ELSEVIER

Signal Processing 70 (1998) 155–176

SIGNAL PROCESSING

Image dissimilarity

Jean-Bernard Martens*, Lydia Meesters[1]

IPO, Center for Research on User-System Interaction, P.O. Box 513, 5600 MB Eindhoven, The Netherlands

Received 30 July 1998

Abstract

In this paper we compare the performance of a number of representative instrumental models for image dissimilarity with respect to their ability to predict both image dissimilarity and image quality, as perceived by human subjects. Two sets of experimental data, one for images degraded by noise and blur, and one for JPEG-coded images, are used in the comparison. © 1998 Elsevier Science B.V. All rights reserved.

Zusammenfassung

In dieser Arbeit vergleichen wir das Verhalten einer Anzahl repräsentativer einsetzbarer Modelle für die Unterschiedlichkeit von Bildern in bezug auf ihre Fähigkeit sowohl Bildunterschiede und Bildqualität vorherzusagen, so wie menschliche Beobachter sie empfinden würden. Zwei Sätze experimenteller Daten, einer für verrauschte und unscharfe Bilder und einer für JPEG-kodierte Bilder, werden im Vergleich verwendet. © 1998 Elsevier Science B.V. All rights reserved.

Résumé

Nous comparons dans cet article les performances d'un certain nombre de modèles instrumentaux pour la dissimilarité d'image vis-à-vis de leur capacité à prédire à la fois la dissimilarité d'image et la qualité d'image comme perçues par des sujets humains. Deux ensembles de données expérimentales, l'un d'images dégradées par du bruit et rendues floues, et l'autre d'images codées par JPEG sont utilisés à des fins de comparaison. © 1998 Elsevier Science B.V. All rights reserved.

Keywords: Image dissimilarity; Image quality metric; Instrumental measures; Objective image quality; Subjective image quality; Vision models

1. Introduction

Reliable experimental techniques are available for measuring the influence of imaging system parameters on the perceived quality of images [11]. In most cases, such an experimental evaluation is only feasible in the design phase of a system, and even then only with a limited number of images and subjects. Therefore, instrumental measures that correlate well with experimental data on image

* Corresponding author. Tel.: 31 40 2475208 or 73; fax: 31 40 2431930; e-mail: jbm@ipo.tue.nl.

[1] Supported by European ACTS AC055 project TAPESTRIES.

PII: S0165-1684(98)00123-6

quality are needed once systems become operational or when systems have to be tested on a large set of images. A proliferation of computational image quality metrics has been witnessed, especially over the past few years [1]. At first, most metrics were limited to monochromatic still images [6,7,16,17,32]. More recently, extensions towards color and image sequences are being proposed [9,15,30,31]. Furthermore, simplified models are also proposed, since the complexity of some models makes them unsuited for most applications (especially applications where quality has to be monitored in real time).

Little has been done to compare the performance of the different models. It remains unclear how well models perform, as well as which model components are mostly responsible for this performance. An extensive comparison of all proposed models is not straightforward. First, this requires implementing all these models, which is not always feasible since specific algorithmic details needed for the implementation are sometimes not available. Second, the models have to be tested on a large database of original and distorted images. For these images, both the test results and the experimental settings (monitor characteristics, viewing distance, experimental procedure, etc.) have to be documented. These experimental results must obviously include ratings of image quality. However, in order to get more insight into how overall quality is influenced by different distortions, this data base may also have to include information about image dissimilarity and/or image quality attributes such as noisiness, perceived blur, blockiness, etc. Such databases are simply not (publicly) available.

One of the more extensive and well-documented experimental data sets that is available at our institute concerns images degraded by noise and blur [12]. Although the explicit goal of most image quality metrics is to predict the visual effect of image-dependent distortions, such as the ones occurring in image coding, it is usually implicitly assumed that they perform at least as well as more traditional measures, such as root-mean-squared error, in case of image-independent distortions (such as noise and blur). One of the goals of this paper is to examine if this assumption can be supported by experimental evidence.

Another data set that we have recently collected concerns images coded by a baseline JPEG-coder [21]. As in the case of the above data set, not only overall image quality ratings were collected, but also ratings on dissimilarity and blockiness (an important image quality attribute) [19]. The dissimilarity scores were specifically intended to be used in this study.

Although the concept of image dissimilarity is very familiar in the context of instrumental measures for image quality, it is fairly uncommon to use it as an experimental paradigm. Most instrumental measures relate image quality to some distance (such as the root-mean-squared error) between the original and the processed image, such that image dissimilarity arises naturally in this context. In a number of recent papers [8,12], it was demonstrated that dissimilarity can also be judged consistently by subjects. Instead of judging quality (differences), subjects were asked to indicate how dissimilar or different they perceived two images to be, and not to base their score on any preference, quality or emotional criteria. Although judging dissimilarities seems more complicated than judging quality differences, it is often experienced as being easier by subjects, most likely because it does not involve value judgements.

2. Perceived dissimilarity and image quality

In a number of recent papers, we have argued for a multidimensional approach towards analyzing and modelling image quality variations within a scene due to variations in imaging system parameters [8,12,13]. This approach recognizes the fact that image quality is often determined by several underlying attributes (such as noise and blur) and uses a multidimensional geometric model to describe the mutual relationships between different perceptual attributes, as well as the relationships between these attributes and overall image quality. In this geometric model, both the original and the degraded/processed versions of an image are represented as positions in a multidimensional space. The dimensionality of this space is determined by the number of *independently varying* perceptual attributes. For instance, in the case where noise and

blur are varied independently, a dimensionality of two is assumed. A number of techniques, usually referred to as multidimensional scaling (MDS) techniques [14,22], have been developed within the field of mathematical psychology to position stimuli (such as the original and processed images) based on experimental data.

The basic assumption of MDS is that all relevant image properties such as overall image quality, image dissimilarity and the strengths of perceptual attributes correlate highly with geometrical properties of the stimulus positions. For instance, distances between the image positions are assumed to be monotonically related to the perceived dissimilarity between the corresponding images. Furthermore, the strength of perceived attributes, as well as overall image quality, are assumed to correlate well with coordinates along different directions in this space and/or with the distance from an 'ideal'-image point.

The mapping from dissimilarity data into stimulus configurations will be used for both experimental data and instrumental measures. We therefore briefly summarize how this mapping is performed. Let us denote by d_{rij} the dissimilarity between stimuli $i,j = 1,\ldots,I$ $(j < i)$ as indicated by subject (or model) $r = 1,\ldots,R$. The fact that the subject/model responses d_{rij} must be monotonically, but not necessarily linearly, related to the sensations of dissimilarity $d^{\star}_{rij}$ is modelled by the power-law relationship

$$d^{\star}_{rij} = a_r \cdot d^{p_r}_{rij}, \tag{1}$$

with exponent p_r, for $r = 1,\ldots,R$. This power-law relationship is quite flexible and supported by many psychophysical studies [26]. The model underlying the MULTISCALE estimation program [22] assumes that the difference between the transformed dissimilarities $d^{\star}_{rij}$ and the Euclidean distance

$$\hat{d}_{ij} = \sum_{m=1}^{M} (x_{im} - x_{jm})^2 \tag{2}$$

between the stimuli i and j in an M-dimensional space belongs to a log-normal error distribution, i.e.,

$$e_{rij} = \log d^{\star}_{rij} - \log \hat{d}_{ij}, \tag{3}$$

is assumed to have a zero-mean normal distribution with standard deviation σ_r. An extensive argumentation in favour of this log-normal error distribution model is given in the original paper by Ramsay [22].

The maximum-likelihood estimation of all parameters (σ_r, a_r, p_r and the positions of all stimuli in an M-dimensional Euclidean space) involves the maximization of the log-likelihood function

$$\log L = -\frac{1}{2} \sum_{r=1}^{R} \left(\frac{S_r}{\sigma_r^2} + D_r \log \sigma_r^2 \right) + C, \tag{4}$$

where C combines all terms that are independent of the parameters to be optimized, D_r is the number of measured dissimilarities for subject/model r and the sum

$$S_r = \sum_{i,j}^{D_r} e^2_{rij} \tag{5}$$

is over all measured dissimilarities for subject/model r. We refer to the original papers by Ramsay for more details on how this maximization can be performed [22]. One of the advantages of a maximum-likelihood estimation is that we cannot only estimate the positions of the stimuli, but also confidence regions for these stimulus positions [23]. The (asymptotic) 95% confidence regions will also be included in some stimulus configurations further on in the paper.

The stimulus positions can be arbitrarily translated and rotated without influencing the distances between the stimuli. Moreover, scaling all coordinates (and hence all distances) by the same factor s does not influence the dissimilarity predictions either, because this can be counteracted by an increase in the proportionality factor a_r of the power-law relationship between original and transformed dissimilarities. If we collect the stimulus positions into an $I \times M$ matrix X, then the above remarks can be mathematically summarized by stating that the transformed stimulus configuration

$$Y = X \cdot (s\,U) + T, \tag{6}$$

is an equally valid stimulus configuration to describe the experimental data, i.e., it will result in the same value for the log-likelihood function. In this formula, s is a scalar factor, U is a unitary matrix (i.e., $U^{\mathrm{T}} \cdot U$

is equal to the identity matrix) and T is a translation matrix (with all identical rows specifying the translation vector). The MULTISCALE program adds extra conditions to the above log-likelihood maximization in order to guarantee a unique solution for the stimulus position matrix X [23].

When comparing two stimulus configurations X and Y, for instance arising from experimental data and model predictions, then the transformation parameters for one of the configurations, say X, must be optimized before determining the distance from and/or correlation with the second configuration Y. The optimum translation can be easily determined by requiring that the column averages after transformation become equal. If the translation matrices T_x and T_y are needed to make the column averages equal to zero for X and Y, respectively, then the optimum overall translation matrix is $T = T_x \cdot s\,U - T_y$.

The scaling factor s and the transformation matrix U for the optimum transformation from $\tilde{X} = X + T_x$ to $\tilde{Y} = Y + T_y$ are determined by minimizing the sum of the squared distances between the stimulus positions, i.e.,

$$d^2 = \text{trace}[(\tilde{X} \cdot s\,U - \tilde{Y})^{\mathrm{T}}(\tilde{X} \cdot s\,U - \tilde{Y})]. \tag{7}$$

Substituting the optimum scaling factor

$$s = \frac{\text{trace}(\tilde{Y}^{\mathrm{T}}\,\tilde{X}\,U)}{\text{trace}(U^{\mathrm{T}}\,\tilde{X}^{\mathrm{T}}\,\tilde{X}\,U)} \tag{8}$$

into this expression results in

$$d^2 = \text{trace}(\tilde{Y}^{\mathrm{T}}\,\tilde{Y})(1 - \rho^2), \tag{9}$$

so that minimizing d^2 is equivalent to maximizing the inner-product correlation

$$\rho = \frac{\text{trace}\ [\tilde{Y}^{\mathrm{T}}\tilde{X}U]}{\{\text{trace}\ [U^{\mathrm{T}}\tilde{X}^{\mathrm{T}}\tilde{X}U] \cdot \text{trace}\ [\tilde{Y}^{\mathrm{T}}\tilde{Y}]\}^{1/2}}. \tag{10}$$

Maximizing this inner-product correlation is in turn equivalent to maximizing the inner product $\text{trace}[\tilde{Y}^{\mathrm{T}}\tilde{X}U]$, since the denominator in the latter expression is constant if U is a unitary transformation. The solution to the inner-product maximization problem is well-known [24,27]. More precisely, the optimum (so-called orthogonal Procrustes) transformation is given by

$$U = P\,Q^{\mathrm{T}}, \tag{11}$$

where $\tilde{X}^{\mathrm{T}}\tilde{Y} = PSQ^{\mathrm{T}}$ is the singular-value decomposition of $\tilde{X}^{\mathrm{T}}\tilde{Y}$, and S is a diagonal matrix of singular values.

Sometimes we may also wish to consider the case where the matrix transformation between two stimulus configurations need not be unitary, but can be an arbitrary linear transformation. The optimum linear transformation matrix between $\tilde{X}$ and $\tilde{Y}$ in this general (unconstrained) case is

$$U = (\tilde{X}^{\mathrm{T}}\tilde{X})^{-1} \cdot \tilde{X}^{\mathrm{T}}\tilde{Y}. \tag{12}$$

The elements of this matrix are equal to the multiple regression coefficients between the columns of $\tilde{X}$ and the columns of $\tilde{Y}$ [24].

Once the stimulus positions have been determined, then perceived quality (and attribute) judgements can, for instance, be correlated with directions in this space. This means that the perceived quality judgements q_{ri} for stimulus i by subject r are modeled by

$$\hat{q}_{ri} = c_{r0} + \sum_{m=1}^{M} c_{rm} \cdot x_{im}, \tag{13}$$

for $r = 1, \ldots, R$ and $i = 1, \ldots, I$, where the vector $(c_{r1}, \ldots, c_{rM})$ is orthogonal to the lines of equal quality for subject r.

2.1. Images with noise and blur

The MDS approach to image quality has been described in detail in [12] for images degraded by noise and blur. We briefly summarize the experimental results of this study which constitute our first data set.

The three scenes that were used in the experiment (two natural scenes, Wanda and Terrace, and one synthetic scene, Mondrian) are reproduced in the original paper. All 16 combinations of four levels of blur (corresponding to no filtering and filtering with binomial filters of length 3, 5 and 9, respectively) and four levels of Gaussian noise (corresponding to noise standard deviations of 0, 7, 10 and 14, respectively) were used in the experiments. A region

of interest of 470 rows and 240 columns was selected from the processed images (of size 512 × 512). This restricted image size allowed simultaneous display of two images on the screen. The display was calibrated to have a gray-value-to-luminance characteristic equal to

$$L = \max[L_{\min}, L_{\max}(g/g_{\max})^{\gamma}], \tag{14}$$

for $0 \leqslant g \leqslant g_{\max}$, with $g_{\max} = 255$, $L_{\min} = 0.2\,\mathrm{cd/m^2}$, $L_{\max} = 60\,\mathrm{cd/m^2}$ and $\gamma = 2.5$. The distance between successive pixels, expressed in degrees of visual angle, was approximately 1 arcmin. Two experiments were performed with these stimuli.

In the first experiment, dissimilarity scores were collected for all combinations of 16 stimuli of the same scene. The subjects were urged to base their score (on an interval scale from 0 to 10) only on how dissimilar or different they perceived the images to be. Five subjects participated in the dissimilarity experiment. The dissimilarity data were mapped into two-dimensional stimulus configurations using the MDS program MULTISCALE described above. The resulting stimulus configurations [12] will be used as a reference for the dissimilarity models further on in this paper.

In the second experiment, seven subjects were asked to rate blur, noisiness and overall quality of the images (on an interval scale from 0 to 10). The stronger the perceived attribute, the higher the score they had to give. All stimuli were repeated four times. We used a model based on Thurstone's law of categorical judgement [29] to map the original interval judgements of the subjects into quality (or attribute) scores on a psychologically linear scale. The in-house software package THURCATD was used for this purpose [3]. The resulting quality scores, which are also reproduced in [12], will be used as reference for the quality predictions of the models in this paper.

2.2. JPEG-coded images

The four natural scenes that were used in this second experiment will be referred to as Boats, Child, Girls and Lighthouse, and were taken from a Kodak PhotoCD demonstration disc. The original color images were converted into monochrome images. A public-domain software package for JPEG encoding and decoding (Independent JPEG Software Group, http://www.jig.org/) was used to code these images at quality levels 60, 40, 30, 25 and 20. Including the original, we hence had six images per scene. A region of interest of 480 rows by 240 columns was selected from the images in order to allow simultaneous display of two images. The uncoded images are shown in Fig. 1. The display

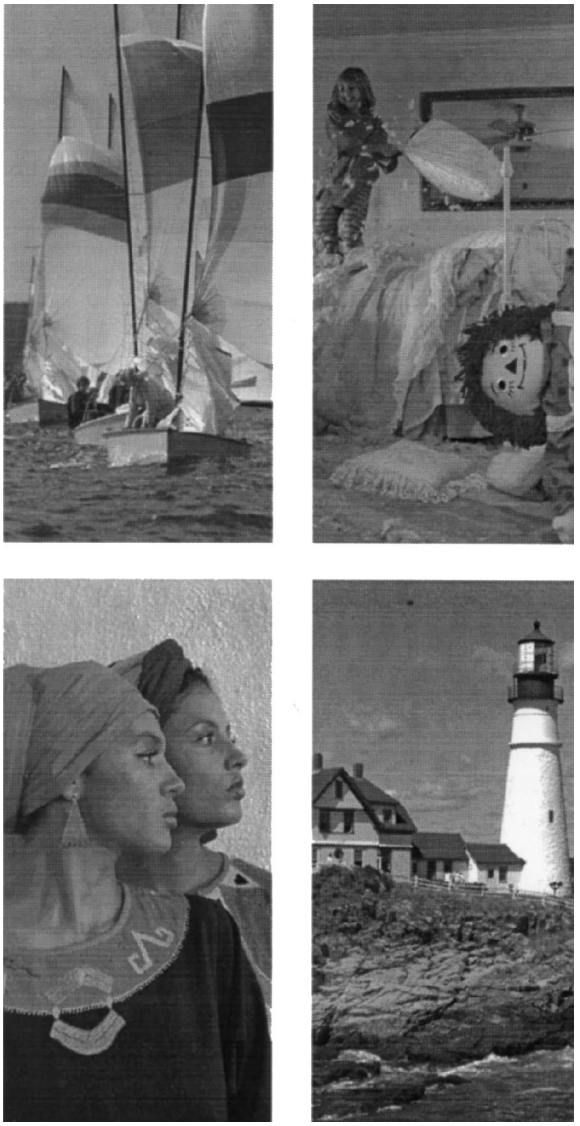

Fig. 1. Images used in the JPEG experiment: 'Boats' (upper left), 'Child' (upper right), 'Girls' (lower left) and 'Lighthouse' (lower right).

was calibrated to have the gray-value-to-luminance characteristic of Eq. (14). In order to allow subjects to adequately judge the small coding differences, the viewing distance was reduced to 80 cm, corresponding to a separation between pixels of approximately 2 arcmin of visual angle.

All 15 distinct combinations of the six stimuli were judged by 10 subjects. The subjects were asked to rate dissimilarity, as well as differences in overall quality and blockiness. The subjective scores for blockiness were intended to be used in another study [19]. More details about the experimental set-up and the data processing can also be found in this reference.

Quality differences (or preferences) were rated on a scale from -5 to 5. Positive and negative values indicated that the image on the right or the left of the screen was preferred, respectively. The absolute value expressed the strength of the quality difference. A value of zero could be used if none of the displayed stimuli was preferred. These preference scores between stimulus combinations were mapped into quality scores for the individual stimuli using the in-house software package DIFSCAL [3]. The resulting quality scores will be used as reference for the quality predictions of the models in this paper.

Similarly as in the case of the previous experiment, dissimilarity was judged on a scale from 0 to 10, and the experimental data were transformed into stimulus configurations using the MULTISCALE estimation program. We will use the resulting 1D stimulus configurations in this paper. We originally started with 2D solutions for the dissimilarity data, but found that the first dimension was very dominant, despite the fact that visual inspection of the JPEG-coded images identified three possible attributes: blockiness, blur and ringing. These impairments are very conspicuous in highly compressed JPEG images, but it was verified in a separate experiment that their strengths decrease in a linearly correlated way when the compression ratio decreases, so that only one independent dimension can be recovered from experiments with this JPEG-baseline-coded stimulus set. One way to control the underlying attributes more independently would be to use a richer data set in the experiment (for instance, JPEG-coded images with quantization matrices that are not only scaled versions of each other).

3. Instrumental dissimilarity measures

Most image quality metrics that have been proposed use as inputs two images, where most often one of the images is the original. The images are assumed to be perfectly alligned. Usually, a model of the early visual pathway is used to map the two separate image inputs I_1 and I_2 into two sets of visual system outputs $P(I_1)$ and $P(I_2)$ [1]. These models typically include aspects of luminance adaptation, decomposition into channels with different frequency/orientation tuning, and masking (both within and across channels). The two sets of visual system outputs are subsequently integrated into a single number $Q[P(I_1),P(I_2)] = d(I_1,I_2)$ that is assumed to be an instrumental measure for the perceived distance between the two images. It is moreover often assumed that the distance from the original image is a measure for the impairment, and that the image quality decreases linearly with the distance from the original. This latter assumption has already been seriously criticized in [25], based on experimental evidence.

Given the amount of assumptions and the number of parameter choices that have to be made in the mapping $I \rightarrow P(I)$, and given the limited experimental evidence on which most of these assumptions and choices are based (most models are tuned based on experimental data for simple patterns such as sinewaves and their overall quality prediction has only marginally been tested against experimental data), it is at least surprising that these 'image-quality metrics' have been so widely adopted, albeit in many varieties. If nothing else, it indicates the need for reliable instrumental measures of image quality.

The assumption that image quality decreases linearly with the distance from the original image is not essential for the above perceived distance measures to be used for image quality modelling. Indeed, similarly as in the case of the MDS approach of the previous section, we could use the results of an instrumental distance measure to position images in a multidimensional space. This would require

not only determining the distances between the original image and all processed images, but also the distances between (a sufficiently large subset of) all image pairs. The stimulus configuration resulting from such an instrumental measure could then be compared against the stimulus configuration resulting from dissimilarity judgements by subjects. Moreover, it could be investigated if such a multidimensional stimulus configuration is better suited for predicting image quality than the original one-dimensional configuration where only the distance from the original image is used.

In order to test the above ideas against experimental data, a representative subset of models had to be selected. We subsequently describe the models that we have implemented in sufficient detail to allow reproduction.

3.1. Sarnoff model

The Sarnoff model is regarded by many as the current de facto standard for an instrumental image quality metric, and many alternative models may be viewed as simplified versions of it [1]. The model for monochromatic still images is fairly well documented [16,17], although some parameters still have to be tuned for the specific viewing conditions. The model version for color image sequences is available as a commercial product [9], which explains why little or no detailed information is published about this latter model. We now describe the model that we have implemented in our study, including all model parameters. The model follows as closely as possible the original model description, although some small but inconsequential modifications (in the optical and pooling filter) have been made to simplify the implementation.

First, a gray-value image is converted into a luminance image according to Eq. (14). The optical filtering by the eye is simulated by a Gaussian filter with a standard deviation of 0.35 arcmin. The sampling distance is 1 arcmin in case of the images degraded by noise and blur, and 2 arcmin in case of the JPEG-coded images. In all filter operations, reflection of the image at the boundaries is applied.

The filtered luminance image is converted into a Gaussian pyramid with seven levels using a separable filter with coefficients (0.05,0.25,0.4,0.25,0.05) [4]. The original images of size 470×240 or 480×240 are extended to size 512×256 (and padded with zeros) before the pyramid is constructed. The Gaussian pyramid is subsequently mapped to a contrast pyramid with five levels, where the contrast at level k is as defined by Peli [20], i.e.,

$$C_k = \frac{G_k - G^i_{k+1}}{G^{ii}_{k+2} + \Phi/4^k}, \tag{15}$$

for $k = 0, \ldots, 4$, where the offset $\Phi = 0.1$ avoids divisions by zero. The image G^i_{k+1} arises by interpolating Gaussian pyramid level G_{k+1} with a separable interpolation filter with coefficients (0.1,0.5,0.8,0.5,0.1). Applying this interpolation twice to Gaussian pyramid level G_{k+2} results in G^{ii}_{k+2}. The images G_k, G^i_{k+1} and G^{ii}_{k+2} all have the same size.

In the next step, the contrast images C_k are mapped into channel responses. For this purpose the contrast images are convolved with four pairs of spatially oriented filters. The filters used are second derivatives of a Gaussian and their Hilbert transform, for four different orientations. The standard deviation of the Gaussian is approximately equal to the sampling distance, and the actual filter tabs are specified in Tables V and VI of the Freeman and Adelson paper on steerable filters [10]. The four oriented responses r_{kl}, for pyramid level index $k = 0, \ldots, 4$ and orientation index $l = 0, \ldots, 3$, are obtained by taking the square root of the sum of the squared filtered images for the Hilbert pairs with the same orientation.

The gains $g_{kl} = g_k$ for the different response channels have to be calibrated such that the peak sensitivities follow the contrast sensitivity function (CSF) of the visual system. This is accomplished by determining the peak response in the different channels to sinewave gratings with varying spatial frequency and with modulation depth equal to the inverse of the CSF for that frequency. The channel gain g_{kl} has to be adjusted such that the maximum response in channel (k,l) is equal to one for the spatial frequency and orientation for which that channel is most sensitive. We used the formula proposed by Barten [2] for an average luminance of $20\,\mathrm{cd/m^2}$ as a mathematical description of the CSF.

Although the above procedure seems straightforward, it runs into practical difficulties for spatial frequencies close to the sampling frequency. Hence, we have only used it to set the channel gains for pyramid levels 2, 3 and 4. The resulting gains were in agreement with the relative sensitivities of frequency channels with different peak frequencies reported in the contrast perception model of Cannon and Fullenkamp [5]. They also propose relative sensitivities for higher-frequency channels that seem more realistic than those resulting from the above calibration procedure, and which we have therefore adopted to set the channel gains for pyramid levels 0 and 1. The resulting channel gains depend on the viewing distance (because the CSF is specified in cycles per degree of visual angle), and therefore we specify these channel gains in Table 1 for sampling distances of 1 and 2 arcmin. Note that the peak sensitivity shifts to a channel with a smaller channel number when the sampling distance increases from 1 to 2 arcmin.

Within-channel visual masking is taken into account by mapping the (amplified) channel responses $a_{kl} = g_{kl} \cdot r_{kl}$, for $k = 0, \ldots, 4$ and $l = 0, \ldots, 3$, through a sigmoid non-linearity of the form

$$T(a) = \frac{(2 + c)\, a^s}{1 + a^{s-l} + c a^{s-w}}, \tag{16}$$

where $s = 1.5$, $l = 0.4$, $w = 0.068$ and $c = 0.1$ are typical values [16]. The 20 masked response images $T(a_{kl})$ are filtered using a pooling filter equal to a uniform filter of size 5×5, and subsequently interpolated to the input image size 512×256 using the interpolation filter mentioned in the contrast pyramid definition.

The resulting response images $R_{kl}^{(1)}(x,y)$ and $R_{kl}^{(2)}(x,y)$ for the first and second input image I_1 and I_2, respectively, are pointwise combined into one distortion image using the Minkowski metric

$$D_{12}(x,y) = \left\{ \sum_{k=0}^{4} \sum_{l=0}^{3} |R_{kl}^{(1)}(x,y) - R_{kl}^{(2)}(x,y)|^Q \right\}^{1/Q} \tag{17}$$

with exponent $Q = 2.4$. If the model is properly tuned, then a value of $D_{12}(x,y) = 1$ indicates 1 JND (just noticeable difference), i.e., a probability of 75% to see a difference between both images at the corresponding location (x,y).

In order to obtain a single dissimilarity number for a pair of images, we can perform a spatial Minkowski integration with exponent P across the distortion image $D_{12}(x,y)$, i.e.,

$$d_S(I_1,I_2) = \left\{ \frac{1}{N} \sum_{x,y} |D_{12}(x,y)|^P \right\}^{1/P}, \tag{18}$$

where N ($= 512 \times 256$) is the number of pixels at pyramid level 0. By varying P from one to infinity, we can vary from taking the average to taking the (scaled) maximum of the distortion image as the relevant distance measure. The 'Sarnoff' instrumental measure $d_S(I_1,I_2)$ that we use in the rest of the paper was calculated with $P = 1$.

3.2. Simplified measures

The original Sarnoff model described in the previous section requires a very large number of computations, and therefore we have also developed a simplified version of it. Especially the orientation and pooling filtering and the full-size interpolation at the end determine the total number of operations. We were interested to know if removing these components would seriously affect the predictions made by the model.

The directional filtering can be avoided by using an alternative method for mapping the contrast images C_k into channel responses. The purpose of using a Hilbert pair of oriented filters is to realize a phase independence for the channel responses, e.g., to make the responses insensitive to the exact position of an edge (with respect to the sampling lattice). Oriented filters are only required in case orientation masking is applied [28], which is not (currently) the case in the Sarnoff model. An alternative, and simpler, way of realizing phase-independent responses is by taking the residue amplitudes (or local standard deviations) [18], i.e.,

$$r_k(x,y) = \{ w(x,y) * C_k^2(x,y) - [w(x,y) * C_k(x,y)]^2 \}^{1/2}, \tag{19}$$

where a Gaussian filter $w(x,y)$ of the same standard deviation as in the case of the directional filtering is used. More specifically, this filter is approximated by a separable 9-tap filter with coefficients $w(4) = w(-4) = 0.00048$, $w(3) = w(-3) = 0.00880$,

$w(2) = w(-2) = 0.06965$, $w(1) = w(-1) = 0.23997$ and $w(0) = 0.36217$. The number of channels hence remains the same as the number of pyramid levels.

A consequence of the alternative channel response mechanism is that the channel gains g_k have to be calibrated accordingly. These alternative channel gains are also given in Table 1 for viewing distances of 1 and 2 arcmin. The amplified channel responses are denoted by $a_k = g_k \cdot r_k$, for $k = 0, \ldots, 4$. Similarly as in the case of the original Sarnoff model, the masked channel responses $T_k = T(a_k)$ are obtained by applying the sigmoid non-linearity in Eq. (16). A pooling filter is also not included in the simplified model. The standard deviation of the Gaussian residue amplitude filter could potentially be increased if response averaging over a larger area is required.

The differences between the masked channel responses $T_k^{(1)}(x,y)$ and $T_k^{(2)}(x,y)$ for both input images can be spatially integrated at each pyramid level k. More specifically, let

$$D_{12}(k) = \left\{ \frac{1}{N_k} \sum_{x,y} |T_k^{(1)}(x,y) - T_k^{(2)}(x,y)|^P \right\}^{1/P}, \qquad (20)$$

where N_k is the number of pixels at pyramid level k, for $k = 0, \ldots, 4$, denote the result of this spatial averaging with a Minkowski metric with exponent P. The overall instrumental measure for the perceived distance between two images I_1 and I_2 is obtained by Minkowski integration across pyramid levels

$$d_{SR}(I_1, I_2) = \left\{ \sum_{k=0}^{4} |D_{12}(k)|^Q \right\}^{1/Q}, \qquad (21)$$

Table 1
The two leftmost columns specify the channel gains g_k for the original Sarnoff model and sampling distances of 1 and 2 arcmin, respectively. The two rightmost columns give similar information for the simplified Sarnoff model

Channel number	Original Sarnoff		Simplified Sarnoff	
	1 arcmin	2 arcmin	1 arcmin	2 arcmin
0	25	60	170	420
1	85	133	450	960
2	125	111	845	885
3	90	68	670	535
4	50	37	385	295

with an exponent equal to Q. Correspondence with the original Sarnoff model implies the choice $P = 1$ and $Q = 2.4$. The need for full-size interpolation is avoided in the simplified model by reversing the order of across-channel integration and spatial integration. The consequence is that no single distortion image is available anymore. However, we are only interested in an overall distortion measure in the current study.

The simplified model retains four basic properties of the original model: contrast analysis at multiple spatial scales, contrast calibration against human CSF, within-channel contrast masking, and Minkowski integration (with different exponents across space and across channels).

Next to the original and simplified Sarnoff models, we have also included two very simple root-mean-square error (RMSE) models. These models operate on the psychometric lightness image $L^\star$ that is obtained from the luminance image L through the CIELAB lightness formula, i.e.,

$$L^\star = \begin{cases} 116(L/L_{max})^{1/3} - 16 \\ \quad \text{for } L/L_{max} \geqslant 0.008856, \\ 903.3(L/L_{max}) \\ \quad \text{for } L/L_{max} < 0.008856. \end{cases} \qquad (22)$$

This lightness image ranges from 0 to 100. For a well-calibrated monitor, this lightness image will typically be very close to the scaled gray-value image. The main reason for using the lightness image is that it allows to incorporate the gray-value-to-luminance characteristic of the monitor. This characteristic will for instance have an influence if the luminance of the monitor saturates at the low or high end, or if the gamma of the monitor deviates significantly from $\gamma = 3$.

Table 2
Images with noise and blur: correlations between (averaged) perceived image quality and distances from the original image (for 4 dissimilarity models and 3 scenes)

Model	Mondrian	Terrace	Wanda
d_S	0.568	0.950	0.830
d_{SR}	0.561	0.895	0.818
$RMSE_L$	0.855	0.951	0.795
$RMSE_{LR}$	0.748	0.931	0.724

The first RMSE model that we use simply determines the RMSE between two lightness images $L_1^\star(x,y)$ and $L_2^\star(x,y)$, i.e.,

$$\mathrm{RMSE_L}(I_1,I_2) = \sqrt{\frac{1}{N}\sum_{x,y}[L_1^\star(x,y) - L_2^\star(x,y)]^2}, \quad (23)$$

where N is the number of image pixels. The second RMSE model does not operate directly on the lightness images, but on the residue amplitudes derived from these lightness images [18]. We used the same 9-tap Gaussian filter as mentioned above in relation to Eq. (19). The resulting RMSE measure will be denoted by $\mathrm{RMSE_{LR}}(I_1,I_2)$ and was proposed in order to see how a measure that reacts to differences in lightness variations compares to a measure like $\mathrm{RMSE_L}(I_1,I_2)$ that reacts to differences in absolute lightness levels.

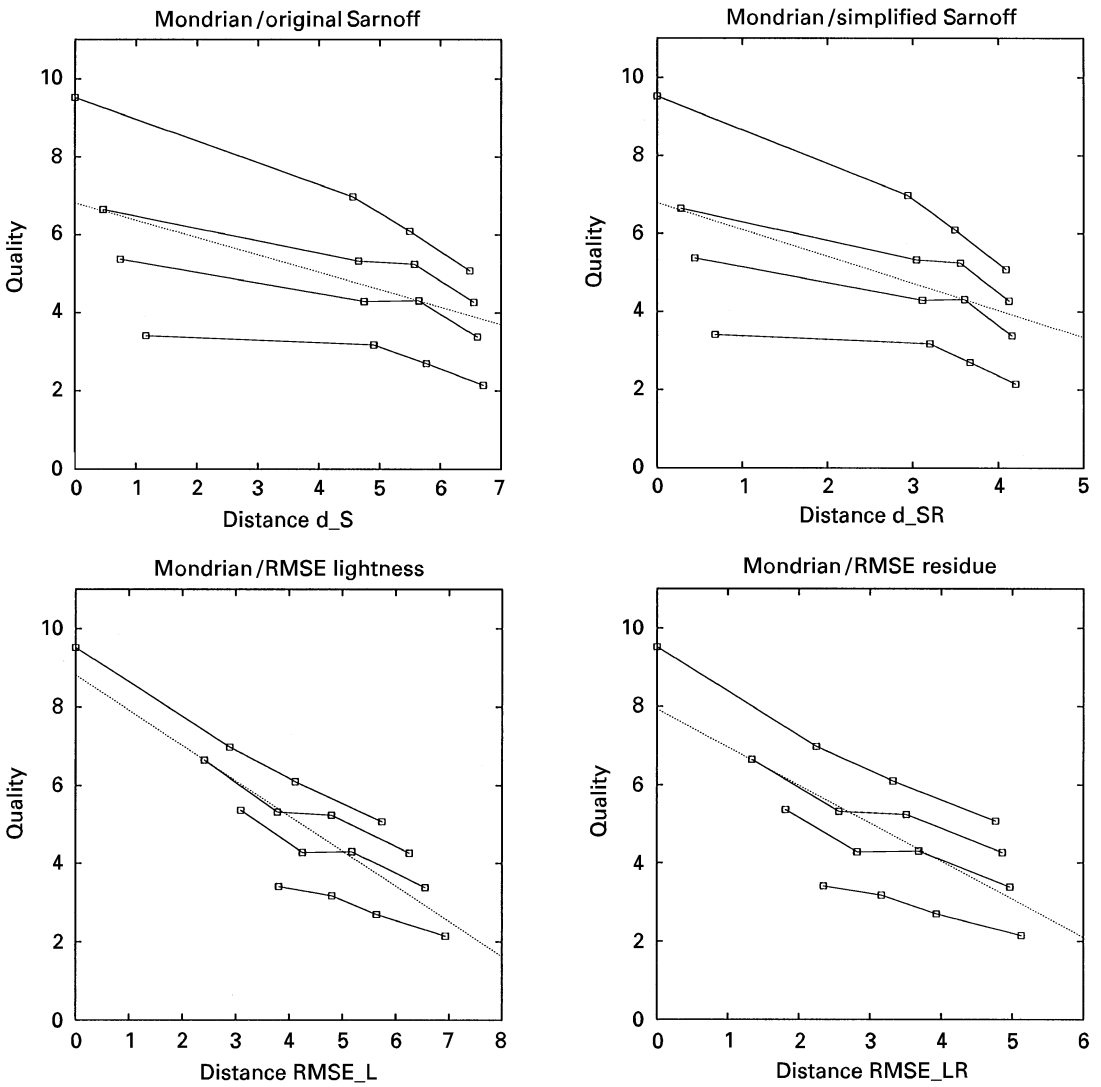

Fig. 2. Subjective quality judgements in the blur/noise-experiment for scene 'Mondrian' versus distances from the reference image for four instrumental models: original Sarnoff (upper left), simplified Sarnoff (upper right), RMSE on lightness (lower left) and RMSE on residue amplitude (lower right). Connected points correspond to the same amount of blur.

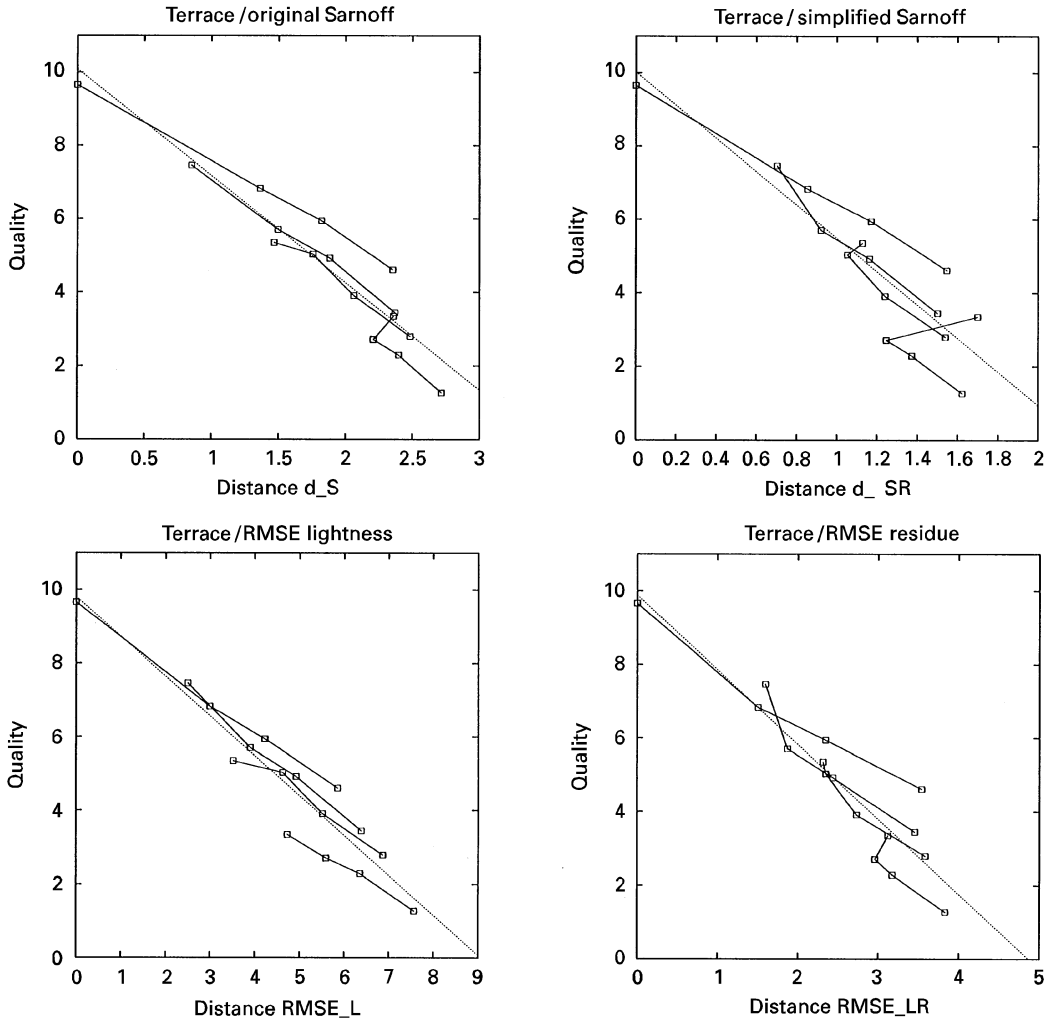

Fig. 3. Subjective quality judgements in the blur/noise-experiment for scene 'Terrace' versus distances from the reference image for four instrumental models: original Sarnoff (upper left), simplified Sarnoff (upper right), RMSE on lightness (lower left) and RMSE on residue amplitude (lower right). Connected points correspond to the same amount of blur.

Table 3
Images with noise and blur: two important parameters from the MULTISCALE estimation of 2-D stimulus configurations, i.e., the power-law exponent p for mapping dissimilarities and the correlation r between the transformed dissimilarities and the distances in the configuration

	Mondrian		Terrace		Wanda	
Model	p	r	p	r	p	r
d_S	1.045	1.000	1.022	0.999	1.089	1.000
d_{SR}	1.018	0.999	0.995	0.998	1.010	1.000
$RMSE_L$	1.046	0.999	1.038	1.000	1.038	1.000
$RMSE_{LR}$	1.008	1.000	1.019	1.000	1.017	1.000

4. Instrumental measures versus experimental data

4.1. Images with noise and blur

In the existing models, the distance from the original image is used as the prediction for quality. In Table 2, we show the correlations between the quality scores (averaged over seven subjects) and these predictions for the four proposed models and the three scenes used in the experiments with noise and blur. Note that the correlations for d_S and d_{SR} are especially low for the 'Mondrian' scene. The simplified Sarnoff model performs worse than the original Sarnoff model, and the RMSE measure on the residue amplitude is worse than the RMSE measure on the lightness values. The complex

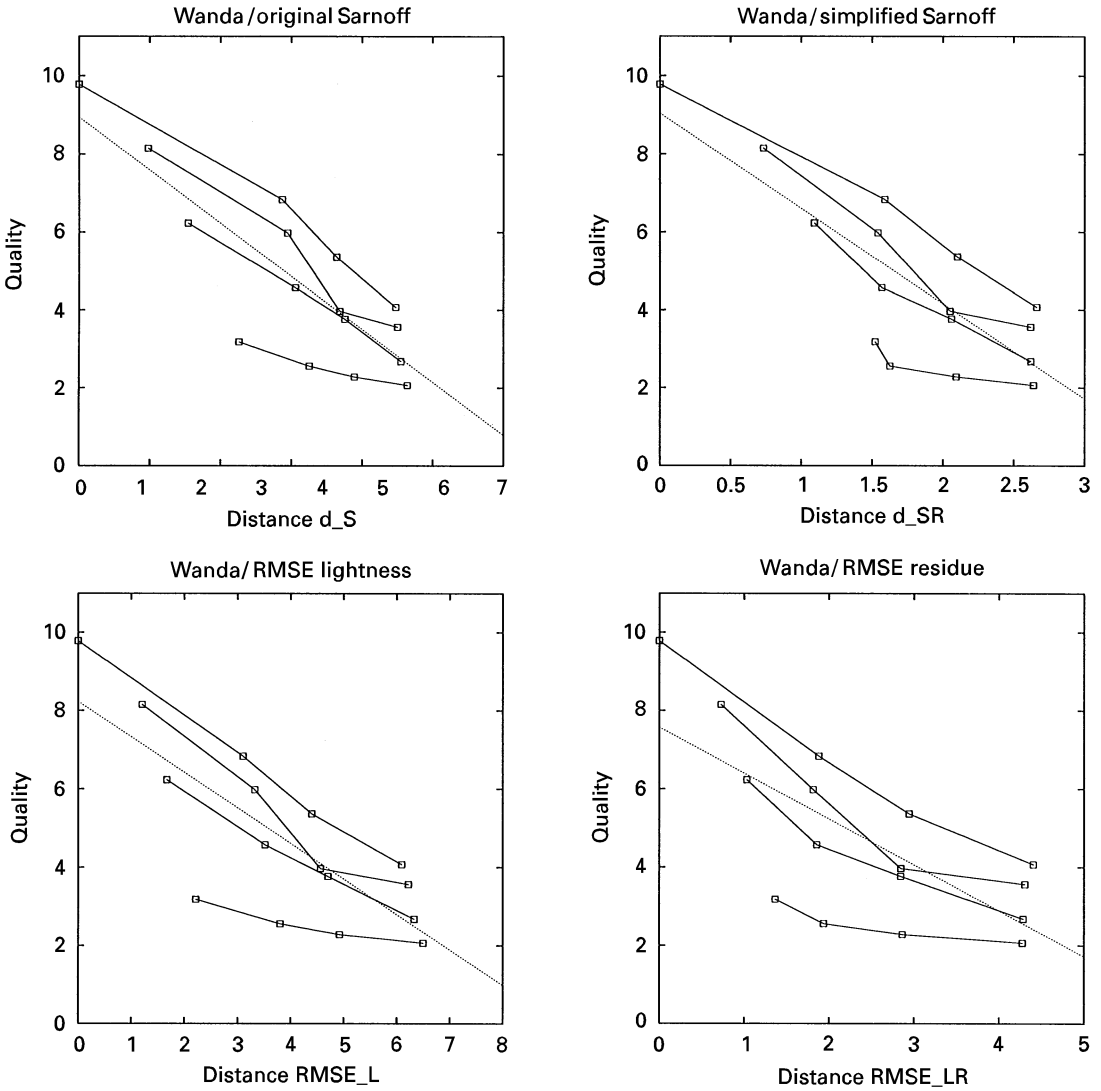

Fig. 4. Subjective quality judgements in the blur/noise-experiment for scene 'Wanda' versus distances from the reference image for four instrumental models: original Sarnoff (upper left), simplified Sarnoff (upper right), RMSE on lightness (lower left) and RMSE on residue amplitude (lower right). Connected points correspond to the same amount of blur.

Sarnoff model only performs better than the simple RMSE measure in case of the 'Terrace' scene.

In Figs. 2–4, we have plotted the relationship between the instrumental measures and the subjective quality scores for the respective scenes 'Mondrian', 'Terrace' and 'Wanda'. The data points corresponding to images with the same amount of blur and varying levels of noise standard deviation have been connected. There is a monotonous relationship between the 1-D model predictions and subjective image quality in case only one distortion is present (i.e., only noise or only blur). For the 'Mondrian' and 'Wanda' scene, the slope of this relationship is much steeper in the case of images degraded by blur. The fact that there is usually less variation in the model predictions for images with different amounts of blur indicates that the quality variations due to blur are typically underestimated by the models (as compared with quality variations due to noise). This latter observation will also be confirmed when the model predictions between all pairs of images are transformed into 2-D stimulus

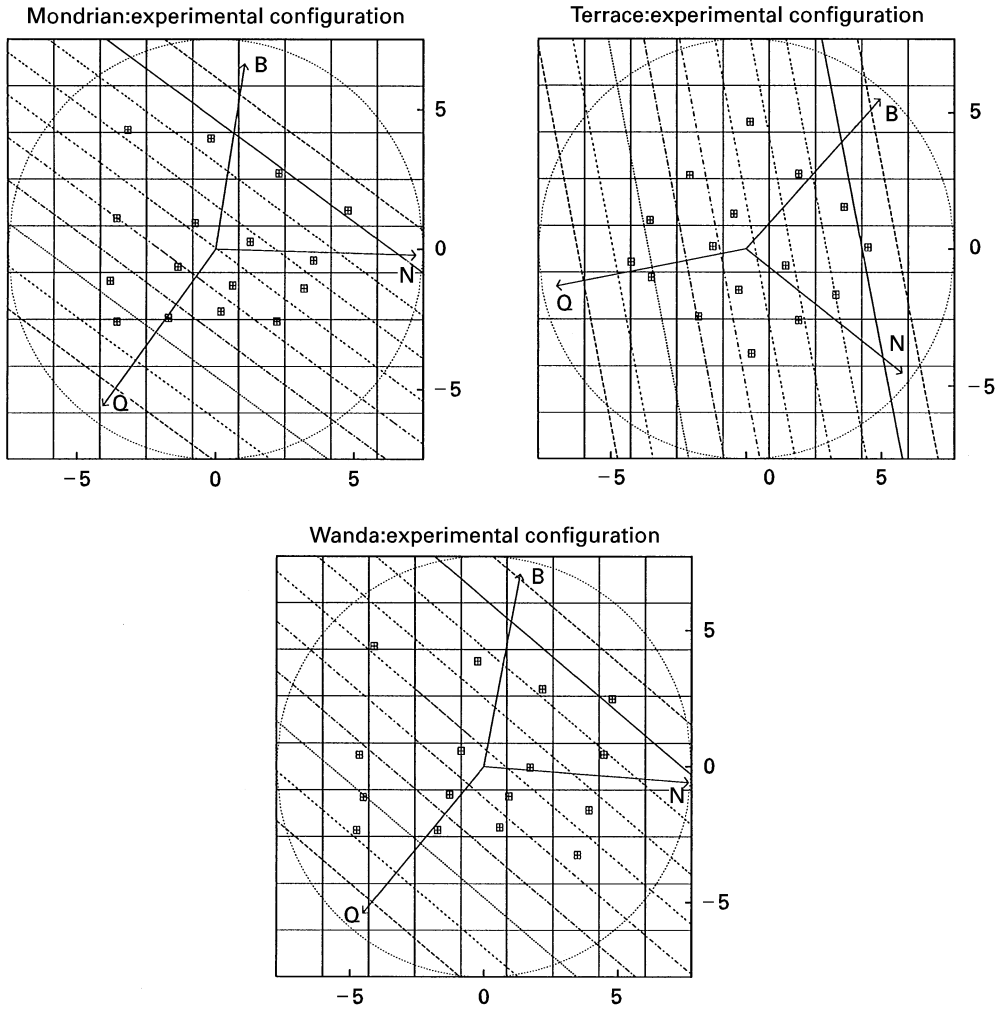

Fig. 5. Experimental stimulus configurations in the blur/noise experiment for scenes 'Mondrian' (upper left), 'Terrace' (upper right) and 'Wanda' (lower middle). The quality, blur and noisiness vectors are indicated by 'Q', 'B' and 'N', respectively.

Table 4
Images with noise and blur: the average distance $\bar{d}$ between stimulus positions and the inner-product correlation coefficient ρ obtained when comparing the stimulus configuration from subjective dissimilarity measurements with the stimulus configurations from model calculations

	Mondrian		Terrace		Wanda	
Model	$\bar{d}$	ρ	$\bar{d}$	ρ	$\bar{d}$	ρ
d_S	1.934	0.804	1.301	0.904	1.519	0.912
d_{SR}	2.026	0.759	1.663	0.840	2.020	0.825
$RMSE_L$	1.036	0.948	0.844	0.963	1.156	0.947
$RMSE_{LR}$	1.145	0.930	1.078	0.937	1.553	0.905

configurations. Note that, in the case of the 'Terrace' scene, some of the algorithms even have problems predicting the right order if both the blur and the noise standard deviation are large.[2]

We used the MULTISCALE estimation program described in Section 2 to estimate 2-D Euclidean configurations from all pairwise distances given by the models. For all four models we find that the predicted distances between the images can be very well described by 2-D Euclidean configurations. This can be judged from Table 3 where it is shown that the correlations r between the transformed model distances (after the power-law transformation) and the Euclidean distances between the image points in the configuration are very close to one for all four metrics. The exponent p is also close to one in all cases, which indicates that a power-law transformation is not even strictly necessary, but that the model distances themselves can be interpreted as Euclidean distances.

We compared the stimulus configurations resulting from the instrumental models with the stimulus configurations obtained from the dissimilarity judgements by subjects [12]. The experimentally obtained stimulus configurations for all three images are reproduced in Fig. 5. The directions for (average) quality, blur and noisiness are also indicated in these figures. The lengths of these attribute vectors are proportional to the communalities (i.e., the squares of the correlation coefficients between the experimental data and the stimulus coordinates in the direction of the quality/attribute vector). Perfect correlation corresponds to the dotted circle. Lines of constant quality, which are orthogonal to the quality vector, are also shown. Note that the noisiness and blur vectors are approximately orthogonal, and that quality is a compromise between blur and noise. The fact that the quality vector has a direction which is closer to the direction of the blur vector than to the direction of the noisiness vector moreover implies that blur is the dominant attribute for quality.

Table 5
Images with noise and blur: correlations between (averaged) perceived image quality and coordinates along an optimized direction in 2-D space for four dissimilarity models, one experimental dissimilarity configuration and three scenes

Model	Mondrian	Terrace	Wanda
d_S	0.963	0.956	0.963
d_{SR}	0.919	0.935	0.842
$RMSE_L$	0.947	0.956	0.933
$RMSE_{LR}$	0.954	0.963	0.951
exp	0.959	0.972	0.953

Comparing the experimental configurations with the configurations derived from model calculations again reveals that the variations due to blur are compressed in the latter models. As described in Section 2, the stimulus configurations in Figs. 6–8 have been optimally transformed (using

[2] In the case of the Sarnoff models, a better fit could probably be obtained by altering the channel gains. However, these channel gains are not considered as free parameters, since they are calibrated based on CSF data, and intended to be fixed for all stimuli.

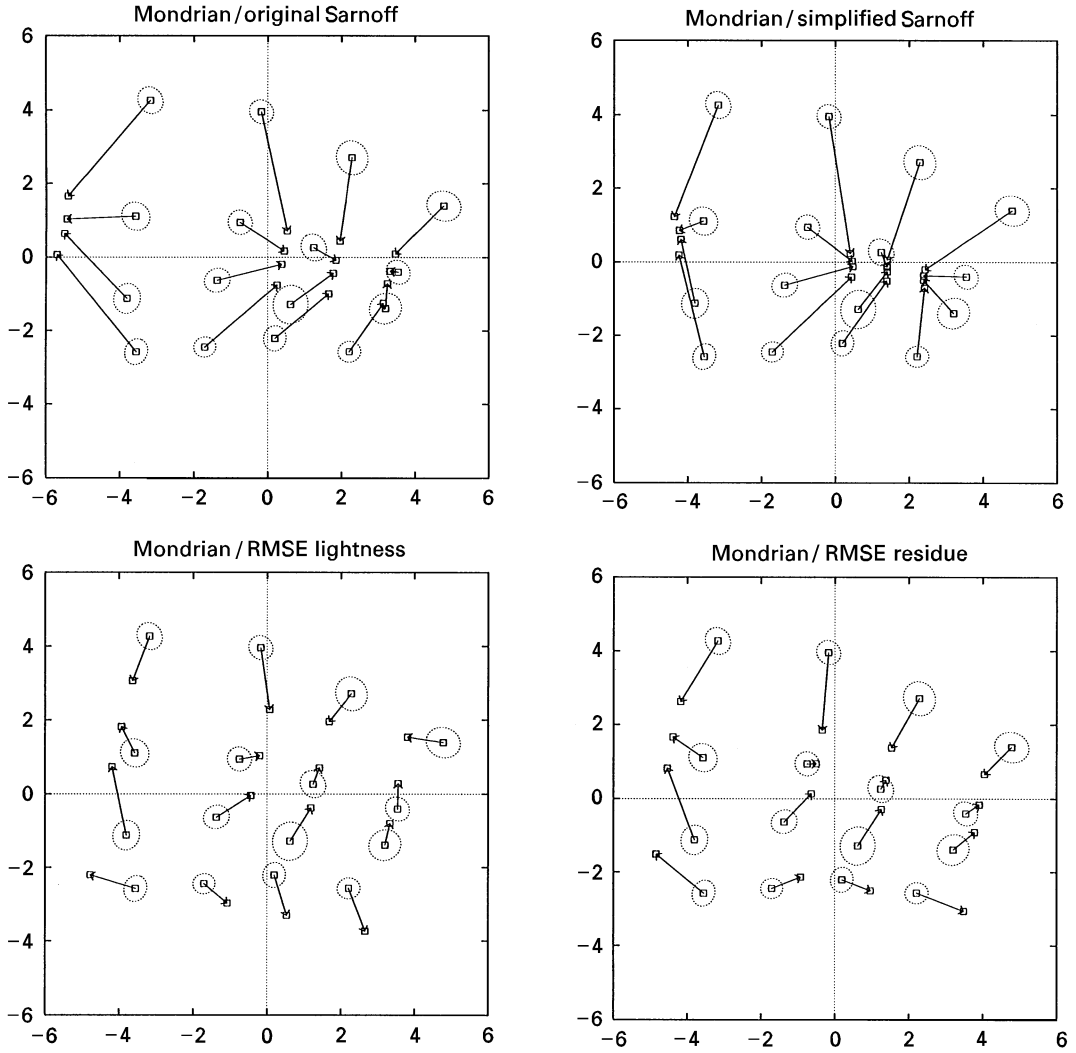

Fig. 6. Experimental stimulus configuration in the blur/noise experiment for scene 'Mondrian' versus stimulus configurations from four instrumental models: original Sarnoff (upper left), simplified Sarnoff (upper right), RMSE on lightness (lower left) and RMSE on residue amplitude (lower right). The experimental data are at the centers of the 95% confidence regions.

scaling, translation and unitary transformation) in order to allow an optimal comparison with the experimental data. The experimental data correspond to the centers of the 95% confidence ellipses, while the transformed model data are the end points of the arrows. None of the instrumental measures is able to make a good prediction for the measured dissimilarities, since the arrows are typically much larger than the sizes of the uncertainty ellipses. The average distance $\bar{d}$ between the image points in both configurations, as well as the inner-product correlations (according to Eq. (9)), are listed in Table 4. The $\mathrm{RMSE_L}$ measure obviously performs better than all other measures since the relative compression of the blur versus noise dimension is least pronounced for this measure.

All configurations can be used to perform quality predictions. We can use Eq. (13) to model the quality direction in 2-D space, and maximize the regression between subjective data and model predictions.

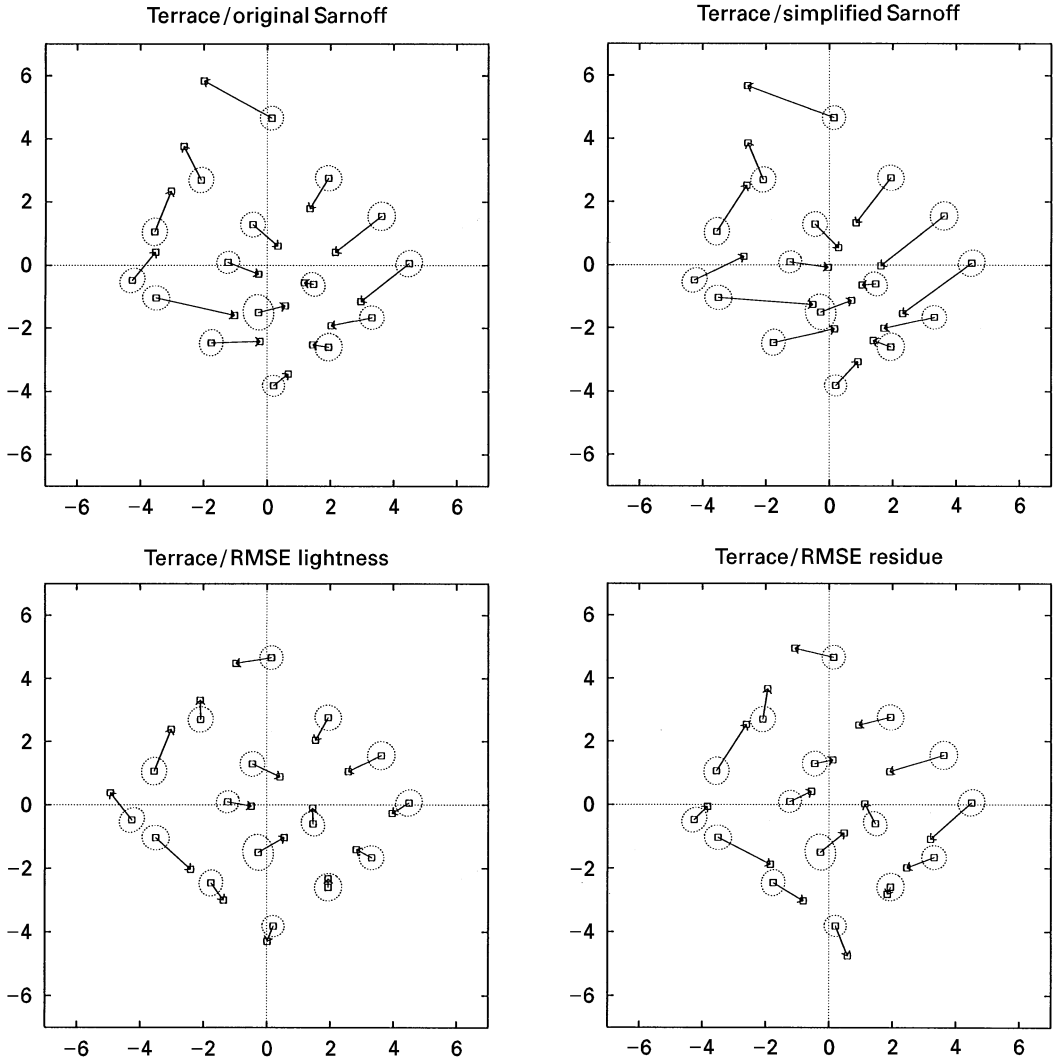

Fig. 7. Experimental stimulus configuration in the blur/noise experiment for scene 'Terrace' versus stimulus configurations from four instrumental models: original Sarnoff (upper left), simplified Sarnoff (upper right), RMSE on lightness (lower left) and RMSE on residue amplitude (lower right). The experimental data are at the centers of the 95% confidence regions.

Since we use an additional parameter in the quality prediction which is able to counteract the above-mentioned relative compression, we expect that all configurations are about equally well suited as basis for a quality model, which is confirmed by the correlation results in Table 5. This table also lists the correlations that are obtained when the configurations resulting from experimental dissimilarity judgements are used as basis for the quality predictions.

In conclusion, also in this alternative 2-D approach, the Sarnoff models are unable to demonstrate a clear benefit over simple RMSE measures. The simplified Sarnoff model performs worse than the original Sarnoff model, both in predicting dissimilarities and in predicting quality.

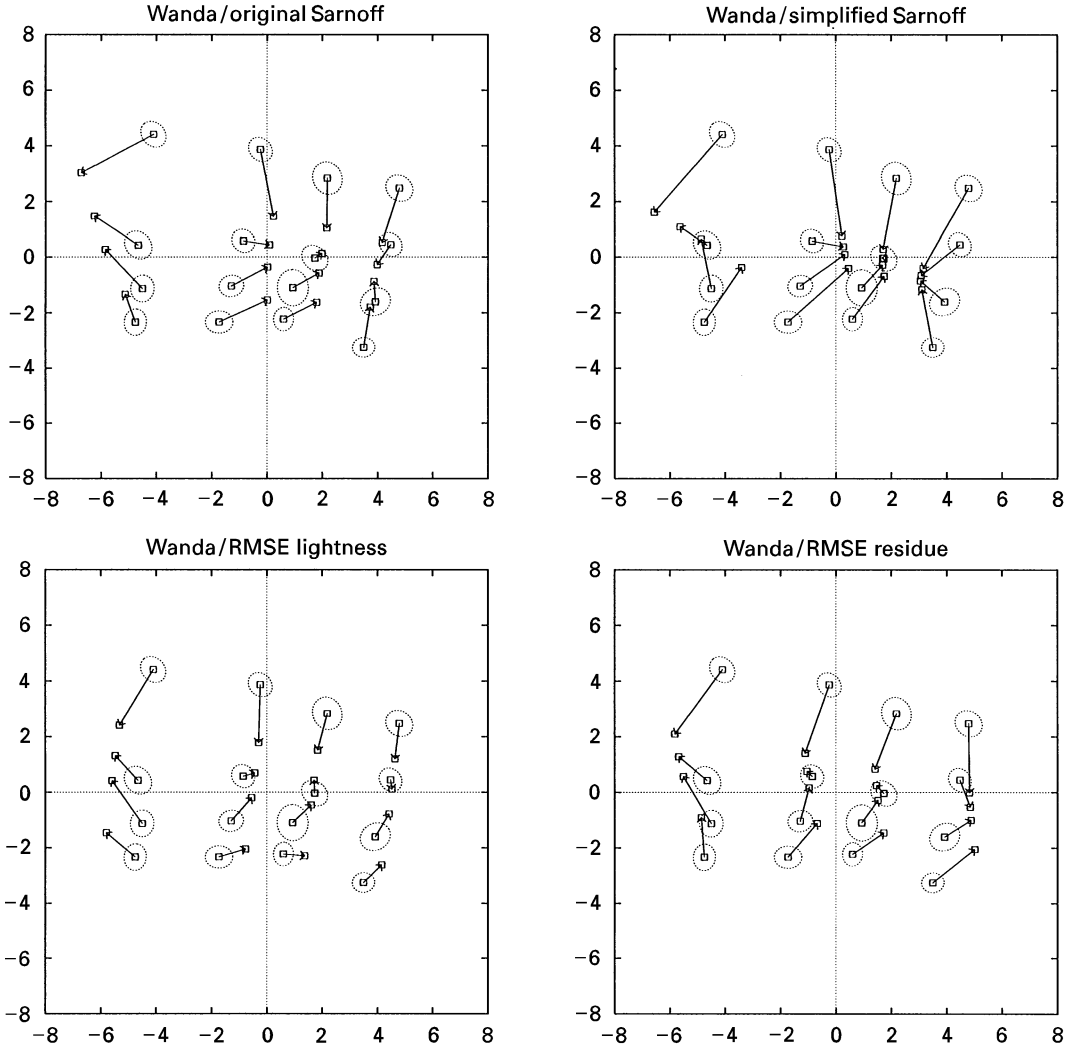

Fig. 8. Experimental stimulus configuration in the blur/noise experiment for scene 'Wanda' versus stimulus configurations from four instrumental models: original Sarnoff (upper left), simplified Sarnoff (upper right), RMSE on lightness (lower left) and RMSE on residue amplitude (lower right). The experimental data are at the centers of the 95% confidence regions.

4.2. JPEG-coded images

Since the quality variations in the JPEG-coded images are accomplished by the systematic variation of one coder parameter (i.e., a scaling factor for the quantization matrix), we expect that a 1-D prediction model will perform much better than in the previous case of two simultaneous but independent distortions. If we use the distance from the original image as the quality prediction for each of the above four models, then we obtain the correlations listed in Table 6. The RMSE measure on the residue amplitude of lightness seems to have the best performance.

In Fig. 9 we have plotted the subjective quality judgements versus the distances from the original image for all four instrumental measures and all test scenes. In order to present all data in one figure, we have scaled all measures except the RMSE. The factors used to scale the different distance measures

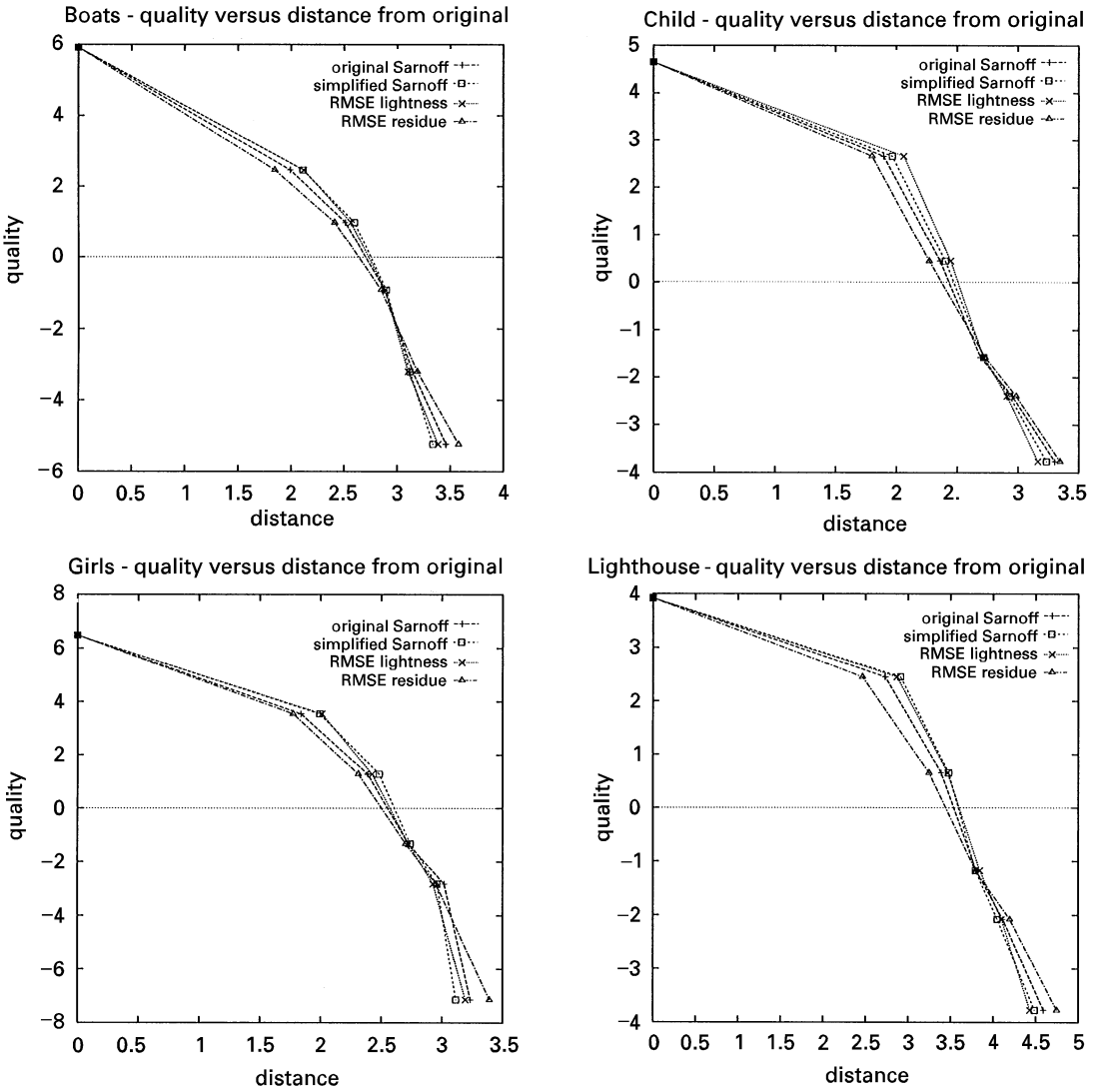

Fig. 9. Subjective quality judgements in the JPEG experiment versus distance from the reference image for four instrumental models and four images: 'Boats' (upper left), 'Child' (upper right), 'Girls' (lower left) and 'Lighthouse' (lower right).

Table 6
JPEG-coded images: correlations between (averaged) perceived image quality and distances from the original image for four dissimilarity models and four scenes. The numbers between brackets are the correlations when the original is excluded

Model	Boats	Child	Girls	Lighthouse
d_S	0.935 (0.978)	0.935 (0.995)	0.897 (0.956)	0.900 (0.997)
d_{SR}	0.910 (0.968)	0.922 (0.997)	0.861 (0.945)	0.872 (0.996)
$RMSE_L$	0.916 (0.978)	0.903 (0.997)	0.874 (0.975)	0.874 (0.995)
$RMSE_{LR}$	0.958 (0.985)	0.951 (0.998)	0.924 (0.984)	0.934 (0.998)

Table 7
JPEG-coded images: multiplication factors used in order to represent distances from the original image on a common scale in Fig. 9 (for four dissimilarity models and four scenes)

Model	Boats	Child	Girls	Lighthouse
d_S	1.973	1.862	1.517	2.484
d_{SR}	2.946	3.053	2.523	3.465
$RMSE_L$	1.000	1.000	1.000	1.000
$RMSE_{LR}$	3.083	3.069	2.908	3.393

are given in Table 7. A closer look at these factors reveals that the distances are much smaller (i.e., closer to 1 JND in the case of the Sarnoff model) than in the previous case of images degraded by noise and blur.

The graphs in Fig. 9 illustrate that the predictions for the coded images are almost on a straight line, but that the quality of the original image is ill-predicted by extrapolating this linear relationship. This failure of the instrumental models could however easily be remedied by putting a threshold on the model predictions. The numbers between parentheses in Table 6 give the correlations between measured quality and model predictions if the original is excluded. The models perform about equally well in this case, although $RMSE_{LR}$ retains the highest correlation values.

We can also construct 1-D Euclidean configurations from all pairwise distances given by the models and all pairwise dissimilarity scores given by the subjects. The correlations in Table 8 indicate that the distances predicted by the models can again be reasonably well described by Euclidean configurations. There is however a noticeable difference between the power-law exponents in Table 8 and those in Table 3, which needs some clarification. The difference between both data sets is that the model distances for the JPEG-coded images are smaller (i.e., close to the visible threshold). For these small values, we need a power-law transformation with exponent larger than one. For larger distances, as in the case of the images degraded by noise and blur, a linear function was adequate. This is in accordance with previous studies, where an

Table 8
JPEG-coded images: two important parameters from the MULTISCALE estimation of 1-D stimulus configurations: the power-law exponent p for mapping dissimilarities and the correlation r between the transformed dissimilarities and the distances in the configuration

	Boats		Child		Girls		Lighthouse	
Model	p	r	p	r	p	r	p	r
d_S	3.688	0.972	3.715	0.987	4.305	0.954	3.508	0.975
d_{SR}	3.148	0.966	3.026	0.975	3.786	0.934	2.835	0.984
$RMSE_L$	5.009	0.973	4.825	0.980	5.026	0.960	4.593	0.973
$RMSE_{LR}$	4.942	0.966	4.049	0.985	3.665	0.988	4.103	0.990

Table 9
JPEG-coded images: the average distance $\bar{d}$ between stimulus positions and the inner-product correlation coefficient ρ obtained when comparing the stimulus configuration from subjective dissimilarity measurements with the stimulus configurations from model calculations (for four dissimilarity models and four scenes)

	Boats		Child		Girls		Lighthouse	
Model	$\bar{d}$	ρ	$\bar{d}$	ρ	$\bar{d}$	ρ	$\bar{d}$	ρ
d_S	0.967	0.979	2.664	0.969	2.979	0.951	3.121	0.981
d_{SR}	0.899	0.970	2.836	0.975	1.006	0.994	4.831	0.958
$RMSE_L$	1.092	0.966	2.868	0.972	3.249	0.917	3.150	0.982
$RMSE_{LR}$	1.640	0.923	3.195	0.962	2.654	0.961	4.733	0.957

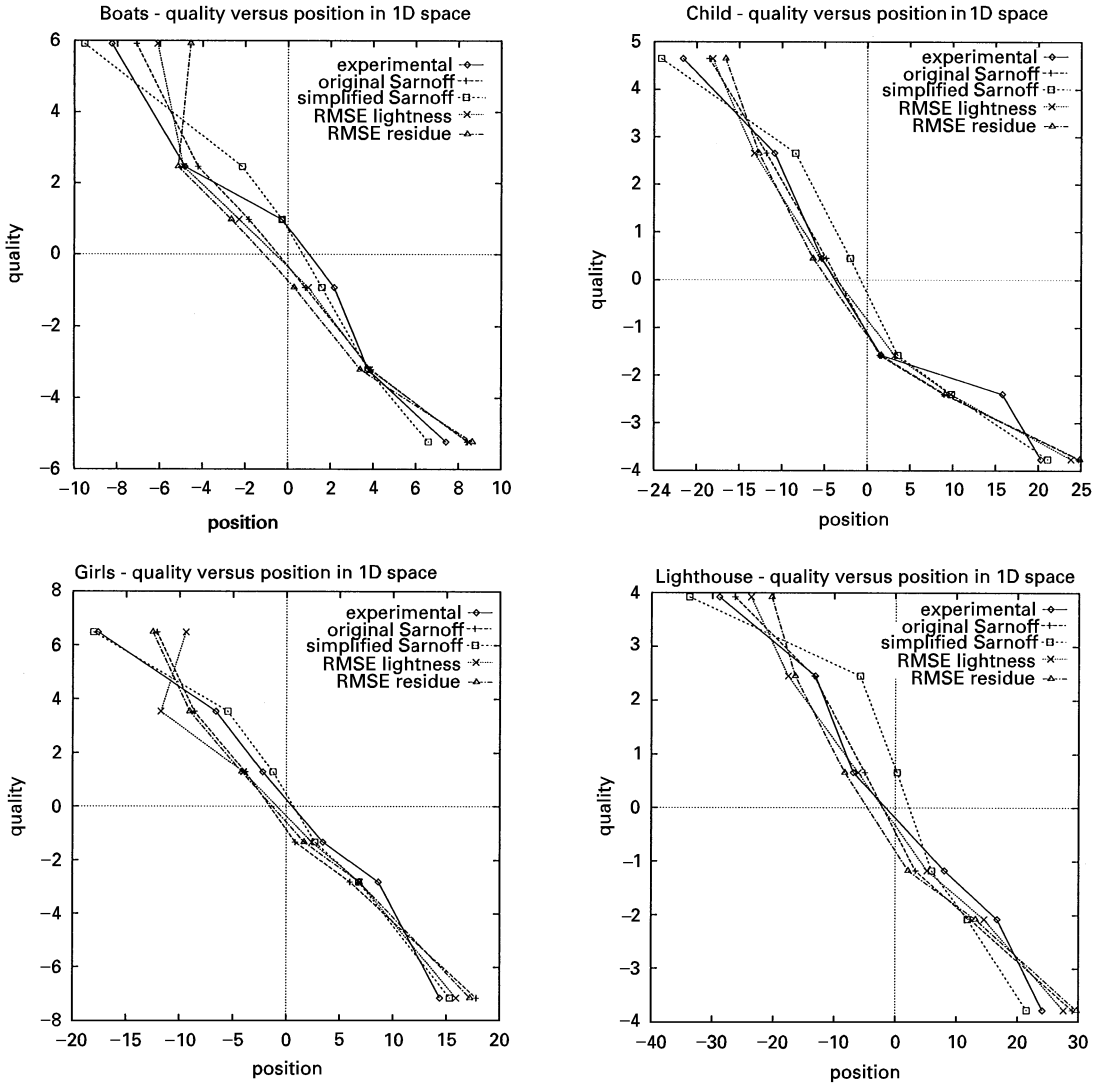

Fig. 10. Subjective quality judgements in the JPEG experiment versus positions in a 1-D space for four instrumental models and 4 images: 'Boats' (upper left), 'Child' (upper right), 'Girls' (lower left) and 'Lighthouse' (lower right).

S-shaped function (which includes threshold behaviour) is used to map an objective parameter (such as standard deviation of the noise or size of the blurring kernel) into a variable that is more linearly related to a visual sensation [13]. For values close to the threshold, such an S-shaped function can be approximated by a power law with an exponent larger than one.

The stimulus configurations resulting from the instrumental models can be compared with the stimulus configuration from experimental dissimilarity judgements. As in the previous section, comparisons are performed after optimal transformations of the model configurations. For the 1-D case, this reduces to an arbitrary linear transformation between stimulus positions. Contrary to the case of images degraded by noise and blur, most instrumental measures are able to make a reasonably accurate prediction for the measured dissimilarities. The average distance $\bar{d}$ between the image

Table 10
JPEG-coded images: correlations between (averaged) perceived image quality and coordinates along an optimum quality direction in 1-D space for four dissimilarity models, one experimental dissimilarity configuration and four scenes

Model	Boats	Child	Girls	Lighthouse
d_S	0.987	0.954	0.987	0.987
d_{SR}	0.982	0.979	0.983	0.942
$RMSE_L$	0.970	0.965	0.963	0.994
$RMSE_{LR}$	0.927	0.942	0.991	0.972
exp	0.986	0.971	0.985	0.992

points in both configurations, as well as the correlations (according to Eq. (9)), are listed in Table 9.

All configurations are about equally well suited for performing linear quality predictions. In Fig. 10 we have plotted the subjective quality judgements versus the stimulus positions for all four instrumental measures and all test scenes. The stimulus positions for the instrumental measures have been linearly transformed to minimize the average distance from the experimental stimulus positions. Table 10 lists the correlations between quality scores and 1-D stimulus coordinates for all quality models. A comparison between Table 10 and Table 6 reveals that the updated models correlate much more linearly with subjective quality. However, this improvement is mostly due to the nonlinear relationship between model dissimilarity and distance (i.e., the threshold mechanism discussed above). This nonlinear threshold transformation mostly influences the position of the original image. Including such a threshold mechanism in the distance measures from the original image would have accomplished a similar performance improvement, as can easily be judged from the graphs in Fig. 9.

5. Conclusions

Many existing models for perceived image quality are based on a distance metric between the original image and the processed/coded versions of it. If the variation in the images can be described as a variation in only one perceptual dimension (as in the case of the JPEG-coded images), then these models can perform well (especially if the nonlinear behaviour close to threshold is properly taken into account).

The dimensionality of a stimulus configuration can be studied using both instrumental measures and dissimilarity judgements by subjects, provided (a large subset of) all distances between pairwise stimulus combinations are determined. An estimation program like MULTISCALE can be used to determine (two- or more-dimensional) stimulus configurations from such pairwise comparisons.

If more than one dimension is needed for the stimulus configurations, then more than one independent psychological dimension is probably involved. The data for combined noise and blur demonstrate that the existing models are not able to balance the different kinds of distortions, so that the distance from the original often does not correlate well with the perceived quality. Nevertheless, the higher-dimensional stimulus configurations resulting from the instrumental measures can still be used as the basis for an image-quality model. An optimum quality direction can be determined by correlating with (a subset of) the subjective data. The direction of the optimum quality vector then determines the relative weight with which the different dimensions contribute in overall quality.

In none of the examined cases could a clear advantage of complicated distance metrics (such as the Sarnoff model) be demonstrated over simple measures such as RMSE.

Acknowledgements

The authors wish to acknowledge the support of the ACTS AC055 project 'Tapestries'.

References

[1] A.J. Ahumada Jr., Computational image-quality metrics: a review, in: SID 93 Digest, Society for Information Display, Santa Ana, CA, 1993. pp. 305–308.
[2] P.G.J. Barten, The sqri method: a new method for the evaluation of visible resolution on a display, Proc. Soc. Inform. Disp. 30 (1987) 253–262.
[3] M.C. Boschman, J.A.J. Roufs, Text quality metrics for visual display units: ii. an experimental survey, Displays 18 (1997) 45–64.

[4] P.J. Burt, E. Adelson, The laplacian pyramid as a compact image code, IEEE Trans. Commun. 31 (April 1983) 532–540.
[5] M.W. Cannon, S.C. Fullenkamp, A transducer model for contrast perception, Vision Res. 31 (1991) 983–998.
[6] S. Daly, Visible differences predictor: an algorithm for the assessment of image fidelity, in: B.E. Rogowitz, (Ed.), Human Vision, Visual Processing, and Digital Display III, Proc. SPIE, Vol. 1666, 1992, pp. 2–15.
[7] S. Daly, Visible differences predictor: an algorithm for the assessment of image fidelity, in: A.B. Watson (Ed.), Digital Images and Human Vision, MIT Press, Cambridge, MA, 1993, pp. 179–206.
[8] B. Escalante-Ramírez, J.B. Martens, H. de Ridder, Multidimensional characterization of the perceptual quality of noise-reduced computed tomography images, J. Visual Commun. Image Representation 6 (December 1995) 317–334.
[9] D.K. Fibush, Practical application of objective picture quality measurements, in: IBC97, International Broadcast Union, 1997, pp. 123–135.
[10] W.T. Freeman, E.H. Adelson, The design and use of steerable filters, IEEE Trans. Pattern Anal. Mach. Intell. 13 (September 1991) 891–906.
[11] ITU, Method for the subjective assessment of the quality of television pictures, Technical Report ITU Recommendation 500-3, International Television Union, Geneva, 1992.
[12] V. Kayargadde, J.B. Martens, Perceptual characterization of images degraded by blur and noise: experiments, J. Opt. Soc. Amer. A 13 (June 1996) 1166–1177.
[13] V. Kayargadde, J.B. Martens, Perceptual characterization of images degraded by blur and noise: model, J. Opt. Soc. Amer. A 13 (June 1996) 1178–1188.
[14] J.B. Kruskal, M. Wish, Multidimensional Scaling, Sage University Paper Series 07-011 on Quantitative Applications in the Social Sciences, Sage Publications, Beverley Hills, CA, 1978.
[15] P. Lindh, C.J. van den Branden Lambrecht, Efficient spatio-temporal decomposition for perceptual processing of video sequences, in: IEEE Internat. Conf. on Image Processing, Vol. III of III, Lausanne, Switzerland, September 1996, pp. 331–334.
[16] J. Lubin, The use of psychophysical data and models in the analysis of display system performance, in: A.B. Watson (Ed.), Digital Images and Human Vision, MIT Press, Cambridge, MA, 1993, pp. 162–178.
[17] J. Lubin, A visual discrimination model for imaging system design and evaluation, in: E. Peli (Ed.), Visual Models for Target Detection and Recognition, World Scientific, River Edge, NJ, 1995.
[18] J.-B. Martens, Adaptive contrast enhancement through residue-image processing, Signal Processing 44 (1995) 1–18.
[19] L. Meesters, J.B. Martens, Blockiness estimation in jpeg-coded images, Internal communication.
[20] E. Peli, Contrast in complex images, J. Opt. Soc. Amer. A 7 (1990) 2032–2040.
[21] W.B. Pennebaker, J.L. Mitchell, JPEG Still Image Compression Standard, Van Nostrand Reinhold, New York, 1993.
[22] J.O. Ramsay, Maximum likelihood estimation in multidimensional scaling, Psychometrika 42 (1977) 241–266.
[23] J.O. Ramsay, Confidence regions for multidimensional scaling analysis, Psychometrika 43 (1978) 145–160.
[24] J.O. Ramsay, J. ten Berge, G.P.H. Styan, Matrix correlation, Psychometrika 49 (September 1984) 403–423.
[25] D.A. Silverstein, J.E. Farell, The relationship between image fidelity and image quality, in: IEEE Internat. Conf. on Image Processing, Lausanne, Switzerland, September 1996, pp. 881–884.
[26] S.S. Stevens, On the psychophysical law, Psychol. Rev. 64 (1957) 153–181.
[27] J.M.F. Ten Berge, Orthogonal procrustes rotation for two or more matrices, Psychometrika 42 (1977) 267–276.
[28] P. Teo, D. Heeger, Perceptual image distortion, in: IEEE Internat. Conf. on Image Processing, Austin, TX, November 1994, pp. 982–984.
[29] W.S. Torgerson, Theory and Methods of Scaling, Wiley, New York, 1958.
[30] C.J. van den Branden Lambrecht, Color moving pictures quality metric, in: IEEE Internat Conf. on Image Processing, Vol. I of III, Lausanne, Switzerland, September 1996, pp. 885–888.
[31] C.J. van den Branden Lambrecht, Perceptual models and architectures for video coding applications, Ph.D. Thesis, Ecole Polytechnique Fédérale de Lausanne, Switzerland, 1996.
[32] C. Zetzsche, G. Hauske, Multiple channel model for the prediction of subjective image quality, in: B.E. Rogowitz (Ed.), Human Vision, Visual Processing, and Digital Display, Proc. SPIE, Vol. 1077, 1989, pp. 209–216.