image quality research. His background comes from psychology, in which he received a master's degree in 2010. His main occupation has been developing and conducting subjective image quality experiments in a collaboration project with Nokia at the University of Helsinki Visual Cognition research group. He has been with a long-lasting project since 2005 and is currently working as a project manager in addition to his efforts toward his doctoral thesis on related topics.

Pirkko Oittinen is a full professor in the Department of Media Technology of Aalto University School of Science. Her research has the mission of advancing visual technologies and raising the quality of visual information to create enhanced user experiences in different usage contexts. The research approach is constructive and seeks to cross disciplinary boundaries. Current research focuses on interrelations between computational characteristics of still and moving images and visual experience.