Vision models for image quality assessment: one is not enough

Roland Brémond
Jean-Philippe Tarel
Eric Dumont
Nicolas Hautière
Université Paris Est, Laboratoire Exploitation, Perception, Simulateurs et Simulations
Institut National de Recherche sur les Transports et leur Sécurité-Laboratoire Central des Ponts et Chaussées
58 Boulevard Lefebvre, 75015 Paris, France
E-mail: roland.bremond@lcpc.fr

Abstract. *A number of image quality metrics are based on psychophysical models of the human visual system. We propose a new framework for image quality assessment, gathering three indexes describing the image quality in terms of visual performance, visual appearance, and visual attention. These indexes are built on three vision models grounded on psychophysical data: we use models from Mantiuk et al. (visual performance), Moroney et al. (visual appearance), and Itti, Koch, and Niebur (visual attention). For accuracy reasons, the sensor and display system characteristics are taken into account in the evaluation process, so that these indexes characterize the image acquisition, processing, and display pipeline. We give evidence that the three image quality indexes, all derived from psychophysical data, are very weakly correlated. This emphasizes the need for a multicomponent description of image quality.* © *2010 SPIE and IS&T.* [DOI: 10.1117/1.3495989]

1 Introduction

The recent development of digital image acquisition technologies leads to better image quality in terms of spatial resolution and sensitivity.[1] At the same time, image display technologies are rapidly changing, achieving better resolution and luminance dynamic range,[2] allowing us to take advantage of the progress in the image acquisition process. This trend leads to a growing need for quantitative evaluation criteria in terms that depend on the application, e.g., movies, video games, radiology, driving simulators, hard-copy printers, etc. The recently created eighth division of the *Commission Internationale de l'Eclairage*[3] is a step in this direction, and a sign that these issues are of interest both in industrial and scientific terms.

Image quality was first considered by painters, then in the technical fields of photography, hard-copy printing, and television. In the field of image processing, image quality metrics based on the human visual system (HVS) were first devoted to include HVS models in the evaluation of image transforms and image distortion (see Fig. 1), and are now widely used for image compression.[4] In 1993, two seminal papers addressed the question from different viewpoints. Daly proposed a visual difference predictor (VDP) that allowed comparison of digital images in terms of visibility for the HVS,[5] while Tumblin and Rushmeier proposed the concept of tone reproduction operator to the field of computer graphics.[6] Since then, a number of important steps have been made in vision science and computer science, allowing a new framework for image quality metrics to be proposed.

A number of vision models allow image comparisons, using what Tumblin and Rushmeier called an *observer model,* to compare the consequences for a given model of the HVS of some changes in an image. In this work, we include state of the art HVS-based image quality metrics in a general framework, assessing the quality of a displayed image with the broadly accepted idea that image quality indexes should be described in terms of the HVS: what do people see/fail to see of these images? What do people stare at? How do the images appear to them? Is it the same looking at the physical scene and the displayed image?

Our image quality indexes are chosen among available models accounting for the main aspects of vision. The main image quality criteria in the field of hard-copy printers and digital photography are based on color appearance models (CAM),[7] while VDP seems to be the most popular in image processing and computer graphics applications. However, these models only address specific aspects of the HVS, and other descriptions of the visual behavior can be found, including vision models inspired from neurosciences.

Looking for an approach consistent across industries and applications, we consider the visual behavior as a starting point to derive the main aspects of vision. This led us to three indexes considering the visual performance, visual appearance, and visual attention aspects of image quality. This choice is based on a coarse overlook at the large variety of available models of the HVS, each taking into account a small part of the actual visual behavior. We selected three indexes derived from broadly used vision models. The VDP describes the HVS in terms of visual performance,[5] the CIECAM02 in terms of visual appearance,[8] and the saliency map in terms of visual attention,[9] with all three models

Paper 09161RR received Sep. 2, 2009; revised manuscript received Jul. 8, 2010; accepted for publication Aug. 16, 2010; published online Dec. 9, 2010

1017-9909/2010/19(4)/043004/14/$25.00 © 2010 SPIE and IS&T.

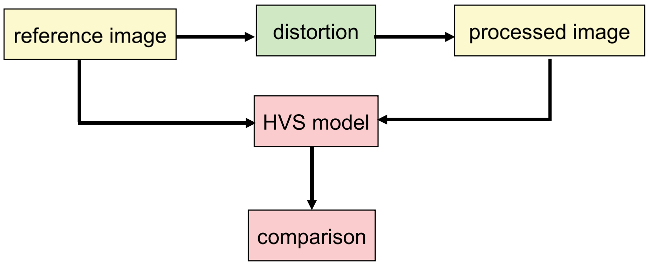


Fig. 1 Standard framework for HVS-based image quality metrics, such as Refs. 5 and 12. In the following figures, yellow boxes refer to images, green boxes refer to image processing, and pink boxes refer to human vision components.

being validated against psychophysical data. However, other models are discussed. The weak correlation of image quality indexes derived from these well-known models is the main contribution of this work, meaning that a multicomponent description of image quality is needed.

In most coding applications, image quality is assessed without the knowledge of sensor and display properties. In the following, however, we consider the full digital image acquisition-processing-display pipeline to assess the quality of the displayed image compared to the real scene, that is, image fidelity with respect to what should be seen. Our framework is based on a comparison between two images: a “world” image I_w, which is the reference scene one tries to reproduce with the image pipeline, and a “displayed” image I_d. The “world” image can be either derived using the sensor properties, or computed with a virtual sensor in the case of physical-based image rendering.[10] Any image processing can happen between the sensor output and the display input, such as image compression, gamut mapping, etc. As the image comparison is set in terms of human vision, the image representation uses photometric and geometric units.

The following sections of this work are divided in three main parts.

- Section 3 addresses the image acquisition issue (Sec. 3.1), that is, the “world” image I_w (sensor input data) estimation from the sensor output data and sensor properties. Then, it addresses image display (Sec. 3.3), that is, the “displayed” image estimation I_d from the input image and the display device properties. Examples of the calibration of a Nikon D200 camera (Sec. 3.2) and a NEC 1701 LCD monitor (Sec. 3.4) are proposed.
- Section 4 addresses image quality and proposes a framework for the comparison between the “world” and “display” images (Sec. 4.1) using three components of image quality dealing with visual performance (Sec. 4.2), visual appearance (Sec. 4.3), and visual attention (Sec. 4.4). The selected vision models allow us to build image quality indexes.
- Section 5 shows, on a database of 53 calibrated images, only weak correlations between the three indexes, which emphasizes the needs for a multicomponent HVS-based image quality metric. A new image quality index is proposed, including components from the three selected models, allowing a user-defined tuning.

2 Related Work

2.1 *Image Quality Metrics*

Engeldrum described the image quality circle from an imaging system designer point of view, allowing the rating of the impact of technological variables on the customer’s subjective preferences when looking at displayed images.[11] Although this approach is restricted to visual appearance quality indexes (sharpness, graininess, lightness, etc.), it highlights two major aspects of image quality relevant to other kind of indexes. The direct approach rates displayed images in terms of perceptive attributes, such as sharpness, etc., without any reference to the physical scene. The comparative approach, which is followed in this work, uses perceptive metrics to rate the difference between the displayed image and a reference image (see Fig. 1).

One of the first image quality metrics in the field of digital images to use a psychophysical model of the HVS was proposed by Mannos and Sakrison, where the contrast sensitivity function (CSF) of the human eye was modeled.[12] Daly[5] (further refined in Ref. 13) and Lubin[14] went deeper into the HVS limits. The VDP takes into account masking effects, threshold versus intensity (TVI) data, and a visual cortex model,[15] while Lubin’s model is closely inspired by the retina physiology. These models take two luminance images as input and compute a visibility map as output, where pixel values are understood as predictions about the visibility of a possible difference between the input images at this location. Recently, Wang et al. proposed a metric based on a HVS property first emphasized in the Gestalt theory: the sensitivity to image structure.[16] Structure similarity is described in statistical terms (covariance between two input images), in addition to a classical comparison of luminance and contrast (image variance) data. Ferwerda and Pellacini proposed a functional difference predictor (FDP), which takes into account the visual task.[17] Ramanarayanan et al.[18] proposed an image quality metric called the visual equivalence, which is less conservative than the VDP in the sense that two images with visible differences can still be felt equivalent in terms of material, illumination, and objects shape.

The aim of these models is to predict whether a difference between two images would be visible for a human observer. Another important aspect of image quality is the visual appearance, which depends on human judgments rather than on visual performances. Brightness is a key aspect of appearance,[6] however, most image quality models focus on color appearance.[7] The CIE proposed the CIECAM97 and CIECAM02 models,[8,19] while the iCAM model was recently proposed for digital image applications.[20] In terms of image quality metrics, these models can be seen as predictive models of the HVS about the visual appearance of the images, and stand on psychophysical data. More issues about visual appearance have been investigated, such as realness and naturalness.[21–23]

Perceptually based image quality metrics also concern the embedded HVS models in computer graphics rendering algorithms. These algorithms are based on the idea that dramatic performance gains can be expected if one avoids spending computer power on issues that are not perceived by human observers.[24,25] Image quality metrics may benefit from some of these advances, and some tone mapping operators (TMO) also use models of the HVS that may contribute to such metrics. By splitting a TMO into a vision model and a display model (see a good example in (Ref. 26), one can extract the first one and treat it as a HVS predictor in a VDP-like framework. For instance, Tumblin and Rushmeier[6] use a brightness model from Stevens' psychophysical data;[27] Ward[28] proposes a linear algorithm based on a visibility threshold model using Blackwell's data[29]; Ferwerda et al.[30] propose a model for visual masking, using data from Legge and Foley;[31] Pattanaik et al.[26] take into account the bandpass mechanisms of spatial vision[32] and use Peli's definition of local contrast,[33] and so on.

2.2 *Psychophysical Assessment of Image Quality*

Psychophysical assessment of HVS-based image quality metrics have been performed both in the field of image compression and imaging system design. However, as stated by Eckert and Bradley, "There seems to be as many psychophysical techniques used to validate metrics as there are metrics."[4] Roughly, these techniques use either rating scales,[34] pair comparisons,[35] both assessing suprathreshold differences, and just noticeable differences (JND) tests,[36] assessing threshold values. The weak correlation between these methods and the sensitivity to the detail of the psychophysical experimental protocol are related to various cognitive aspects of the task, such as search strategy, learning effects, detail of the instructions given to the observers, etc.[37]

The direct psychophysical evaluation of image quality, without any reference to any image quality metric, is an increasing topic in image rendering,[38] which puts into focus the lack of ready-to-use HVS-based image quality metric for computer graphics applications. This trend began with Meyer et al.[39] and uses either image comparisons, low dynamic range (LDR) versus high dynamic range (HDR) display device comparisons,[40] or image displays versus physical scenes.[22,23,41] The psychophysical methodology of these experiments mostly uses judgment evaluations, and sometimes a visual performance.[42]

The fact that no image quality metric could be built, to date, from such experiments can be easily understood if one recalls that HVS models are made of limited data compared to the HVS complexity in actual situations, so that researchers feel the need for experimental evidence when trying to assess the image quality with respect to complex aspects of human vision.

2.3 *Computational Models of Visual Attention*

Visual attention is a major topic in today's neuroscience.[43,44] Selective visual attention drives the gaze direction and saccadic motion toward salient items. Two mechanisms are combined to complete this process: image-based bottom-up mechanisms are preattentive and data driven (that is, image driven) to select the most salient area, while top-down mechanisms introduce task-dependent biases as well as prior knowledge on the image content and object relations.

Computational models of visual attention originate from Koch's hypothesis that a unique saliency map takes into account the low level features of the HVS for the selection of spatial visual attention.[45] It is a computational approach of Treisman's feature integration theory[46] for the bottom-up attentional process. Itti, Koch, and Niebur derived a very popular computational model from this hypothesis.[47] The model makes a prediction, using a retinal image as input data, about where the focus of attention will shift next. It is a HVS model validated against oculometric data[9] (see also Ref. 48 for the psychometric validation of a dynamic version of this model[49]). Instead of predicting visual appearance or performance, it aims at predicting the visual behavior in a more natural way.

Most models of bottom-up attention use the concept of saliency map, which gathers in a single map of the saliency of each spatial location on the retina.[50] This map is computed from the input image using early visual features of the vision process (center-surround mechanisms, color opponency, etc.). Even if this unique saliency map was not based, at first, on physiological data, the early vision image processing steps are designed to mimic the biological information processing, e.g., the winner-takes-all selection in the thalamus nuclei. Biologically plausible computational models of visual attention have been used since in computer graphics applications[51] as well as image compression[52] and robotics.[53]

These models are relevant for image quality assessment, because they allow us to predict an important part of the visual behavior: what area of an image attracts the observer's attention? This makes the saliency map a good candidate to build an image quality metric on.

3 Image Pipeline

Using HVS-based image quality metrics needs an image representation in units compatible with a vision model, such as luminance, chromaticity, XYZ, or LMS. Although in some cases such as image coding, standard matrices (such as sRGB) are available to convert RGB signal to XYZ signal, a more accurate calibration of the sensor and display devices allows better relevance of the image quality assessment. Thus, we split the image pipeline into three steps to properly assess the image quality with reference to the scene captured by the sensor. The first step is the estimation of the input stimulus I_w from the sensor output, knowing the sensor properties (Fig. 2). Then, image processing steps such as

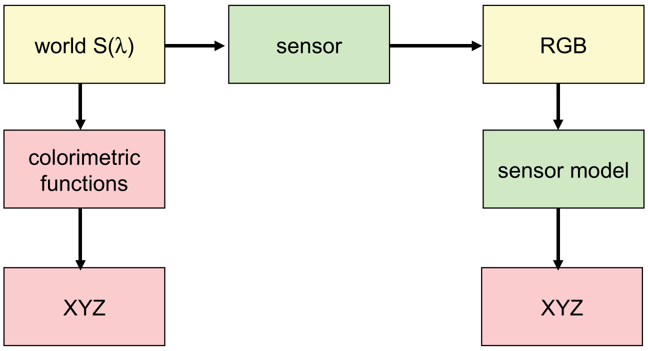


Fig. 2 A sensor model allows us to estimate the input XYZ tristimulus from the sensor output RGB values.

compression and tone mapping can be applied to the images. Finally, the image I_d is displayed on a screen or printed. In this section, we focus on practical ways to compute the I_w and I_d images. When necessary, pixel indexes (i, j) are removed for easier reading.

3.1 *Estimation of the Sensor Input*

The I_w image is estimated using a reference sensor, which may be the acquisition sensor in the image pipeline one wants to assess, or a high quality sensor if one wishes to compare images captured with various sensors against the same reference image I_w. Another reason why one may use a reference sensor different from the pipeline sensor is that an accurate estimation of I_w can be useful when using a low quality sensor in the image pipeline.

The full calibration of a sensor means that one can estimate the photometric and colorimetric values of the input signal, say in XYZ units, from the output data (RGB values). However, full accuracy is not possible unless the camera is a true metrological sensor, that is, the spectral sensitivities of the filters linearly depend on the CIE colorimetric functions $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, and $\bar{z}(\lambda)$. In the general case, the calibration can be estimated,[54,55] allowing us to compute I_w in XYZ units from the sensor RGB output image. However, simpler approximations can be proposed. In the following, we divide the calibration into two steps:

- a radiometric calibration, allowing the transformation (on each channel) of the RGB signal into an intensity signal with a linear response with respect to the input luminance
- a colorimetric calibration of the linearized (corrected) sensor, allowing the estimation of the XYZ values from the intensity values on each channel, knowing the spectral sensitivities of the sensor filters.

The radiometric calibration may be done following Ref. 56 (provided that the response function fits a power function) or Ref. 57 (with a polynomial approximation). For a linear sensor, like most charge-coupled device (CCD) sensors, this step may be unnecessary. Then, the colorimetric calibration can be done using the spectral sensibilities of the filters, which can be measured with a monochromator. In the following, these sensitivity functions are denoted $C_k(\lambda)$, where λ is the wavelength and k is the channel (R, G, and B). The sensor output at pixel (i, j) on channel k is:

$$C_{i,j,k} = \int_\lambda S_{i,j}(\lambda) C_k(\lambda) d\lambda, \tag{1}$$

where $S_{i,j}(\lambda)$ is the input spectral distribution. The trichromatic XYZ input stimulus at the same pixel is computed from $S(\lambda)$ using the CIE colorimetric functions:

$$X = K \int_\lambda \bar{x}(\lambda) S(\lambda) d\lambda, \quad Y = K \int_\lambda \bar{y}(\lambda) S(\lambda) d\lambda, \quad \text{and}$$

$$Z = K \int_\lambda \bar{z}(\lambda) S(\lambda) d\lambda,$$

with $K = 683\ lm.W^{-1}$. From Eq. (1), one may wish to compute a linear transform from RGB to XYZ units. To do so, Glassner proposed to reduce the function space for $S(\lambda)$ from infinity (due to the infinite number of wavelength values) to a 3-D space:[58]

$$S(\lambda) = \sum_{g=1}^{3} a_g F_g(\lambda). \tag{2}$$

The estimation of $S(\lambda)$ from the sensor output is fully addressed in Ref. 59. We address here a strong restriction of this problem: estimating $[X, Y, Z]$ from $[R, G, B]$. To this end, estimating $S(\lambda)$ is an intermediary step, and errors on $S(\lambda)$ may be of little importance if they lead to small errors on $[X, Y, Z]$. Equation (1) can be rewritten: $[R, G, B]^T = \mathbf{Ma}$ and $[X, Y, Z] = \mathbf{Na}$, where $\mathbf{a} = [a_1, a_2, a_3]^T$, and:

$$M_{1,j} = \int_\lambda F_j(\lambda) R(\lambda) d\lambda, \quad N_{1,j} = K \int_\lambda F_j(\lambda) \bar{x}(\lambda) d\lambda,$$

$$M_{2,j} = \int_\lambda F_j(\lambda) G(\lambda) d\lambda, \quad N_{2,j} = K \int_\lambda F_j(\lambda) \bar{y}(\lambda) d\lambda,$$

$$M_{3,j} = \int_\lambda F_j(\lambda) B(\lambda) d\lambda, \quad N_{3,j} = K \int_\lambda F_j(\lambda) \bar{z}(\lambda) d\lambda.$$

Once a function family F_k is chosen, the XYZ values may be computed from:

$$[X, Y, Z]^T = \mathbf{NM}^{-1} [R, G, B]^T. \tag{3}$$

In the following, we call $\mathbf{T} = \mathbf{NM}^{-1}$ the transform matrix. Wandell[60] uses a constant function, plus a sine and a cosine for F_1, F_2, and F_3 with good results, while Ref. 59 proposes a more complete analysis of $S(\lambda)$ estimation. In the next section, we followed Heikkinen et al.[59] and tested several function families, including some with more than three functions in the basis (with a regularization), looking for an optimal $\mathbf{T}$ matrix for a Nikon D200 camera.

3.2 *Application: Nikon D200 Camera*

In this section, we propose a practical example of how the reference image I_w can be computed. A reference physical scene was captured under five controlled illuminations using a Munsell SpectraLight III light box (Velbert, Germany) in a dark room (walls painted in black, no windows): a illuminant (incandescent light), D65 illuminant (average sky), Horizon, TL 84, and Cool White. A Greta McBeth Color Checker chart was included in the scene (Fig. 3). For each light source, a three-channel 12-bit raw image was recorded with a D200 Nikon digital camera. The radiometric linearity of the sensor was first checked with fair results using the 12 gray-level patches around the white square in the middle of the chart. The spectral distribution of the light sources were measured with a spectrocolorimeter Minolta CS 1000. Three were continuous (Horizon, D65, and A), and two were discontinuous (TL 84 and Cool White), see Fig. 4(a).

The normalized spectral sensitivities of the camera filters were measured with a monochromator Optronic (Orlando, Florida) 740A/D [Fig. 4(b)], allowing us to compute $\mathbf{T}$ matrix with various function families: the D200 sensitivity functions; the CIE $\bar{x}$, $\bar{y}$, and $\bar{z}$ colorimetric functions; polynomial basis with degs 2, 4, 8, and 16; Gaussian functions with 4, 8, and 30 translations; Fourier basis with 1, 2, and 3 periods (with and without a linear function); and sRGB. Some among the tested matrix were computed using more than three functions in the basis, with a regularization (see also, Ref. 59), that is, minimizing $\|[R, G, B]^T - \mathbf{Ma}\|^2 + q\|\mathbf{a}\|^2$

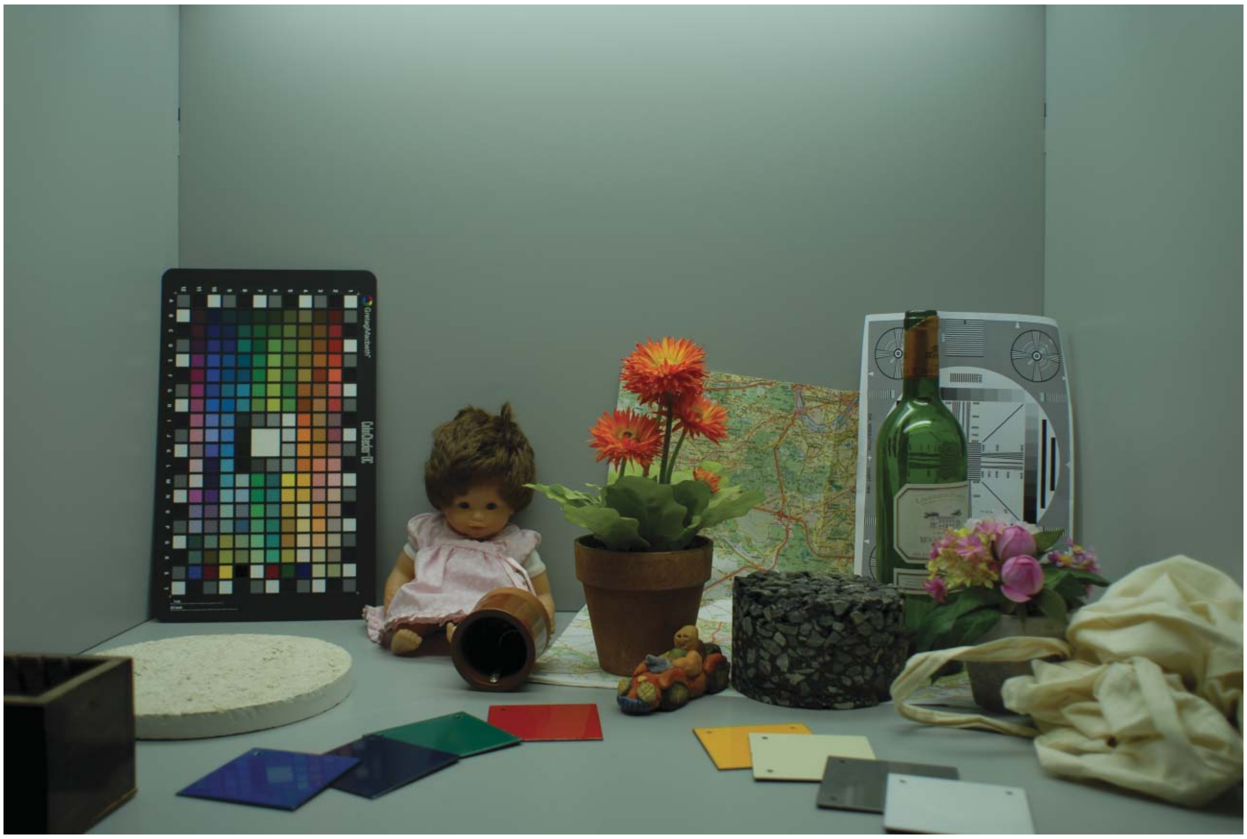


Fig. 3 The *Little Lucie* image under the *Cool White* light source (JPEG photograph).

with respect to **a**, which leads to:

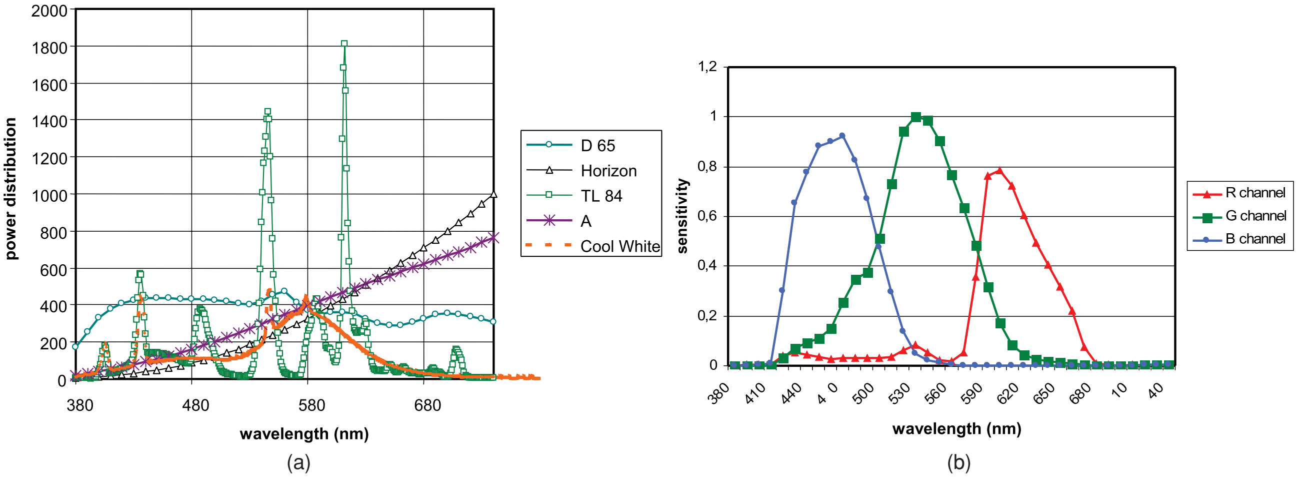
$$\mathbf{M}^T[R, G, B]^T = (\mathbf{M}^T\mathbf{M} + q\mathbf{I})\mathbf{a}. \tag{4}$$

(**I** is for the identity matrix), and finally:

$$\mathbf{T} = \mathbf{N}(\mathbf{M}^T\mathbf{M} + q\mathbf{I})^{-1}\mathbf{M}^T. \tag{5}$$

For each matrix, we computed a mean error, comparing the computed (x, y) colors of the 180 patches of the color checker to the estimated (x, y) colors of the same patches in the image. This was done for the five *Little Lucie* raw images (with the five light sources). Using $q = 1$ (**T** is not very sensitive to q), we found that the best basis for the five light sources corresponds to the D200 RGB sensibility functions. It means that the **T** matrix is built by assuming that the input spectra are linear combinations of the sensor sensitivity curves [Fig. 4(b)]. In the following, we use the **T** matrix computed from these functions:

$$\mathbf{T} = k_L \begin{bmatrix} 99.086 & 20.185 & 8.472 \\ 42.821 & 73.659 & -14.630 \\ 4.156 & -13.290 & 123.251 \end{bmatrix}, \tag{6}$$



(a)

(b)

Fig. 4 (a) Spectral measurements of the Spectra Light III sources. (b) Normalized spectral sensitivities of the D200 sensors.

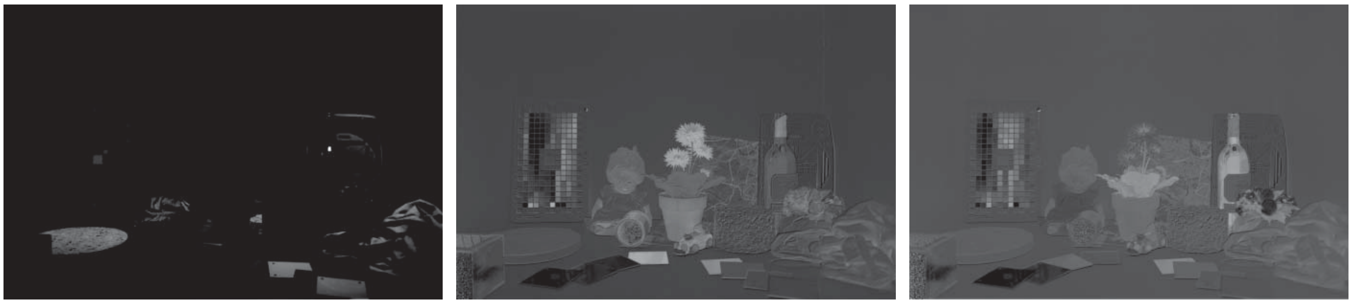


Fig. 5 Luminance (left) and chromaticity (middle and right) estimates of the *Little Lucie* D65 image, computed from the raw data of a D200 Nikon camera. The normalized luminance image looks dark because of the specular reflection on the bottle.

where k_L is a luminance normalization factor. Measuring the white patch luminance in the chart allows us to compute this factor without knowing the camera settings (shutter speed, aperture, etc.). For instance, we found $k_L = 8.22$ for the D65 *Little Lucie* image. Figure 5 shows the XYZ estimation, with this matrix, for the D65 raw image.

3.3 *Estimation of the Display Output*

Display devices expect, as input, the sensor output data, i.e., an array of digital values on RGB channels. The display technology (CRT, LCD, DLP, etc.) transforms these digital values into displayed luminance and chromaticities. The perception of such images depends on geometric conditions (e.g., pixel angular size) and lighting conditions (e.g., surround luminance).

In the proposed framework, the key point is to estimate the photometric and colorimetric properties of the image that are actually seen by an observer. The displayed image I_d is described in XYZ units, as was the world image I_w, allowing further comparison (see Fig. 6). A photometric and colorimetric calibration of the display device is needed,[61,62] measured in a situation as close as possible to the viewing conditions. Thus, the display device model allows us to compute the displayed XYZ signal from the RGB digital values at each pixel. For simplicity reasons, we used a gain-offset-gamma model:

$$\begin{bmatrix} X - X_0 \\ Y - Y_0 \\ Z - Z_0 \end{bmatrix} = \begin{bmatrix} \frac{x_R}{y_R} & \frac{x_G}{y_G} & \frac{x_B}{y_B} \\ 1 & 1 & 1 \\ \frac{z_R}{y_R} & \frac{z_G}{y_G} & \frac{z_B}{y_B} \end{bmatrix} \begin{bmatrix} g_R R^{\gamma_R} \\ g_G G^{\gamma_G} \\ g_B B^{\gamma_B} \end{bmatrix}, \qquad (7)$$

where (X_0, Y_0, Z_0) is the offset, γ_k is the gamma factors, g_k is the gain, and (x_k, y_k) is the chromaticity of channel k [we denote $z_k = 1 - (x_k + y_k)$ in Eq. (7) for easier reading]. We call $\mathbf{D}$ the matrix, allowing us to rewrite Eq. (7):

$$[X, Y, Z]^T = [X_0, Y_0, Z_0]^T + \mathbf{D}[R^{\gamma_R}, G^{\gamma_G}, B^{\gamma_B}]^T. \qquad (8)$$

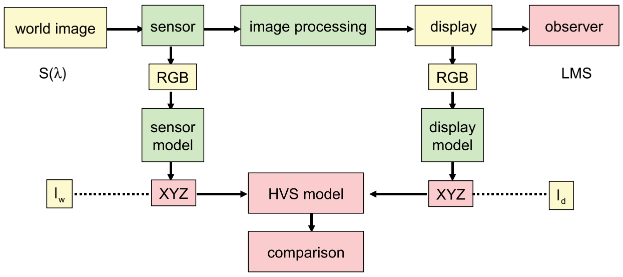


Fig. 6 The image pipeline: comparison of the I_w and I_d images.

3.4 *Application: NEC 1701 Liquid Crystal Display Monitor*

The 12-bit raw images from the Nikon D200 (Sec. 3.2) are processed by the camera firmware, leading to 8-bit JPEG images. Next, these images are displayed on a NEC 1701 LCD monitor, which was characterized in our laboratory, so that a gain-offset-gamma model could be applied to the RGB data to estimate the XYZ displayed values of I_d (see Fig. 7). We used the following matrix $\mathbf{D}$:

$$\mathbf{D} = \begin{bmatrix} 74.95 & 51.39 & 24.62 \\ 38.90 & 107.09 & 11.35 \\ 11.56 & 15.47 & 134.56 \end{bmatrix}, \qquad (9)$$

with $\gamma_R = 2.80$, $\gamma_G = 2.99$, and $\gamma_B = 2.97$. The luminance of the black is 0.4 cd/m^2.

Figure 7 shows the same physical components as in Fig. 5 for the displayed image. A direct comparison of these luminance and chromaticity images is not straightforward, as it depends on the media (monitor, printed paper, etc.), so that perceptual metrics are needed. However, some differences are visible, such as the light reflection on the wine, bottle which is attenuated in the displayed JPEG image.

Note that without sensor and display calibration, computing error maps between I_w and I_d would produce biased results. What is the bias? It can be computed, with the estimated transfer matrices:

$$I_w - I_d = k_L\mathbf{T}[R, G, B] - ([X_0, Y_0, Z_0]^T + \mathbf{D}[R^{\gamma_R}, G^{\gamma_G}, B^{\gamma_B}]^T), \qquad (10)$$

so that $I_w = I_d$ is impossible in the general case [see the $\mathbf{T}$ and $\mathbf{D}$ matrix values in Eq. (6) and (9)].

4 Human Visual System-Based Image Quality Metrics

4.1 *Components of Image Quality*

Current HVS-based image quality metrics use models of the HVS that take into account a limited part of human vision, thus evaluating the image quality in terms of the selected

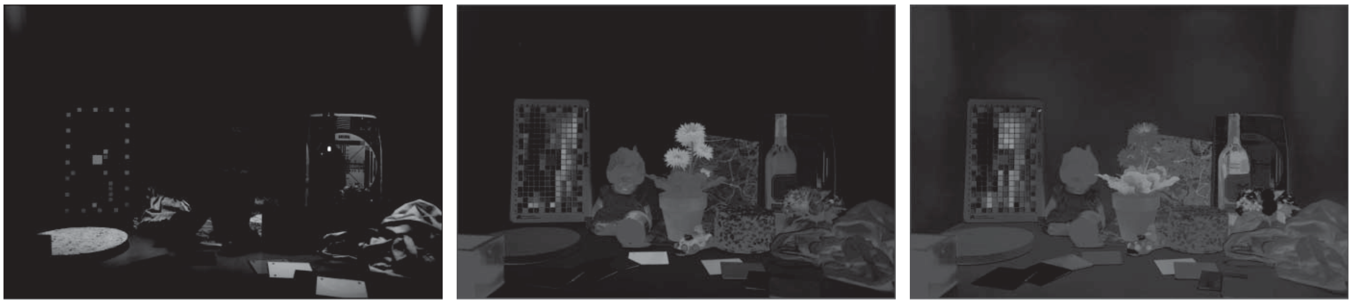


Fig. 7 Luminance (left) and chromaticity (middle and right) estimates of the *Little Lucie* D65 JPEG image displayed in a dark room on a NEC 1701 LCD monitor.

visual process. For instance, the VDP uses a CSF model, a TVI, and a visual masking model, while the CIECAM02 is a color appearance model.

These vision models are validated against psychophysical data. However, the lack of a global HVS model in the current state of the art in vision science leads to a specific problem for HVS-based image quality metrics: one cannot expect that a single HVS model would account for all dimensions of image quality.[63]

We propose to cope with this limit using a multicomponents description of the image fidelity to the actual scene, taking into account three key issues in human vision that are image based (making them relevant for an image quality metric). This framework addresses various aspects of human vision in natural situations, however without any bias due to the task or semantic image content. The last restriction applies, to our sense, to any future HVS-based image quality metric, because whatever these task-dependent biases, they are the same in the "world" reference condition and in the "display" condition.

We used three criteria in subjective and objective aspects of the human visual behavior as a basis for image quality metric selection; however, we do not claim that more criteria should not be added to these. In Secs. 4.2, 4.3, and 4.4, specific models are chosen for each of these components. They can be replaced by more accurate models without changing the framework. The three image quality criteria are as follows.

1. Observers see **the same things** in the reference situation (the real world) and in the displayed situation. That is to say, no significant difference is found when comparing between what observers see/fail to see in images I_w and I_d.
2. The reference and displayed images **look the same** to the observers. That is to say, no significant difference is found between the appearance of I_w and I_d as far as observers can judge.
3. Observers **look at the same things** in the reference situation and in the displayed image. That is to say, no significant difference is found when comparing gaze positions in I_w and I_d.

From these visual behavior components, three indexes are proposed and derived from vision models, and allow us to compute error maps from I_w and I_d (Fig. 8).

1. The visual performance index (VPI) refers to visual performances in a psychophysical sense, such as visibility. This index should take into account, as far as possible, data about the CSF, TVI, visual masking, visual adaptation, mesopic vision, disability glare, etc. Indexes computed from Refs. 5, 8, 14, 16, 24, 26, and 64 are relevant here.
2. The visual appearance index (VAI) refers to human judgments in a psychophysical sense, such as brightness and color. The most popular model of this kind is the CIECAM02[8]; however, other approaches are possible, such as iCAM,[20] brightness rendering,[6] visual equivalence,[18] or combining the many "nesses" of image quality through psychometric scaling.[37]
3. The visual saliency index (VSI) refers to the image-based bottom-up aspects of visual attention. Indexes computed from Refs. 9, 47, and 65 are relevant here.

We give evidence in the following that these three components of the visual behavior are poorly correlated. Each index emphasizes one aspect of image fidelity. To illustrate these differences, we selected three partial models of the HVS to build the indexes. We used Daly's VDP for the visual performance index, CIECAM02 for the visual appearance index, and the saliency map of Ref. 47 for the visual attention index, given that these models are broadly accepted and used in the

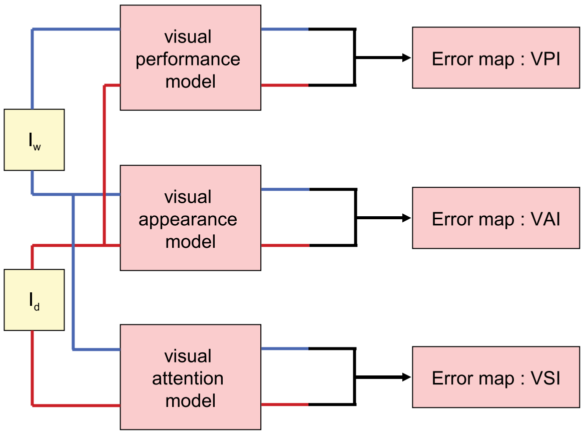


Fig. 8 Framework of the comparison between the "world" image I_w and the displayed image I_d using three quality indexes: a visual performance index (VPI), a visual appearance index (VAI), and a visual saliency index (VSI).

image community and are validated against psychophysical data. We do not mean to rate these models, only to propose an implementation of our theoretical framework.

Global indexes can be computed from the local indexes through a spatial summation, e.g., mean square error, mean absolute difference, or any Minkowski metric. However, we follow Daly's opinion[5] that this spatial summation is a hazardous step.

4.2 *Visual Performance*

We use Daly's VDP as a visual performance index[5] in its HDR version.[13] This model takes as input two luminance images (color information is not processed), and produces a map in which a pixel value is understood as the probability that the difference between the two images at this pixel is visible. The HDR-VDP[13] is a recent evolution of Daly's model, which can be considered the reference benchmark for this class of models. It extends the VDP to high dynamic range images, including high contrast vision psychophysical models to the previous framework.

The computation follows several steps. The first one uses an optical transfer function (OTF) to model the light diffusion in the retina and converted in just noticeable difference (JND) units. This first step models the nonlinearity of the HVS. The second step is a contrast sensibility function (CSF) filter. The cortex transform (modified from Ref. 15) creates a multichannel representation of the image using radial and orientation filters. Psychophysical data modeling visual masking is also taken into account.[31] The last step uses a psychometric function[66] to compute detection probabilities for each subband, and the final probability is computed from the partial detection probabilities.

4.3 *Visual Appearance*

The visual appearance index is computed from the newest generation of the CIECAM models, the CIECAM02.[8] It is a consensus among TC 8-01 expert groups of the CIE, aiming to predict the color appearance. The model needs some inputs: surround luminance (average, dim, or dark), adaptation luminance, and white point in the "world" and "display" conditions. The perceptual attribute correlates are the hue angle *h,* eccentricity factor *e*, hue composition *H,* lightness *J*, brightness *Q,* chroma *C*, and saturation *s,* while *a* and *b* give a Cartesian color representation. From these attributes, Luo, Cui, and Li[67] computed a perceptive color difference:

$$\Delta E' = [(\Delta J')^2 + (\Delta a')^2 + (\Delta b')^2]^{\frac{1}{2}}, \tag{11}$$

where J', a', and b' are derived from J, a, and b. In the following, the visual adaptation and the reference white, which are needed to compute the CIECAM color attributes, are set respectively to the mean color in the image and to the color of the white patch in the chart (when present) or to five times the mean color otherwise (see Ref. 68).

4.4 *Visual Attention*

The visual saliency index is computed using the saliency map in Ref. 47. Like other computational models of visual attention, this algorithm uses RGB inputs combined to compute luminance and opponent colors channels. Thus, we used an analogy between the RGB channels and the physiological LMS channels. The physiological encoding of color in the magno-, parvo-, and koniocellular pathways can be described in terms of M + L for the luminance (magnocellular) channel, L-M for the red-green (parvocellular) channel, and S-(M + L)/2 for the blue-yellow (koniocellular) channel.[69] Thus, we computed the LMS values from the XYZ values using the CIECAM02 matrix:

$$\begin{bmatrix} L \\ M \\ S \end{bmatrix} = \begin{bmatrix} 0.7328 & 0.4296 & -0.1624 \\ -0.7036 & 1.6975 & 0.0061 \\ 0.0030 & 0.0136 & 0.9834 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}, \tag{12}$$

and computed the opponent color channels from these LMS values in the algorithm in Ref. 47.

The saliency map is understood as a probability distribution across space, thus it is normalized to 1 (this normalization was not included in the original algorithm). Then, the VSI is computed as the absolute difference between the two normalized saliency maps.

4.5 *Model Implementation and Units*

In the following, we use online implementations when available: the HDR-VDP is available from the Max Planck Institute website (MPI)[70]; a Saliency Toolbox for Matlab is available online[71,72]; and we implemented the CIECAM02 following Moroney et al.[8] to compute the color attributes in an image; and Luo, Cui, and Li[67] to compute the color difference [Eq. (11)].

The units of these three indexes are heterogeneous: the VPI computes, for each pixel, a value between 0 and 1, meaning a probability of difference detection; while the VAI computes a just noticeable difference (JND) from CIECAM02, with values in $[0,\infty[$; and the VSI computes an absolute difference, also in $[0,\ 1]$. There is no obvious way to compare the units or the scales of these outputs. Thus, building a multicomponent index out of these three indexes should take this heterogeneity into account (see Sec. 5.3). A simpler way to use these indexes is comparing the same index for two conditions.

Coming back to the image pipeline, our image quality evaluation framework compares the "world" image I_w (estimated in Sec. 3.1 from a sensor properties) and the displayed image I_d (estimated in Sec. 3.3 from the display device properties). Both are described in photometric units. Three error maps are computed using standard vision models validated against psychophysical data. We compare I_w and I_d in terms of visual performance, visual appearance, and visual saliency.

5 Results

What is the need for a multicomponent description of image quality? If the VPI, VAI, and VSI indexes would roughly measure the perceptive "error" between the reference image and the displayed image, they should strongly covariate. Moreover, if these errors were correlated, they should be correlated as errors, i.e., the correlation should be positive, and an error near zero for index *i* should correspond to an error near zero for index *j*.

In this section, we give evidence that the index covariation is, at best, small (in the statistical sense of an effect size), meaning that the underlying processes of human vision are only weakly related to each other. Thus, the indexes measure truly different quality components, which emphasizes the need for a multicomponent description of image quality.

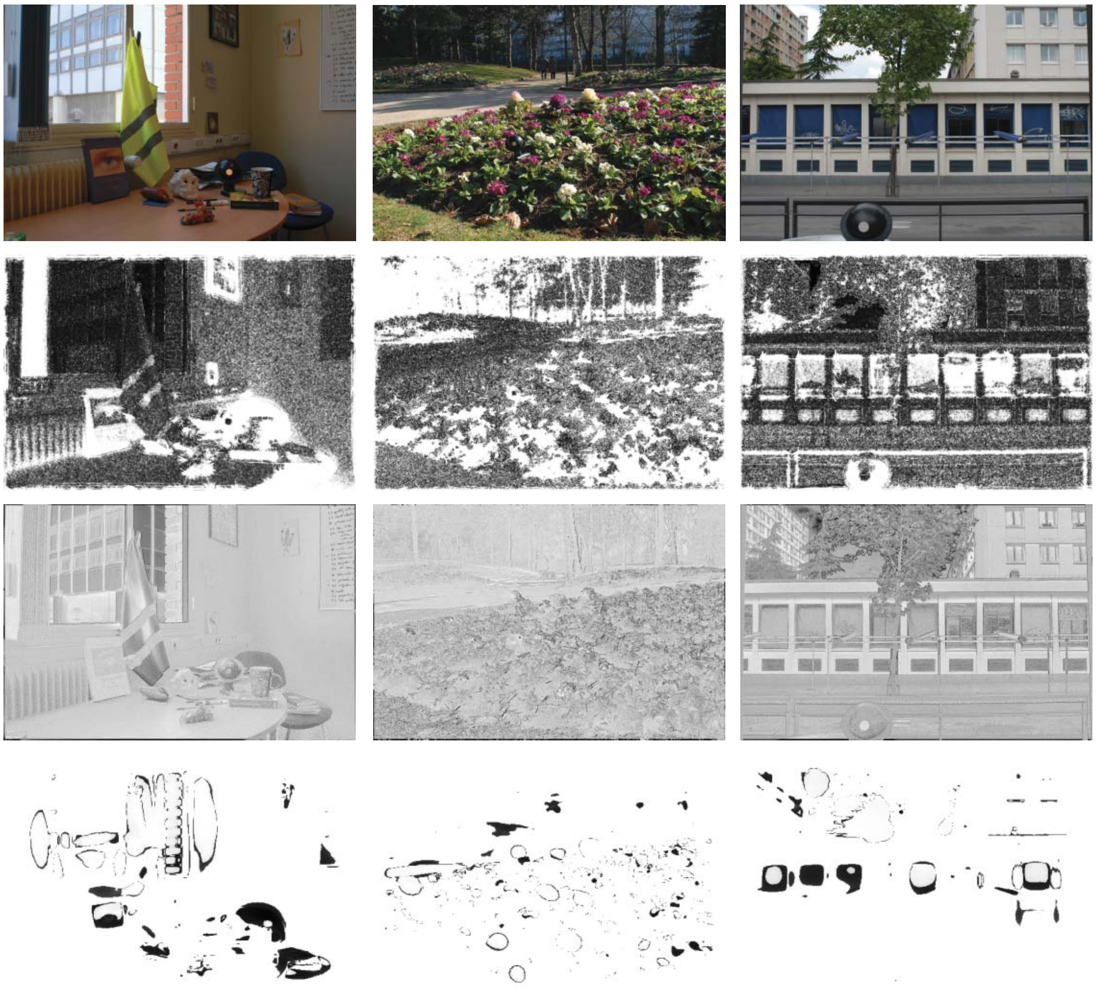


Fig. 9 Normalized error maps computed on sample images from our database, when comparing the reconstructed I_w images and the displayed JPEG I_d images. First raw: JPEG images. Raw 2: VPI error maps. Raw 3: VAI error maps. Raw 4: VSI error maps. Higher errors appear dark.

Moreover, in many cases, improving the image quality according to one quality index lowers the image quality in the sense of another one (significant negative correlations).

5.1 *Image Database*

52 photographs were taken under unknown illumination (both indoor and outdoor), but including a luminance reference (see examples in Fig. 9, first raw). The black integrating sphere in these images includes a constant light (240 *cd/m*^{2}, measured with a video-photometer Minolta CA-S20W). This trick allowed us (the sensor being linear) to compute the true luminance for the light source area, using two photographs for each scene, with the light source "on" and "off." For the CIECAM02, we set the adaptation luminance to the mean luminance, and the white luminance to five times the adaptation luminance.[68] The *Little Lucie* image was added to the image database, which resulted in a set of 53 images.

For each of these 53 images, two images were compared. I_w was built from the raw image and I_d was built from the JPEG image (we used the default compression level, set to 6), displayed on a LCD monitor. Figure 9 gives some examples of the VPI, VAI and VSI error maps.

5.2 *Statistical Analysis*

The Pearson product-moment correlation coefficient r allows us to compare two datasets in search of a possible linear correlation. The Spearmann coefficient p, considering ranks instead of data values, checks for a possible monotonic function explaining the dependence between the variables. Considering two images I_w and I_d, Pearson and Spearman coefficients r and p were computed to assess the correlation between two image quality indexes (VPI versus VAI; VPI versus VSI; or VAI versus VSI) on this image at the pixel level.

p-values were computed to see whether these coefficients are significantly greater than 0 (one-tailed test, with a significance criteria set to $a = 0.02$ in the following). When relevant (positive and significant correlation), an effect size (ES) was computed on the correlation coefficients r and p.[73] We considered that in psychophysical science, which is relevant for vision models, a statistical effect can be rated as *small* when $0.3 < r \leq 0.5$ (that is, the explained variance is between 9% and 25%), *medium* when $0.5 < r \leq 0.707$ (the explained variance is between 25% and 50%), and *large* when $r > 0.707$ (the explained variance is above 50%). We rated a correlation as *not relevant* (NR) when it was significantly lower than 0.30, in a statistical sense (again, the significance

Table 1 Comparison between the image quality indexes on a database of 53 calibrated images. Pearson and Spearman coefficients *r* and *p* are computed, as well as the effect size. S: *small.* M: *medium.* L: *large.* NR: *not relevant.* neg: *negative.* Part 1: indoor images.

	VPI versus VAI				VPI versus VSI				VAI versus VSI			
Image	*r*	ES	*p*	ES	*r*	ES	*p*	ES	*r*	ES	*p*	ES
Little Lucie	0.325	S	0.194	NR	−0.094	Neg	−0.103	Neg	−0.019	Neg	−0.070	Neg
Indoor 1	0.130	NR	0.101	NR	0.083	NR	0.088	NR	0.054	NR	0.050	NR
Indoor 2	0.359	S	0.381	S	0.079	NR	0.242	NR	0.063	NR	0.366	S
Indoor 3	−0.160	Neg	−0.279	Neg	−0.047	Neg	−0.073	Neg	0.029	NR	0.237	NR
Indoor 4	0.153	NR	0.145	NR	0.038	NR	0.083	NR	0.031	NR	0.282	NR
Indoor 5	0.198	NR	0.161	NR	0.046	NR	0.022	NR	0.015	NR	0.015	NR
Indoor 6	0.486	S	0.502	M	0.120	NR	0.370	S	0.096	NR	0.374	S
Indoor 7	0.353	S	0.401	S	0.077	NR	0.239	NR	0.045	NR	0.367	S
Indoor 8	0.066	NR	0.079	NR	−0.095	Neg	−0.186	Neg	0.041	NR	0.251	NR
Indoor 9	0.232	NR	0.171	NR	0.004	NR	0.174	NR	0.014	NR	0.154	NR
Indoor 10	0.443	S	0.551	M	0.085	NR	0.259	NR	0.046	NR	0.346	S
Indoor 11	0.370	S	0.217	NR	0.139	NR	0.103	NR	0.173	NR	0.101	NR
Indoor 12	−0.135	Neg	−0.463	Neg	−0.176	Neg	0.118	NR	0.064	NR	−0.268	Neg
Indoor 13	0.190	NR	0.064	NR	0.105	NR	−0.005	Neg	0.062	NR	0.327	S
Indoor 14	0.261	NR	0.240	NR	0.043	NR	0.308	S	0.041	NR	0.160	NR
Indoor 15	0.411	S	0.461	S	0.082	NR	0.332	S	0.062	NR	0.403	S
Indoor 16	0.296	NR	0.217	NR	0.107	NR	0.152	NR	0.067	NR	0.255	NR
Indoor 17	0.293	NR	0.408	S	−0.040	Neg	0.305	S	−0.024	Neg	0.130	NR
Indoor 18	0.190	NR	0.059	NR	−0.014	Neg	−0.025	Neg	0.017	NR	−0.099	Neg
Indoor 19	0.072	NR	0.072	NR	−0.069	Neg	0.268	NR	−0.032	Neg	−0.013	Neg
Indoor 20	0.363	S	0.460	S	0.036	NR	0.385	S	0.035	NR	0.229	NR
Indoor 21	0.340	S	0.441	S	−0.006	Neg	0.313	S	−0.019	Neg	0.340	S
Indoor 22	0.111	NR	−0.037	Neg	−0.029	Neg	0.221	NR	−0.041	Neg	−0.108	Neg

criteria was set to $a = 0.02$). Note that the proposed approach is optimistic, as two monotonic functions explaining the data from images *A* and *B* have no reason to be identical. However, we show in the following that even this approach does not lead to strong correlations.

The JPEG image size was 1950×1308 pixels. Due to computing power issues in the recursive computation of pixel ranking (for the Spearman coefficient computation), we used a subsampling of the index data (one pixel over eight in horizontal and vertical directions). Thus, the degree of freedom (*DOF*) in the statistical tests was $N - 2 = 39{,}607$, which is why most tests are significant in the following. The true issue of the statistical analysis is the effect size of the correlation (see Tables 1 and 2).

The first step of the statistical analysis was the computation of Pearson correlation coefficients. From the $53\times3 = 159$ *r* values (53 images×3 comparisons: VPI/VAI, VPI/VSI, and VAI/VSI), we found 97 positive *r* correlations (45 for the VPI/VAI comparisons, 21 for the VPI/VSI, and 31 for the VAI/VSI). Due to the high *DOF* value, only five positive *r* values (out of 97) were not significantly >0. From the 92 significant positives, no *medium* or *large* effect size (ES) effect was found, in the statistical sense proposed before, and only 17 *small* ES were found, that is, correlations explaining at least 9% of the total variance (all 17 for VPI/VAI comparisons).

As expected, the Spearman coefficients most of the time improved the correlation coefficients and the effect size:

Table 2 Comparison between the image quality indexes on a database of 53 calibrated images. Part 2: outdoor images.

	VPI versus VAI				VPI versus VSI				VAI versus VSI			
Image	r	ES	ρ	ES	r	ES	ρ	ES	r	ES	ρ	ES
Outdoor 1	−0.020	Neg	−0.039	Neg	−0.115	Neg	−0.145	Neg	−0.006	Neg	0.231	NR
Outdoor 2	0.101	NR	0.087	NR	−0.057	Neg	0.099	NR	0.024	NR	−0.135	Neg
Outdoor 3	−0.079	Neg	−0.139	Neg	0.004	NR	0.243	NR	0.015	NR	−0.260	Neg
Outdoor 4	−0.032	Neg	−0.015	Neg	−0.214	Neg	0.258	NR	−0.001	Neg	0.067	NR
Outdoor 5	0.085	NR	0.081	NR	0.030	NR	−0.104	Neg	0.003	NR	−0.027	Neg
Outdoor 6	0.160	NR	0.169	NR	−0.118	Neg	0.339	S	0.031	NR	0.112	NR
Outdoor 7	0.370	S	0.395	S	0.121	NR	0.227	NR	0.049	NR	0.480	S
Outdoor 8	0.300	S	0.432	S	−0.076	Neg	0.250	NR	−0.041	Neg	0.254	NR
Outdoor 9	0.089	NR	−0.076	Neg	−0.011	Neg	0.019	NR	−0.022	Neg	−0.051	Neg
Outdoor 10	0.146	NR	0.069	NR	−0.084	Neg	−0.058	Neg	−0.027	Neg	0.133	NR
Outdoor 11	0.071	NR	0.114	NR	−0.061	Neg	−0.042	Neg	0.003	NR	0.037	NR
Outdoor 12	0.150	NR	0.160	NR	0.041	NR	−0.003	Neg	−0.018	Neg	0.024	NR
Outdoor 13	0.153	NR	0.169	NR	−0.011	Neg	0.021	NR	0.055	NR	0.053	NR
Outdoor 14	0.083	NR	0.053	NR	−0.013	Neg	−0.057	Neg	0.009	NR	0.069	NR
Outdoor 15	0.320	S	0.312	S	0.086	NR	0.044	NR	0.109	NR	0.267	NR
Outdoor 16	0.128	NR	0.079	NR	−0.014	Neg	0.080	NR	−0.023	Neg	−0.166	Neg
Outdoor 17	0.203	NR	0.118	NR	−0.013	Neg	0.077	NR	−0.003	Neg	−0.040	Neg
Outdoor 18	0.388	S	0.469	S	−0.047	Neg	0.327	S	−0.017	Neg	0.227	NR
Outdoor 19	0.353	S	0.353	S	−0.188	Neg	−0.075	Neg	−0.066	Neg	−0.146	Neg
Outdoor 20	0.303	S	0.397	S	−0.009	Neg	0.124	NR	−0.001	Neg	0.229	NR
Outdoor 21	−0.055	Neg	−0.250	Neg	−0.124	Neg	0.227	NR	−0.004	Neg	−0.202	Neg
Outdoor 22	0.405	S	0.450	S	−0.014	Neg	0.410	S	0.047	NR	0.366	S
Outdoor 23	0.197	NR	0.180	NR	−0.043	Neg	0.113	NR	0.020	NR	0.108	NR
Outdoor 24	0.109	NR	0.107	NR	−0.059	Neg	−0.165	Neg	0.015	NR	−0.287	Neg
Outdoor 25	0.235	NR	0.288	NR	0.068	NR	0.098	NR	−0.024	Neg	−0.005	Neg
Outdoor 26	−0.021	Neg	−0.125	Neg	−0.089	Neg	0.222	NR	−0.063	Neg	−0.186	Neg
Outdoor 27	0.417	S	0.492	S	−0.023	Neg	0.058	NR	−0.016	Neg	0.092	NR
Outdoor 28	−0.001	Neg	−0.073	Neg	−0.038	Neg	−0.042	Neg	−0.032	Neg	−0.008	Neg
Outdoor 29	0.077	NR	−0.006	Neg	−0.076	Neg	0.063	NR	−0.018	Neg	−0.002	Neg
Outdoor 30	0.244	NR	0.223	NR	0.050	NR	0.180	NR	0.011	NR	0.172	NR

116 over 159 positive correlations were found, and all were significantly > 0. Still, no *large* effect (i.e., explaining more than 50% of the variance) was found, and only two *medium* effects and 14 *small* effects were found for the VPI/VAI comparisons. Meanwhile, nine *small* effects were found for both the VPI/VSI and the VAI/VSI comparisons (no *medium* or *large* effect was found for these comparisons). Note that a 100% correlation is expected for p when an increasing

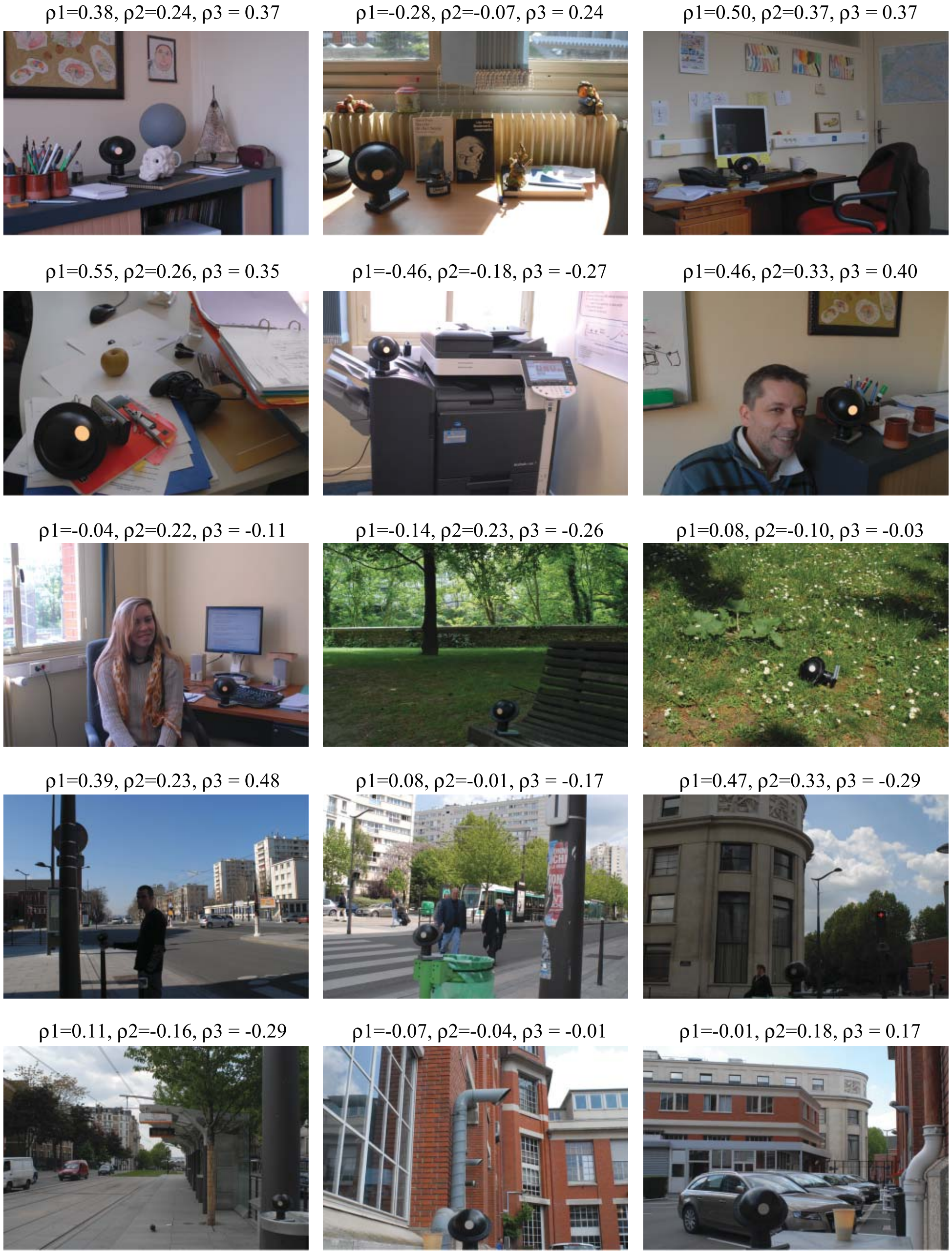

ρ1=0.38, ρ2=0.24, ρ3 = 0.37
ρ1=-0.28, ρ2=-0.07, ρ3 = 0.24
ρ1=0.50, ρ2=0.37, ρ3 = 0.37
ρ1=0.55, ρ2=0.26, ρ3 = 0.35
ρ1=-0.46, ρ2=-0.18, ρ3 = -0.27
ρ1=0.46, ρ2=0.33, ρ3 = 0.40
ρ1=-0.04, ρ2=0.22, ρ3 = -0.11
ρ1=-0.14, ρ2=0.23, ρ3 = -0.26
ρ1=0.08, ρ2=-0.10, ρ3 = -0.03
ρ1=0.39, ρ2=0.23, ρ3 = 0.48
ρ1=0.08, ρ2=-0.01, ρ3 = -0.17
ρ1=0.47, ρ2=0.33, ρ3 = -0.29
ρ1=0.11, ρ2=-0.16, ρ3 = -0.29
ρ1=-0.07, ρ2=-0.04, ρ3 = -0.01
ρ1=-0.01, ρ2=0.18, ρ3 = 0.17

Fig. 10 Spearman correlation coefficient values when comparing the image quality indexes VPI and VAI (ρ_1), VPI and VSI (ρ_2), VAI and VSI (ρ_3).

function maps an error index dataset to the other on a given image.

Negative correlations show that in some cases the image quality in the sense of one index tends to increase when the image quality, in the sense of another index, decreases. This should not happen if the image quality indexes would rate a general purpose image quality. Moreover, the number of *negative* correlations in our database was high. Considering r, we found eight negative values (out of 53) for VPI/VAI, 32 for the VPI/VSI (more than half the total number of images), and 22 for the VAI/VSI comparisons. For p, we found 11 significant negative correlations for VPI/VAI, 14 for VPI/VSI, and 18 for VAI/VSI. Figure 10 shows images from the database with various values (both positive and negative) for the indexes correlations.

A specific pattern suggests that the between-index correlations are not very consistent. In some cases, we found that $r\rho < 0$, which means that the best linear fit on the index data has a positive slope, while the best linear fit on the rank data has a negative slope (or the reverse). This situation is anecdotic for the VPI/VAI comparisons (it happens three times on 53 image comparisons), but becomes more frequent for the VPI/VSI comparisons (26 times) and for the VAI/VSI comparisons (16 times).

Another pattern emerges when comparing the three image quality indexes: while the correlations between VPI/VSI and

VAI/VSI are very weak (only nine *small* effects over 53 for both image comparisons for the Spearman coefficient), the correlation between VPI and VAI is somehow more consistent (16 over 53 images result in a *small* or *medium* effect size). Our understanding is that these indexes are mainly based on local variations, and uniform areas lead to very small errors both in the VDP framework and CIECAM framework. Conversely, the saliency map tends to select low frequency information, relevant for peripheral vision, and is thus less sensible to high frequency properties such as local uniformity.

Some limits of the previous statistical analysis should be emphasized. First, the analysis was restricted to the image quality indexes correlations, and some expected properties of the indexes were not checked. For instance, a direct computation can show whether or not low error values in the sense of one index (say, the ten first percentiles) occur at pixels with low error values for the other two indexes. Due to the large number of significant negative correlations, we did not feel the need for such a detailed analysis. Another issue is the size of the image database. One may consider that a database of 53 images is quite small, and we agree to some extend. However, we were limited by the fact that a public domain image database was not available with calibrated raw versus JPEG images. Thus, we had to build the database by ourselves with a complex protocol, including two photographs with a reference luminance in the field of view (see Sec. 5.1). Finally, photographs taken in various situations showed consistent results with no significant positive *large* and only two *medium* ES over the 318 computed correlations (r and ρ for the three kinds of index comparisons). This quantitative result suggests that even if a larger database of calibrated raw images would improve the present results, the limited size of our database was fair enough for the limited purpose of this work.

The statistical analysis shows that the VPI, VAI, and VSI are only weakly correlated, either using a linear or a nonlinear model. This makes sense in terms of visual behavior: the corresponding vision models do not consider the same aspect of vision, and our results show that the three selected components of the visual behavior are almost independent. This is more than an intuitive result: the data suggest that these three components of the visual behavior derive from distinct components of the human visual system.

5.3 *Multicomponent Index*

The fact that negative correlations can occur between the three image quality indexes makes it critical for the user to check whether improving one aspect of image quality does decrease image quality as a whole. Therefore, we propose a new index for image quality assessment based on the previous three.[8,13,47] To get the same range for the three indexes, we have mapped the VAI into [0,1]:

$$Q = \lambda_1 \text{VPI} + \lambda_2 [1 - \exp(\text{VAI})] + \lambda_3 \text{VSI}, \qquad (13)$$

where (λ_1, λ_2, and λ_3) are user-defined parameters that depend on the application.

6 Conclusion

We propose a framework to assess the quality of the image acquisition, processing, and display pipeline with respect to the visual behavior of observers. We use three components: visual performance is concerned with what is visible in the images, that is, with visibility thresholds; visual appearance is concerned with subjective judgments about the images; and visual attention is concerned with where people look in the images. The computation of the error maps uses a photometric estimation of two scenes: a reference scene I_w scanned by a sensor, and a displayed scene I_d. The data show very weak correlations between the three image quality indexes, in terms of size effect, calling for a multicomponent description of image quality. One such index is proposed, including user-defined parameters.

The proposed image pipeline evaluation depends on the reference image (input stimulus) that is chosen for the acquisition step. As the evaluation process depends on the input data, it cannot be said, strictly speaking, that we assess the image pipeline, but rather the pipeline for a given input, which is not suited for most practical applications. To extend our approach to a system diagnosis independent of the input data, one can use a reference scene database (rather than an image database), and perform some kind of a benchmark on this database, e.g., with mean indexes over all the reference scenes. Therefore, the development of calibrated raw image databases should be encouraged.

Some recommendations emerge. First, a universal image quality metric seems beyond the range of current knowledge, and possibly unavailable because of the various (and weakly correlated) components of the human visual behavior. Although specific applications of image quality assessment may select a vision model as being more relevant, it may help to check if, for instance, an image processing tuned with respect to this vision model (e.g., visual performance) leads, or not, to drawbacks for alternative quality indexes (visual appearance and visual attention). This was the rationale for the proposed multicomponent index.

Second, image quality addresses the fidelity between a displayed image and a physical scene. Assessing one step of the pipeline, as is usually done in image coding evaluation, can lose an important fidelity issue. We suggest to take into account, when possible, the sensor and display characteristics in image fidelity assessment.

Although further developments are necessary to include dynamic visual components to our framework,[49,74] our approach can be useful in the field of TV, movies, and video games, as well as in vision research. In our opinion, the main industrial application is the calibration of a display system (LCD, plasma, video-projection, printer, etc.) with the knowledge of the image acquisition system properties, allowing us to choose the best possible settings in a perceptive sense.

Specific applications can be proposed, such as image quality assessment on a specific criterion, while checking the consequences on other quality indexes; easy calibration settings for display devices, for a better rendering quality in terms of visual perception; better design of image processing, such as codecs and TMOs by using the knowledge of the acquisition and display device properties; and of course, direct assessment of the quality of a given image pipeline in terms of visual perception.

Acknowledgments

Thanks to Lucie Brémond for the baby doll on the picture, Vincent Ledoux for the integrating sphere idea, Roger Hubert and Giselle Paulmier for the photometric and spectrometric measurements, Josselin Petit for the computations using the saliency toolbox in Ref. 47, and Henri Panjo and Ariane Tom for their help on the statistical analysis.

References

1. S. Nayar and T. Mitsunaga, *Proc. IEEE Conf. Computer Vision Patt. Recog.* **1**, 472–479 (2000).
2. H. Seetzen, W. Heidrich, W. Stuerzlinger, G. Ward, L. Whitehead, M. Trentacoste, A. Ghosh, and A. Vorozcovs, *ACM Trans. Graphics* **23**, 760 (2004).
3. T. Newman, *Proc. 24th CIE*, pp. 5–10 (1999).
4. M. P. Eckert and A. P. Bradley, *Signal Process.* **70**, 177 (1998).
5. S. Daly, "The visible differences predictor: an algorithm for the assessment of image fidelity," in *Digital Images and Human Vision*, A. B. Watson, Ed., pp. 179–206, MIT Press, Cambridge, MA (1993).
6. J. Tumblin and H. Rushmeier, *IEEE Computer Graphics Appl.* **13**, 42 (1993).
7. M. D. Fairchild, *Color Appearance Models*, 2d ed., John Wiley and Sons, New York (2005).
8. N. Moroney, M. D. Fairchild, R. W. G. Hunt, C. J. Li, M. R. Luo, and T. Newman, *Proc. IS&T-SID 10th Color Imag. Conf.*, pp. 23–27 (2002).
9. L. Itti and C. Koch, *Vision Res.* **40**, 1489 (2000).
10. D. P. Greenberg, K. E. Torrance, P. Shirley, J. Arvo, J. A. Ferwerda, S. N. Pattanaik, E. Lafortune, B. Walter, S. C. Foo, and B. Trumbore, *Proc. ACM SIGGRAPH*, pp. 477–494 (1997).
11. P. G. Engeldrum, *J. Imaging Sci. Technol.* **48**, 446 (2004).
12. J. L. Mannos and D. J. Sakrison, *IEEE Trans. Info. Theory* **IT-4**, 525 (1974).
13. R. Mantiuk, S. Daly, K. Myszkowski, and H. P. Seidel, "Predicting visible differences in high dynamic range images: model and its calibration," *Proc. SPIE* **5666**, 204–214 (2005).
14. J. Lubin, "A visual discrimination model for imaging system design and development," in *Vision Models for Target Detection and Recognition*, E. Peli, Ed., pp. 245–283, World Scientific, Singapore (1995).
15. A. B. Watson, *Computer Vision Image Process.* **39**, 311 (1987).
16. Z. Wang, A. Bovik, H. Sheik, and E. Simoncelli, *IEEE Trans. Image Process.* **13**, 600 (2004).
17. J. A. Ferwerda and D. Pellacini, in *Asilomar Conference on Signal, Systems and Computers*, pp. 1388–1392 (2003).
18. G. Ramanarayanan, J. Ferwerda, B. Walter, and K. Bala, *ACM Trans. Graphics* **26**, Art. No. 76 (2007).
19. CIE/ISO, *Tech. Rep.*, CIE publication 131 (1998).
20. J. Kuang, G. M. Johnson, and M. D. Fairchild, *J. Visual Communi. Image Represent.* **18**, 406 (2007).
21. F. Drago, K. Myszkowski, T. Annen, and N. Chiba, *Proc. Eurograph.* (2003).
22. K. Masaoka, M. Emoto, M. Sugawara, and Y. Nojiri, "Comparing realness between real objects and images at various resolutions," *Proc. SPIE* **6492**, 1F (2007).
23. J. Kuang, H. Yamaguchi, C. Liu, G. M. Johnson, and M. D. Fairchild, *ACM Trans. Appl. Perception* **4**, art. 9 (2007).
24. M. Ramasubramanian, S. N. Pattanaik, and D. P. Greenberg, *Proc. ACM SIGGRAPH*, pp. 73–82 (1999).
25. R. Dumont, F. Pellacini, and J. A. Ferwerda, *ACM Trans. Graphics* **22**, 152 (2003).
26. S. N. Pattanaik, J. A. Ferwerda, M. D. Fairchild, and D. P. Greenberg, in *Proc. ACM SIGGRAPH*, pp. 287–298 (1998).
27. J. Stevens and S. Stevens, *J. Opt. Soc. Am.* **53**, 375 (1963).
28. G. Ward, "A contrast based scale-factor for image display," in *Graphic Gems IV*, pp. 415–421, Academic Press Professional, San Diego, CA (1994).
29. CIE, Tech. Report, CIE publication 19/2 (1981).
30. J. A. Ferwerda and S. N. Pattanaik, *Proc. ACM SIGGRAPH*, pp. 143–152 (1997).
31. G. E. Legge and J. M. Foley, *J. Opt. Soc. Am.* **70**, 1458 (1980).
32. F. Campbell and J. Robson, *J. Physiol.* **197**, 551 (1968).
33. E. Peli, *J. Opt. Soc. Am. A* **7**, 2032 (1990).
34. CCIR, "Method for the subjective assessment of the quality of television pictures," *Recommendation* 500–3, ITU, Geneva (1986).
35. H. A. David, *The Method of Paired Comparison*, Charles Griffin and Co. Ltd (1969).
36. M. P. Eckert, "Lossy compression using wavelets, block DCT, and lapped orthogonal transforms optimized with perceptual model," *Proc. SPIE* **3031**, 339–351 (1995).
37. P. G. Engeldrum, *Psychometric Scaling: a Toolkit for Imaging Systems Development*, Imcotek Press, Winchester, MA (2000).
38. A. McNamara, *Computer Graphics Forum* **20**, 211 (2001).
39. G. W. Meyer, H. E. Rushmeier, M. F. Cohen, D. P. Greenberg, and K. E. Torrance, *ACM Trans. Graphics* **5**, 30 (1986).
40. P. Ledda, A. Chalmers, T. Troscianko, and H. Seetzen, *Proc. ACM SIGGRAPH*, pp. 640–648 (2005).
41. H. Rushmeier, G. Ward, C. Piatko, P. Sanders, and B. Rust, *Proc. Eurograph. Rendering Workshop* (1995).
42. J. Grave and R. Brémond, *ACM Trans. Appl. Perception* **5**, 1 (2008).
43. *Neurobiology of Attention*, L. Itti, G. Rees, and J. K. Tsotsos, Eds., Elsevier, New York (2005).
44. E. I. Knudsen, *Ann. Rev. Neurosci.* **30**, 57 (2007).
45. C. Koch and S. Ullman, *Human Neurobiol.* **4**, 219 (1985).
46. A. M. Treisman and G. Gelade, *Cognitive Psychol.* **12**, 97 (1980).
47. L. Itti, C. Koch, and E. Niebur, *IEEE Trans. Patt. Anal. Mach. Intell.* **20**, 1254 (1998).
48. L. Itti, *IEEE Trans. Image Process* **13**, 1304 (2004).
49. L. Itti, N. Dhavale, and F. Pighin, *Proc. SPIE* **5200**, 64–78 (2003).
50. L. Itti and C. Koch, *Nature Rev. Neurosci.* **2**, 194 (2001).
51. C. H. Lee, A. Varshney, and D. W. Jacobs, *ACM Trans. Graphics* **24**, 659 (2005).
52. C. Privitera and L. Stark, "Focused JPEG encoding based upon automatic preidentified regions of interest," *Proc. SPIE* **3644**, 552–558 (1999).
53. S. Baluja and D. A. Pomerleau, *Robotics Auto. Syst.* **22**, 329 (1997).
54. ISO, "Graphic technology and photography. Colour characterization of digital still cameras (DSC)," ISO/WD 17321 (2006).
55. F. Martinez-Verdu, J. Pujol, and P. Capilla, *J. Imag. Sci. Technol.* **47**, 279 (2003).
56. S. Mann and R. Picard, *Proc. IST 48th Ann. Conf.*, pp. 422–428 (1995).
57. T. Mitsunaga and S. K. Nayar, *Proc. IEEE Conf. Computer Vision Patt. Recog.* **1**, 374–380 (1999).
58. A. S. Glassner, *IEEE Computer Graphics Appl.* **9**, 95 (1989).
59. V. Heikkinen, R. Lenz, T. Jetsu, J. Parkkinen, M. Hauta-Kasari, and T. Jskelinen, *J. Opt. Soc. Am. A* **25**, 2444 (2008).
60. B. A. Wandell, Tech. Report 86844, NASA Ames Research Center, Moffett Field, CA (1985).
61. CIE, Tech. Report, CIE publication 122 (1996).
62. E. A. Day, L. A. Taplin, and R. S. Berns, *Color Res. Appl.* **29**, 365 (2004).
63. D. A. Silverstein and J. E. Farrell, *IEEE Proc. Intl. Conf. Image Process.*, pp. 881–884 (1996).
64. J. A. Ferwerda, S. N. Pattanaik, P. Shirley, and D. P. Greenberg, *Computer Graphics* **30**, 249 (1996).
65. J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo, *Artif. Intell.* **78**, 507 (1995).
66. J. Nachmias, *Vision Res.* **21**, 215 (1981).
67. M. R. Luo, G. Cui, and C. Li, *Color Res. Appl.* **31**, 320 (2006).
68. R. W. G. Hunt, *The Reproduction of Colour,* 5th ed., Fountain Press Ltd., Kingston Upon Thames, UK (1995).
69. K. R. Gegenfurtner and D. C. Kiper, *Ann. Rev. Neurosci.* **26**, 181 (2003).
70. MPI, see http://www.mpi-inf.mpg.de/resources/hdr/vdp/, Max Planck Institute website.
71. D. Walther, see http://www.saliencytoolbox.net/ (2006).
72. D. Walther and C. Koch, *Neural Net.* **19**, 1395 (2006).
73. J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, Academic Press, San Diego, CA (1977).
74. S. Winkler, *Digital Video Quality: Vision Models and Metrics*, Wiley, New York (2005).

Biographies and photographs of the authors not available.