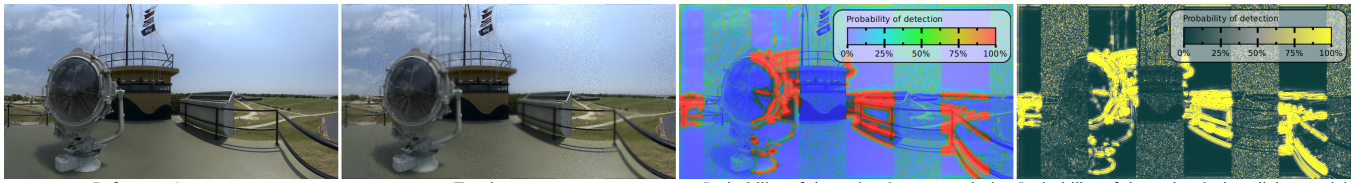
HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions

Rafał Mantiuk*
Bangor University

Kil Joong Kim
Seoul National University

Allan G. Rempel
University of British Columbia

Wolfgang Heidrich
University of British Columbia



Reference Image | Test image | Probability of detection (screen, color) | Probability of detection (print, dichromatic)

Figure 1: *Predicted visibility differences between the test and the reference images. The test image contains interleaved vertical stripes of blur and white noise. The images are tone-mapped versions of an HDR input. The two color-coded maps on the right represent a probability that an average observer will notice a difference between the image pair. Both maps represent the same values, but use different color maps, optimized either for screen viewing or for gray-scale/color printing. The probability of detection drops with lower luminance (luminance sensitivity) and higher texture activity (contrast masking). Image courtesy of HDR-VFX, LLC 2008.*

Abstract

Visual metrics can play an important role in the evaluation of novel lighting, rendering, and imaging algorithms. Unfortunately, current metrics only work well for narrow intensity ranges, and do not correlate well with experimental data outside these ranges. To address these issues, we propose a visual metric for predicting visibility (discrimination) and quality (mean-opinion-score). The metric is based on a new visual model for all luminance conditions, which has been derived from new contrast sensitivity measurements. The model is calibrated and validated against several contrast discrimination data sets, and image quality databases (LIVE and TID2008). The visibility metric is shown to provide much improved predictions as compared to the original HDR-VDP and VDP metrics, especially for low luminance conditions. The image quality predictions are comparable to or better than for the MS-SSIM, which is considered one of the most successful quality metrics. The code of the proposed metric is available on-line.

CR Categories: I.3.0 [Computer Graphics]: General—;

Keywords: visual metric, image quality, visual model, high dynamic range, visual perception

Links: DL PDF WEB

*e-mail: mantiuk@bangor.ac.uk

ACM Reference Format
Mantiuk, R., Kim, K., Rempel, A., Heidrich, W. 2011. HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Trans. Graph. 30*, 4, Article 40 (July 2011), 13 pages. DOI = 10.1145/1964921.1964935 http://doi.acm.org/10.1145/1964921.1964935.

1 Introduction

Validating results in computer graphics and imaging is a challenging task. It is difficult to prove with all scientific rigor that the results produced by a new algorithm (usually images) are statistically significantly better than the results of another state-of-the-art method. A human observer can easily choose which one of the two images looks better; yet running an extensive user study for numerous possible images and algorithm parameter variations is often impractical. Therefore, there is a need for computational metrics that could predict a *visually significant difference* between a test image and its reference, and thus replace tedious user studies.

Visual metrics are often integrated with imaging algorithms to achieve the best compromise between efficiency and perceptual quality. A classical example is image or video compression, but the metrics have been also used in graphics to control global illumination solutions [Myszkowski et al. 1999; Ramasubramanian et al. 1999], or find the optimal tone-mapping curve [Mantiuk et al. 2008]. In fact any algorithm that minimizes root-mean-square-error between a pair of images, could instead use a visual metric to be driven towards visually important goals rather than to minimize a mathematical difference.

The main focus of this work is a calibrated visual model for scenes of arbitrary luminance range. Handling a wide range of luminance is essential for the new high dynamic range display technologies or physical rendering techniques, where the range of luminance can vary greatly. The majority of the existing visual models are intended for very limited luminance ranges, usually restricted to the range available on a CRT display or print [Daly 1993; Lubin 1995; Rohaly et al. 1997; Watson and Ahumada Jr 2005]. Several visual models have been proposed for images with arbitrary dynamic range [Pattanaik et al. 1998; Mantiuk et al. 2005]. However, these so far have not been rigorously tested and calibrated against experimental data. The visual model derived in this work is the result of testing several alternative model components against a set of psychophysical measurements, choosing the best components, and then fitting the model parameters to that data. We will refer to the newly proposed metric as the HDR-VDP-2 as it shares the origins and the HDR capability with the original HDR-VDP [Mantiuk et al. 2005]. However, the new metric and its components constitute a complete overhaul rather than

an incremental change as compared to the HDR-VDP. As with its predecessor, the complete code of the metric is available at `http://hdrvdp.sourceforge.net/`.

The proposed visual model can be used as the main component in the *visual difference predictor* [Daly 1993; Lubin 1995], which can estimate the probability at which an average human observer will detect differences between a pair of images (scenes). Such metrics are tuned towards near-threshold just-noticeable differences. But the straightforward extension of that visual model can also be used for predicting overall *image quality* [Wang and Bovik 2006] for distortions that are much above the discrimination threshold. We show that the proposed quality metric produces results on par with or better than the state-of-the-art quality metrics.

The main contribution of this work is a new visual model that:

- generalizes to a broad range of viewing conditions, from scotopic (night) to photopic (daytime) vision;
- is a comprehensive model of an early visual system that accounts for the intra-ocular light scatter, photoreceptor spectral sensitivities, separate rod and cone pathways, contrast sensitivity across the full range of visible luminance, intra- and inter-channel contrast masking, and spatial integration;
- improves the predictions of a suprathreshold quality metric.

The main limitation of the proposed model is that it predicts only luminance differences and does not consider color. It is also intended for static images and does not account for temporal aspects.

2 Related work

Psychophysical models. Psychophysical measurements have delivered vast amounts of data on the performance of the visual system and allowed for the construction of models of early vision. Although human vision research focuses mostly on simple stimuli such as Gabor patches, there have been several attempts to develop a general visual model for complex images. Two such models that are widely recognized are the Visual Difference Predictor [Daly 1993] and the Visual Difference Metric [Lubin 1995]. More recent research was focused on improving model predictions [Rohaly et al. 1997; Watson and Ahumada Jr 2005], predicting differences in color images [Lovell et al. 2006], in animation sequences [Myszkowski et al. 1999], and high dynamic range images [Mantiuk et al. 2005].

Visual models for tone-mapping. Sophisticated visual models have been proposed in the context of tone-mapping high dynamic range images [Ferwerda et al. 1996; Pattanaik et al. 2000; Pattanaik et al. 1998]. The model of Pattanaik et al. [1998] combines the elements of color appearance and psychophysical models to predict changes in scene appearance under the full range of illumination conditions. However, since these models are intended mostly for visualization, they have not been rigorously tested against the psychophysical and color appearance data and are not intended to be used as visual metrics.

Quality metrics predict subjective judgment about the severity of an image distortion [Wang and Bovik 2006]. They are meant to predict the overall image quality, which is correlated with the results of subjective quality assessment experiments [ITU-R-BT.500-11 2002]. Although the quality measurements vary greatly from psychophysical visual performance measurements, many quality metrics employ visual models similar to those found in the visual difference predictors. However, the recent work in quality assessment favors statistical metrics, such as structural similarity metrics [Wang et al. 2004; Wang et al. 2003].

Feature invariant metrics. The assumption behind structural similarity metrics is that people are more sensitive to certain types of distortions than to others. For example, changes in material and illumination properties of a scene may be noticeable in terms of a just noticeable difference (JND), but non-relevant in terms of overall image quality [Ramanarayanan et al. 2007]. Another example is changes in the shape of a tone-curve, which often remain unnoticed unless they introduce visible contrast distortions. Dynamic-range independent metrics [Aydin et al. 2008; Aydin et al. 2010] rely on the invariance of the visual system to the changes in tone-curve and allow comparing tone-mapped images to a high-dynamic-range reference. These metrics, however, have not been rigorously tested against experimental data and are mostly meant to give good qualitative results in terms of visualized distortion maps, rather than quantitative predictions, such as a mean opinion score or the probability of detection.

3 Visual difference predictor

The overall architecture of the proposed metric, shown in Figure 2, mimics the anatomy of the visual system, but does not attempt to match it exactly. Our first priority was an accurate fit to the experimental data, second the computational complexity, and only then a plausible modeling of actual biological mechanisms.

The visual difference predictor consist of two identical visual models: one each for processing a test image and a reference image. Usually a test image contains and a reference image lacks a feature that is to be detected. For example, for visibility testing it could be a windshield view with and without a pedestrian figure. For measuring compression distortions the pair consists of an image before and after compression.

Since the visual performance differs dramatically across the luminance range as well as the spectral range, the input to the visual model needs to precisely describe the light falling onto the retina. Both the test and reference images are represented as a set of spectral radiance maps, where each map has associated spectral emission curve. This could be the emission of the display that is being tested and the linearized values of primaries for that display. For convenience, we predefined in our implementation several default emission spectra for typical displays (CRT, LCD-CCFL, LCD-RGB-LED) as well as the D65 spectrum for gray-scale stimuli specified in the luminance units of cd/m^2. The pre-defined spectrum for a typical CRT is shown in Figure 4.

The following sections are organized to follow the processing flow shown in Figure 2, with the headings that correspond to the processing blocks.

3.1 Optical and retinal pathway

Intra-ocular light scatter. A small portion of the light that travels through the eye is scattered in the cornea, lens, inside the eye chamber and on the retina [Ritschel et al. 2009]. Such scattering attenuates the high spatial frequencies but more importantly it causes a light pollution that reduces the contrast of the light projected on the retina. The effect is especially pronounced when observing scenes of high contrast (HDR) containing sources of strong light. The effect is commonly known as *disability glare* [Vos and van den Berg 1999] and has been thoroughly measured using both direct measurement methods, such as the double-pass technique [Artal and Navarro 1994], and using psychophysical measurement, such as the equivalent veiling luminance method [van den Berg et al. 1991].

We model the light scatting as a modulation transfer function (**MTF**) acting on the input spectral radiance maps $L[c]$:

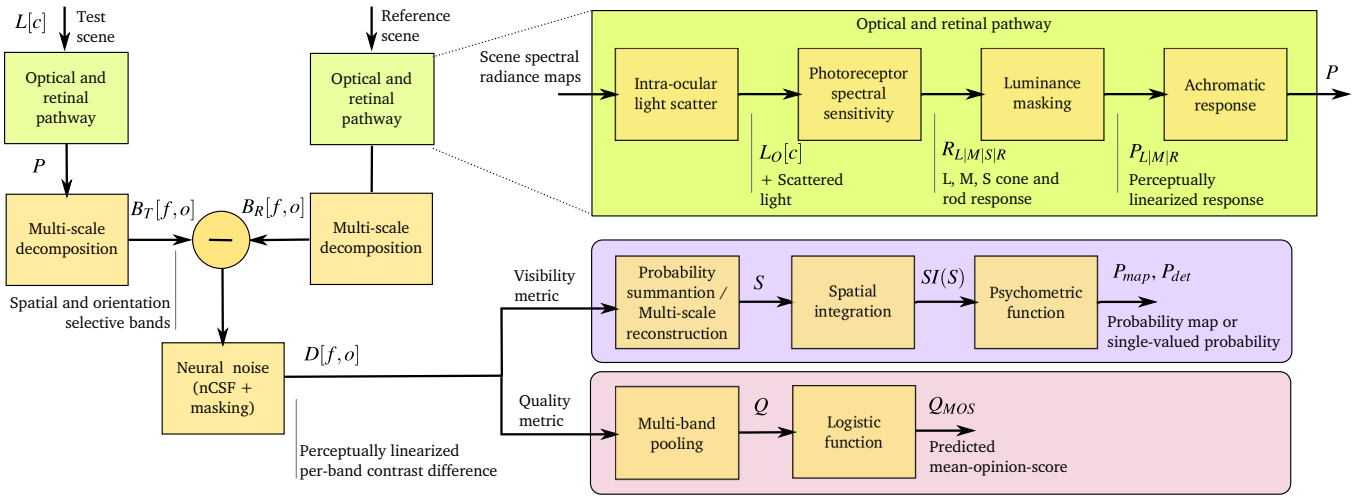


Figure 2: *The block-diagram of the two visual metrics for visibility (discrimination) and quality (mean-opinion-score) predictions and the underlying visual model. The diagram also summarizes the symbols used throughout the paper.*

$$\mathscr{F}\{L_O\}[c] = \mathscr{F}\{L\}[c] \cdot \mathbf{MTF}. \quad (1)$$

The $\mathscr{F}\{\cdot\}$ operator denotes the Fourier transform. For better clarity, we omit pixel or frequency coordinates from the equations and use upper case symbols for images and bold-font symbols for images in the Fourier domain. $[\cdot]$ denotes an index to the set of images, which is the index of the input radiance map, c, in the equation above.

We experimented with several glare models proposed in the literature, including [Ijspeert et al. 1993; Artal and Navarro 1994; Marimont and Wandell 1994; Vos and van den Berg 1999; Rovamo et al. 1998]. We found that only Vos and van den Berg's model [1999] could approximately fit all experimental data (Color glare at scotopic levels, discussed in Section 5), and even that fit was not very good. Vos and van den Berg's model is also defined as the glare-spread-function in the spatial domain, which makes it difficult to use as a digital filter due to the high peak at 0°. To achieve a better match to the data, we fit a generic MTF model, proposed by Ijspeert et al. [1993]:

$$\mathbf{MTF} = \sum_{k=1..4} a_k \, e^{-b_k \rho}, \quad (2)$$

where ρ is the spatial frequency in cycles per degree. The values of all parameters, including a_k and b_k, can be found on the project web-site and in the supplementary materials. Figure 3 shows the comparison of our fitted model with the most comprehensive glare model from the CIE-99 135/1 report [Vos and van den Berg 1999]. To account for the cyclic property of the Fourier transform, we construct an MTF kernel of double the size of an image and we pad the image with the average image luminance or a user supplied surround luminance value.

Note that our MTF is meant to model only low-frequency scattering and it does not predict high frequency effects, such as wavelength dependent chromatic aberrations [Marimont and Wandell 1994] and diffraction, which is limited by the pupil size.

Most studies show little evidence for the wavelength dependency of the intra-occular light scatter [Whitaker et al. 1993] (except for chromatic aberration), and therefore the same MTF can be used for each input radiance map with different emission spectra. A more accurate model could account for a small wavelength dependence

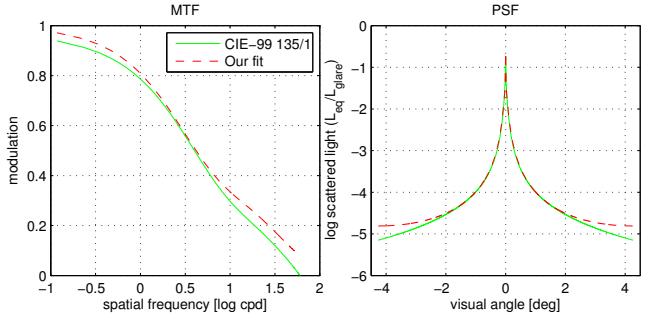


Figure 3: *Comparison of our fitted intra-occular light scatter model with the model from the CIE-99 135/1 report [Vos and van den Berg 1999]. Left panel shows the modulation transfer function of the eye and the right panel its corresponding point spread function. The MTF for the CIE-99 135/1 glare spread function has been computed by creating a densely sampled digital filter and applying the inverse Fourier transform.*

caused by the selective transmission through the iris and the sclera, as reported by van den Berg et al. [1991].

Photoreceptor spectral sensitivity curves describe the probability that a photoreceptor senses a photon of a particular wavelength. Figure 4 shows the sensitivity curves for L-, M-, S-cones, based on the measurements by Stockman and Sharpe [2000], and for rods, based on the data from [CIE 1951]. We use both data sets for our model. When observing light with the spectrum $f[c]$, the expected fraction of light sensed by each type of photoreceptors can be computed as:

$$v_{L|M|S|R}[c] = \int_\lambda \sigma_{L|M|S|R}(\lambda) \cdot f[c](\lambda) d\lambda, \quad (3)$$

where σ is the spectral sensitivity of L-, M-, S-cones or rods, and c is the index of the input radiance map with the emission spectra $f[c]$. We use the index separator $|$ to denote several analogous equations, each with different index letter. Given N input radiance maps, the total amount of light sensed by each photoreceptor type is:

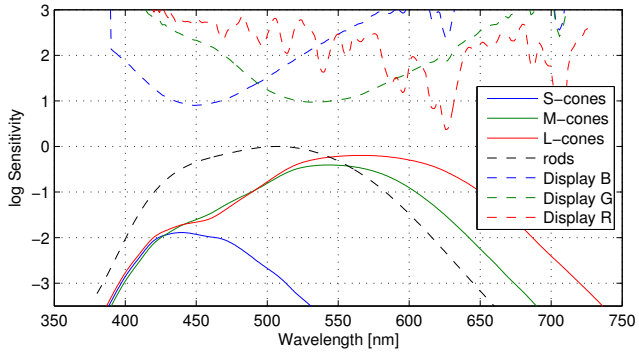


Figure 4: *Spectral sensitivities of the photoreceptors from [Stockman and Sharpe 2000] and [CIE 1951] (bottom). The curves in the upper part of the plot show the measured emission spectra for a CRT display (inverted and arbitrarily scaled to distinguish from the bottom plots).*

$$R_{L|M|S|R} = \sum_{c=1}^{N} L_O[c] \cdot v_{L|M|S|R}[c]. \tag{4}$$

Luminance masking. Photoreceptors are not only selective to wavelengths, but also exhibit highly non-linear response to light. The ability to see the huge range of physical light we owe mostly to the photoreceptors, whose gain control regulates sensitivity according to the intensity of the incoming light. The effect of these regulatory processes in the visual system are often described as *luminance masking*.

Most visual models assume a global (i.e. spatially-invariant) state of adaptation for an image. This is, however, an unjustified simplification, especially for scenes that contain large variations in luminance range (e.g. HDR images). The proposed visual model accounts for the local nature of the adaptation mechanism, which we model using a non-linear transducer function $t_{L|M|R}$:

$$P_{L|M|R} = t_{L|M|R}(R_{L|M|R}), \tag{5}$$

where $P_{L|M|R}$ is a photoreceptor response for L-, M-cones and rods. We omit modeling the effect of S-cones as they have almost no effect on the luminance perception. The transducer is constructed by Fechner's integration [Mantiuk et al. 2005]:

$$t_{L|M|R}(r) = s_{peak} \int_{r_{min}}^{r} \frac{1}{\Delta r_{L|M|R}(\mu)} d\mu = s_{peak} \int_{r_{min}}^{r} \frac{s_{L|M|R}(\mu)}{\mu} d\mu, \tag{6}$$

where r is the photoreceptor absorbed light ($R_{L|M|R}$), r_{min} is the minimum detectable intensity (10^{-6} cd/m^2), $\Delta r(r)$ is the detection threshold, $s_{L|M|R}$ are L-, M-cone, and rod intensity sensitivities. s_{peak} is the adjustment for the peak sensitivity of the visual system, which is the main parameter that needs to be calibrated for each data set (refer to Section 5). The transducer scales the luminance response in the threshold units, so that if the difference between r_1 and r_2 is just noticeable, the difference $t(r_1) - t(r_2)$ equals to 1.

To solve for the transducer functions $t_{L|M|R}$, we need to know how the sensitivity of each photoreceptor type ($s_{L|M|R}$) changes with the intensity of sensed light. We start by computing the combined sensitivity of all the photoreceptor types:

$$s_A(l) = s_L(r_L) + s_M(r_M) + s_R(r_R), \tag{7}$$

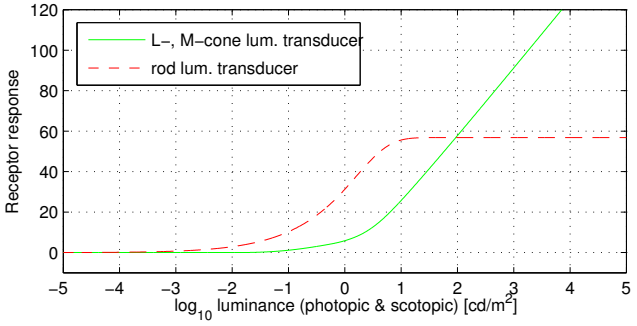


Figure 5: *The luminance transducer functions for cones ($t_L = t_M$) and rods (t_R).*

where l is the photopic luminance. Such combined sensitivity is captured in the CSF function (see Section 4), and is approximated by the peak contrast sensitivity at each luminance level [Mantiuk et al. 2005]:

$$s_A(l) = \max_{\rho} \left(\mathbf{CSF}(\rho, l) \right), \tag{8}$$

where ρ is the spatial frequency and l is adapting luminance. This assumes that any variation in luminance sensitivity is due to photoreceptor response, which is not necessarily consistent with biological models, but which simplifies the computations.

We did not find a suitable data that would let us separately model L- and M-cone sensitivity, and thus we need to assume that their responses are identical: $s_L = s_M$. Due to strong interactions and an overlap in spectral sensitivity, measuring luminance sensitivity of an isolated photoreceptor type for people with normal color vision (trichromats) is difficult. However, there exists data that lets us isolate rod sensitivity, s_R. The data comes from the measurements made for an achromat (person with no cone vision) [Hess et al. 1990, p. 392]. Then the cone sensitivity is assumed to be the difference between the normal trichromat and the achromat contrast sensitivity. Given the photopic luminance $l = r_L + r_M$ and assuming that $r_L = r_M = 0.5l$, we can approximate the L- and M-cone sensitivity as:

$$s_{L|M}(r) = 0.5 \left(s_A(2r) - s_R(2r) \right). \tag{9}$$

The luminance transducer functions for cones and rods, derived from the sensitivity functions, are shown in Figure 5.

The luminance transducer functions $t_{L|M|R}(R)$ make an assumption that the adapting luminance is spatially varying and is equal to the photopic ($R_L + R_M$) or scotopic (R_R) luminance of each pixel. This is equivalent to assuming a lack of spatial maladaptation, which is a reasonable approximation given the finding on spatial locus of the adaptation mechanism [He and MacLeod 1998; MacLeod et al. 1992]. Since numerous mechanisms contribute to overall adaptation (fast neural, slow photo-chemical, pupil contractions) it is difficult to determine the spatial extent of the local adaptation or the minimum size of a feature to which we can adapt. However, the fast adaptation mechanisms, which let us perceive scenes of high dynamic range, are mostly spatially restricted and occur before the summation of cone signals in the horizontal cells [Lee et al. 1999]. In particular, the rapid adaptation mechanism (within $20 - 200$ *ms*) is believed to reside within individual photoreceptors or to operate on signals from individual receptors [He and MacLeod 1998; MacLeod et al. 1992]. Although individual rods exhibit considerable gain changes, the scotopic vision controls its adaptation mainly through post-receptoral mechanisms, lo-

cated presumably in the bipolar cells. The spatial adaptation pool for rods in the human retina is about 10 minutes of arc in diameter [Hess et al. 1990, p. 82], which in most cases is sufficiently small to assume an adaptation luminance equal to $R_L + R_M$ and R_R.

Photoreceptor response is more commonly modeled by an S-shaped function (on a log-linear plot), known as the Michaelis-Menten or Naka-Rushton equation. Such S-shaped behavior, however, can be observed only in the experiments in which a stimulus is briefly flashed after adapting to a particular luminance level, causing maladaptation. Since we assume a stable adaptation condition, there is no need to consider loss of sensitivity due to temporal maladaptation.

Achromatic response. To compute a joint cone and rod achromatic response, the rod and cone responses are summed up:

$$P = P_L + P_M + P_R. \tag{10}$$

The equally weighted sum is motivated by the fact that L- and M-cones contribute approximately equally to the perception of luminance. The contribution of rods, P_R, is controlled by the rod sensitivity transducer t_R so no additional weighting term is necessary. This summation is a sufficient approximation, although a more accurate model should also consider inhibitive interactions between rods and cones.

3.2 Multi-scale decomposition

Both psychophysical masking studies [Stromeyer and Julesz 1972; Foley 1994] and neuropsychological recordings [De Valois et al. 1982] suggest the existence of mechanisms that are selective to narrow ranges of spatial frequencies and orientations. To mimic the decomposition that presumably happens in the visual cortex, visual models commonly employ multi-scale image decompositions, such as wavelets or pyramids. In our model we use the steerable pyramid [Simoncelli and Freeman 2002], which offers good spatial frequency and orientation separation. Similar to other visual decompositions, the frequency bandwidth of each band is halved as the band frequency decreases. The image is decomposed into four orientation bands and the maximum possible number of spatial frequency bands given the image resolution.

We initially experimented with the Cortex Transform [Watson 1987] and its modification [Daly 1993], including the refinements by [Lukin 2009]. The Cortex Transform is used in both the VDP [Daly 1993] and the HDR-VDP [Mantiuk et al. 2005]. However, we found that the spectrally sharp discontinuities where the filter reaches the 0-value cause excessive ringing in the spatial domain. Such ringing introduced a false masking signal in the areas which should not exhibit any masking, making predictions for large-contrast scenes unreliable. The steerable pyramid does not offer as tight frequency isolation as the Cortex Transform but it is mostly free of the ringing artifacts.

3.3 Neural noise

It is convenient to assume that the differences in contrast detection are due to several sources of noise [Daly 1990]. We model overall noise that affects detection in each band as the sum of the signal independent noise (neural CSF) and signal dependent noise (visual masking). If the f-th spatial frequency band and o-th orientation of the steerable pyramid is given as $B_{T|R}[f,o]$ for the test and reference images respectively, the noise-normalized signal difference is

$$D[f,o] = \frac{|B_T[f,o] - B_R[f,o]|^p}{\sqrt{N_{nCSF}^{2p}[f,o] + N_{mask}^2[f,o]}}. \tag{11}$$

The exponent p is the gain that controls the shape of the masking function. We found that the value $p = 3.5$ gives good fit to the data and is consistent with the slope of the psychometric function. The noise summation in the denominator of Equation 11 is responsible for the reduced effect of signal-independent noise, observed as flattening of the CSF function for suprathreshold contrast.

Neural contrast sensitivity function. The signal dependent noise, N_{nCSF}, can be found from the experiments in which the contrast sensitivity function (CSF) is measured. In these experiments the patterns are shown on a uniform field, making the underlying band-limited signal in the reference image equal to 0. However, to use the CSF data in our model we need to discount its optical component that has been already modeled as the MTF of the eye, as well as the luminance-dependent component, which has been modeled as the photoreceptor response. The neural-only part of the CSF is found by dividing it by the MTF of the eye optics (Equation 1) and the joint photoreceptor luminance sensitivity s_A (Equation 7). Since the noise amplitude is inversely proportional to the sensitivity, we get:

$$N_{nCSF}[f,o] = \frac{1}{nCSF[f,o]} = \frac{\mathbf{MTF}(\rho, L_a)\, s_A(L_a)}{\mathbf{CSF}(\rho, L_a)}. \tag{12}$$

ρ is the peak sensitivity for the spatial frequency band f, which can be computed as

$$\rho = \frac{n_{ppd}}{2^f}, \tag{13}$$

where n_{ppd} is the angular resolution of the input image given in pixels per visual degree, and $f = 1$ for the highest frequency band. L_a is adapting luminance, which we compute for each pixel as the photopic luminance after intra-occular scatter (refer to Equation 4): $L_a = R_L + R_M$.

Our approach to modeling sensitivity variations due to spatial frequency assumes a single modulation factor per visual band. In practice this gives a good approximation of the smooth shape of the CSF found in experiments, because the filters in the steerable decomposition well interpolate the sensitivities for the frequencies between the bands. The exception is the lowest frequency band (base-band), whose frequency range is too broad to model sensitivity differences for very low frequencies. In case of the base-band, we filter the bands in both the test and reference images with the nCSF prior to computing the difference in Equation 11 and set $N_{nCSF} = 1$. For the base-band we assume a single adapting luminance equal to the mean of L_a, which is justified by a very low resolution of that band.

Some visual models, such as the VDP or the HDR-VDP, filter an image by the CSF or nCSF before the multi-scale decomposition. This, however, gives worse per-pixel control of the CSF shape. In our model the CSF is a function of adapting luminance, L_a; thus the sensitivity can vary greatly between pixels in the same band due to different luminance levels.

Contrast masking. The signal-dependent noise component N_{mask} models contrast masking, which causes lower visibility of small differences added to a non-uniform background. If a pattern is superimposed on another pattern of similar spatial frequency and orientation, it is, in the general case, more difficult to detect [Foley 1994]. This effect is known as *visual masking* or *contrast masking* to differentiate it from *luminance masking*. Figure 11 (left) shows a typical characteristic obtained in the visual masking experiments together with the fit from our model. The curves show that if a masking pattern is of the same orientation and spatial frequency as the target (*intra-channel masking*, 0°), the target detection threshold first decreases (*facilitation*) and then gets elevated (*masking*) with increasing masker contrast. The facilitation, however, disappears when the masker is of different orientation (*inter-channel masking*, 90°).

Inter-channel masking is still present, but has lower impact than in the intra-channel case. Although early masking models, including those used in the VDP and the HDR-VDP, accounted mostly for the intra-channel masking, findings in vision research give more support to the models with wider frequency spread of the masking signal [Foley 1994; Watson and Solomon 1997]. We follow these findings and integrate the activity from several bands to find the masking signal. This is modeled by the three-component sum:

$$\begin{aligned} N_{mask}[f,o] = & \frac{k_{self}}{n_f}\left(n_f B_M[f,o]\right)^q + \\ & \frac{k_{xo}}{n_f}\left(n_f \sum_{i=O\setminus\{o\}} B_M[f,i]\right)^q + \\ & \frac{k_{xn}}{n_f}\left(n_{f+1} B_M[f+1,o] + n_{f-1} B_M[f-1,o]\right)^q, \end{aligned} \tag{14}$$

where the first line is responsible for self-masking, the second for masking across orientations and the third is the masking due to two neighboring frequency bands. k_{self}, k_{xo} and k_{xn} are the weights that control the influence of each source of masking. O in the second line is the set of all orientations. The exponent q controls the slope of the masking function. The biologically inspired image decompositions, such as the Cortex Transform [Watson 1987], reduce the energy in each lower frequency band due to a narrower bandwidth. To achieve the same result with the steerable pyramid and to ensure that all values are in the same units before applying the non-linearity q, the values must be normalized by the factor

$$n_f = 2^{-(f-1)}. \tag{15}$$

The $B_M[f,o]$ is the activity in the band f and orientation o. Similarly as in [Daly 1993] we assume *mutual-masking* and compute the band activity as the minimum from the absolute values of test and reference image bands:

$$B_M[f,o] = min\left\{|B_T[f,o]|, |B_R[f,o]|\right\} nCSF[f,o]. \tag{16}$$

The multiplication by the N_{nCSF} unifies the shape of the masking function across spatial frequencies, as discussed in detail in [Daly 1993].

3.4 Visibility metric

Psychometric function. Equation 11 scales the signal in each band so that the value $D = 1$ corresponds to the detection threshold of a particular frequency- and orientation-selective mechanism. The values $D[f,o]$ are given in contrast units and they need to be transformed by the psychometric function to yield probability values P:

$$P[f,o] = 1 - \exp(\log(0.5)\, D^{\beta}[f,o]), \tag{17}$$

where β is the slope of the psychometric function. Although the commonly reported value of the masking slope is $\beta = 3.5$ [Daly 1993], we need to account for the masking gain control p in Equation 11 and set it to $\beta = 3.5/p = 1$. The constant $\log(0.5)$ is introduced in order to produce $P = 0.5$ when the contrast is at the threshold ($D = 1$).

Probability summation. To obtain the overall probability for all orientation- and frequency-selective mechanisms it is necessary to sum all probabilities across all bands and orientations. The probability summation is computed as in [Daly 1993]

$$P_{map} = 1 - \prod_{(f,o)} (1 - P[f,o]). \tag{18}$$

After substituting the psychometric function from Equation 17, we get:

$$\begin{aligned} P_{map} &= 1 - \prod_{(f,o)} \exp(\log(0.5) D^{\beta}[f,o]) \\ &= 1 - \exp\left(\log(0.5) \sum_{(f,o)} D^{\beta}[f,o]\right). \end{aligned} \tag{19}$$

It is important to note that the product has been replaced by summation. This lets us use the reconstruction transformation of the steerable pyramid to sum up probabilities from all bands. Such reconstruction involves pre-filtering the signal from each band, upsampling and summing up all bands, and is thus the counterpart of the sum of all band-differences in Equation 19. Therefore, in practice, instead of the sum we use the steerable pyramid reconstruction $\mathscr{P}^{-1}$ on the differences with the exponent equal to the psychometric function slope β:

$$P_{map} = 1 - \exp\left(\log(0.5)\, SI(\mathscr{P}^{-1}(D^{\beta}))\right), \tag{20}$$

where SI is the spatial integration discussed below.

P_{map} gives a spatially varying map, in which each pixel represents the probability of detecting a difference. To compute a single probability for the entire image, for example to compare the predictions to psychophysical data, we compute the maximum value of the probability map: $P_{det} = \max\{P_{map}\}$. Such a maximum value operator corresponds to the situation in which each portion of an image is equally well attended and thus the most visible difference constitutes the detection threshold. A similar assumption was also used in other studies [Daly et al. 1994].

Spatial integration. Larger patterns are easier to detect due to *spatial integration*. Spatial integration acts upon a relatively large area, extending up to 7 cycles of the base frequency [Meese and Summers 2007]. Since our data does not capture the extent of the spatial integration, we model the effect as the summation over the entire image

$$SI(S) = \frac{\sum S}{\max(S)} \cdot S, \tag{21}$$

where $S = \mathscr{P}^{-1}(D^{\beta})$ is the contrast difference map from Equation 20. The map S is modulated by the effect of stimuli size (the fraction in the equation) so that the maximum value, which is used for the single probability result P_{det}, is replaced with the sum. The sum should be interpreted as another probability summation, which acts across the spatial domain rather than across the bands. Such a summation is consistent with other models, which employ spatial pooling as the last stage of the detection model [Watson and Ahumada Jr 2005].

The simple model above is likely to over-predict the spatial integration for suprathreshold stimuli, which is found to decline when the pattern is masked by another pattern of the same characteristic (intra-channel masking), but is still present when the masker has a different characteristic (inter-channel masking). Meese and Summers [2007] summarized these findings and demonstrated that they can be explained by the models of visual masking. However, we found that their model cannot be used to detect more than one target and thus it is not suitable for predicting differences in complex images.

3.5 Implementation details

Visualization. Obtaining a single-valued probability of detection is important for many applications, but it is often necessary to inspect

how the distortions are distributed across an image. For this purpose we created several visualization schemes that use color maps to represent the probability of detection for each pixel (P_{map}). Two examples of such a visualization are shown in Figure 1. If good color reproduction is possible (i.e. the maps are shown on a display), the color-coded probability map is superimposed on top of a context image, which is a luminance-only, contrast-compressed version of the test image. The context image can be disabled if the map needs to be readable on a gray-scale print-out and lightness variations must represent the probabilities. Both tri- and dichromatic color maps are available, where the latter type reduces ambiguities for color deficient observers.

Timing. The run-time of the metric is linear in the number of image pixels, taking 16 seconds to compare 1M pixel images using unoptimized matlab code on one core of a 2.8 GHz CPU (see supplementary).

4 Contrast sensitivity function

The contrast sensitivity function (CSF) is the main reference in our model for visual performance across the luminance range. It determines luminance masking, static noise in cortical bands, and it normalizes the masking signal across bands. Therefore, it is essential to use an accurate model of contrast sensitivity.

We experimented with the two most comprehensive CSF models, proposed by Daly [1993] and Barten [1999]. But we found that they give poor fits to both our experimental data and other data sets, including ModelFest [Watson and Ahumada Jr 2005]. For that reason we decided to fit a custom CSF that is more suitable for predicting visible differences. This does not imply that the CSF models by Daly and Barten are less accurate, but rather that their functions may capture conditions that are different from visual inspection of static images.

As it is not the focus of this work, we only briefly describe the experiment in which we measured CSF for a large range of luminance. We made these measurements because we found available psychophysical data either incomplete, for example lacking the lower frequency part or the full range of luminance levels, or the conditions used for measurements were very different from the conditions in which images are compared. For example, most CSF measurements are conducted for an artificial pupil and a short flicker, which differs greatly from the conditions in which images are usually compared. We also did not want to combine measurements from several studies as they often use very different experimental conditions and protocols.

The stimuli consisted of vertical sine-gratings attenuated by the Gaussian envelope. The σ of the Gaussian constituted size of the object, which was 1.5 visual deg. for most stimuli. To collect data on spatial integration, additional sizes of 0.5 and 0.15 visual deg. were measured for 1 and 8 cycles-per-degree (cpd). The stimuli design was inspired by the ModelFest data set. The background luminance varied from 0.02 to 150 cd/m^2. The luminance levels below 10 cd/m^2 were achieved by wearing modified welding goggles in which the protective glass was replaced with neutral density filters (Kodak Wratten Gelatin) with either 1.0 or 2.0 density value. Although the maximum tested background luminance is only 150 cd/m^2, the changes of the CSF shape above that level are minimal (compare the plots for 20 and 150 cd/m^2 in Figure 6). The frequency range of the sine grating varied from from 0.125 to 16 cpd (cycles per degree). The stimuli were shown on a 24" LCD display with 10-bit panel and RGB LED backlight (HP LP2480zx). Two additional bits were simulated by spatio-temporal dithering so that the effective bit-depth was 12 bits per color channel. Stimuli were

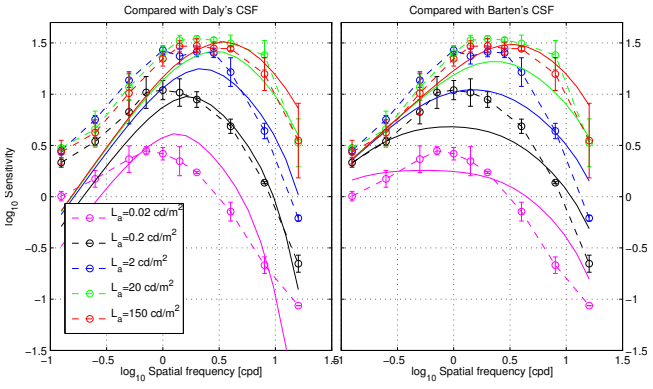


Figure 6: *The CSF measurement results compared to two popular CSF models by Daly [1993] and Barten [1999]. The Dashed lines represent our measurements and the solid lines the two compared models. The model parameters, including stimuli size, were set to match our measurements. The predictions of both models differ from the measurements probably because of different experiment conditions.*

observed from a fixed distance of 93 cm, which gave an angular resolution of 60 pixels per visual degree. The display was calibrated using a photo-spectrometer. The display white point was fixed at D65.

The procedure involved a 4-alternative-forced-choice (4AFC) experiment in which an observer was asked to choose one of the four stimuli, of which only one contained the pattern. We found 4AFC more efficient and faster in convergence than 2AFC because of the lower probability of correct guesses. The stimuli were shown side-by-side on the same screen and the presentation time was not limited. We used this procedure as more appropriate for the task of finding differences in images than temporal flicker intervals used in most threshold measurements. The QUEST procedure [Watson and Pelli 1983] with a fixed number of trials (from 20 to 30, depending on the observer experience) was used to find the threshold. The data was collected for five observers. Each observer completed all the tests in 3–4 sessions of 30–45 minutes.

The results of the experiment compared to the CSF models by Daly and Barten are shown in Figure 6. The effect of stimuli size is shown in Figure 8 together with the fit of the full visual model. Figure 6 shows that Daly's CSF model predicts much lower sensitivity for low-frequency patterns. Both Daly's and Barten's CSF models predict much lower sensitivity for 2 cd/m^2 and a much larger sensitivity difference between 20 cd/m^2 and 150 cd/m^2. The inconsistency between the models and the data result in poor predictions for those luminance levels. Since adjusting the model parameters did not improve the fits, we decided to fit a new CSF model, which is specifically intended for visual difference metrics.

The CSF model is based on a simplified version of Barten's CSF [Barten 1999, Eq. 3.26]:

$$\mathbf{CSF}(\rho) = p_4 s_A(l) \frac{\mathbf{MTF}(\rho)}{\sqrt{(1+(p_1\rho)^{p_2}) \cdot \left(1-e^{-(\rho/7)^2}\right)^{-p_3}}}, \quad (22)$$

where ρ is the spatial frequency in cycles-per-degree and $p_{1...4}$ are the fitted parameters. The parameters are fitted separately for each adaptation luminance L_a (see supplementary). For the luminance values in between the measured levels we interpolate the parameters using the logarithmic luminance as the interpolation coefficient.

$s_A(l)$ is the joint luminance-sensitivity curve for cone and rod photoreceptors, and is given in Equation 8. This sensitivity is modeled as

$$s_A(l) = p_5 \left(\left(\frac{p_6}{l} \right)^{p_7} + 1 \right)^{-p_8}. \quad (23)$$

The parameters p_4 and p_5 are adjusted so that the CSF divided by the $s_A(l)$ peaks at 1. This let us use s_A directly in Equation 9. The model also yields *nCSF* value needed in Equation 12 when MTF and s_A are set to 1.

5 Calibration and validation

The value of the visual model largely depends on how well it can predict actual experimental data. Therefore, we took great care to fit the model to available and relevant psychophysical data. We also ran extensive experiments to capture critical characteristics of the visual system.

Since the model is not invertible, the *calibration* involves an iterative optimization (the simplex search method), in which parameters are adjusted until the best prediction for a threshold data set is found ($SI(S)$ closest to 1). Each pair of images in such a data set is generated so that the contrast between them is equal to the detection or discrimination threshold.

To *validate* each data fitting, we use a different procedure, in which contrast between test and reference images is reduced or amplified until the metric results in $P_{det} = 0.5$. Then, the factor by which the contrast has been modified is considered as a metric error. Both calibration and validation are computationally expensive procedures, which require running the full metric thousands of times. Therefore, the calibration was performed on a cluster of about 10 CPU cores. The running time varied from an hour to a few days, depending on the complexity of the calibration task.

Following [Watson and Ahumada Jr 2005] we report fitting error as a root-mean-square-error (RMSE) of the contrast difference between the mean measurement and the model prediction given in dB contrast units[1]. For the data sets which provided information about the distribution of the measurements we also report the χ^2 statistics. Intuitively, values of χ^2 close or below 1 indicate good fit to the experimental data.

Given the degrees of freedom that the visual metric offers, it is relatively easy to fit each data set separately. However, the main challenge of this work is to get a good fit for *all* data sets using *the same* calibration parameters. The only parameter that was adjusted between calibration and validation steps was the peak sensitivity, s_{peak} (Equation 6), which usually varies between data sets due to differences in experimental procedures or individual variations [Daly et al. 1994]. The values of the peak sensitivity parameter as well as more detailed results can be found in the supplementary materials.

Several measures have been taken to reduce the risk of over-fitting the model. Each component of the model (CSF, OTF, masking) was fitted separately using the data that isolated a particular phenomenon. For example, the glare data can be predicted by both the nCSF and the MTF, but is relevant only for the model of the MTF. This measure also reduced the degrees of freedom that need to be calibrated for each fitting. For all fittings the ratio of data points to the degrees of freedom was at least 4:1.

The images used for calibration and testing were either recreated using the data from original publications or generated from the experimental stimuli modulated by the contrast detection threshold.

[1] dB contrast = $20 \cdot \log_{10}(\Delta L / L)$.

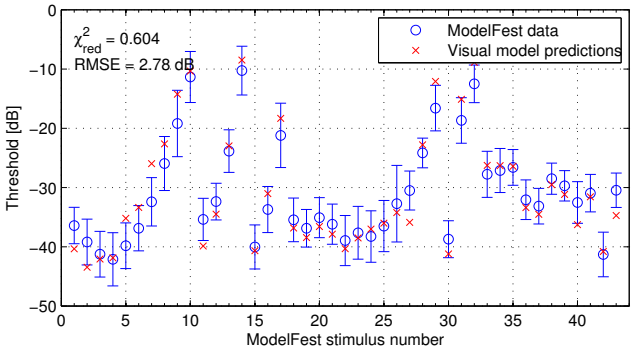


Figure 7: *Visual model predictions for the* **ModelFest** *data set. Error bars denote standard deviation of the measurements. The R value is the prediction mean square root error and χ^2_{red} is the reduced chi-square statistic.*

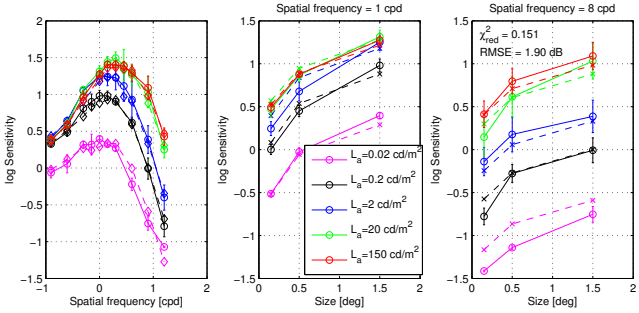


Figure 8: *Visual model predictions for the* **CSF for wide luminance range** *data set. The two plots on the right show sensitivity variation due to stimuli size.*

In all cases the test and reference images are stored as OpenEXR images using 32-bit floating point pixel format. Some more details on the stimuli can be found in the supplementary materials.

ModelFest is a standard data set created to calibrate and validate visual metrics, containing 43 small detection targets at 30 cd/m^2 uniform background [Watson and Ahumada Jr 2005]. Figure 7 shows the predictions of the proposed metric for this data set. The mean prediction error is 2.8 dB, which is higher than for the models explicitly calibrated for the ModelFest data (≈1 dB, refer to [Watson and Ahumada Jr 2005]). However this value is still much lower than the standard deviation of the ModelFest measurements (3.79 dB).

CSF for wide luminance range. Since the ModelFest stimuli are shown on a 30 cd/m^2 background only, they are not sufficient to calibrate our metric, which needs to work for all luminance levels. We used our CSF measurements, which are discussed in Section 4, to validate detection across the luminance and frequency range, as well as to test the spatial integration (Equation 21). As shown in Figure 8, the metric well predicts the loss of sensitivity due to frequency, luminance and size variations. Although good predictions can be expected as this data set was used to derive the CSF for the visual model, this test checks the integrity of the entire metric. The plot reveals some prediction fluctuations throughout frequencies, which are caused by the non-linearities applied to the signal after the steerable-pyramid decomposition. Such non-linearities affect the signal that is split into two separate bands and then reconstructed.

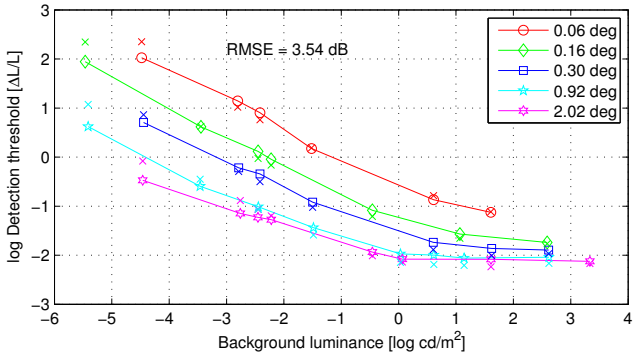


Figure 9: *Visual model predictions for the* **threshold versus intensity curve** *data set. These are the detection thresholds for circular patterns of varying size on a uniform background field [Blackwell 1946].*

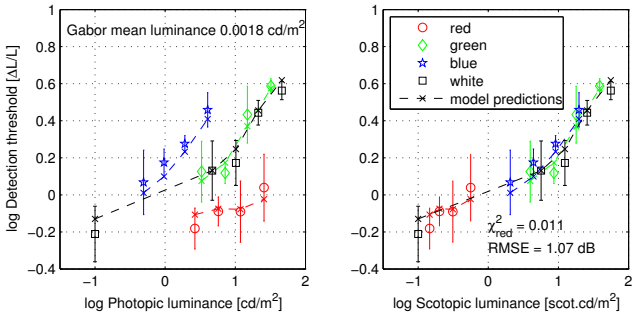


Figure 10: *Visual model predictions for the* **color glare at scotopic levels** *data set. The leftmost point represent the threshold without glare source.*

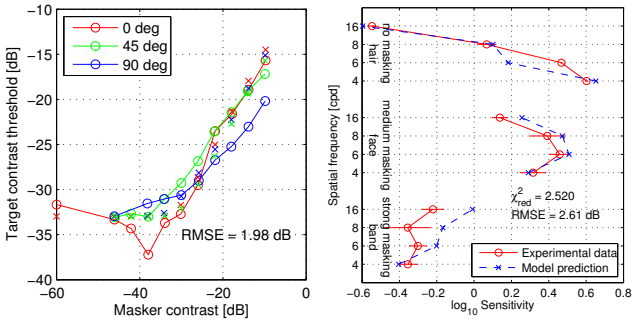


Figure 11: *Visual model predictions for the* **Foley's masking data** *[Foley 1994] on the left and* **CSF flattening** *data set on the right. The degree values in the legend correspond to the orientation of the masking pattern with respect to the test pattern (0 deg. - vertical).*

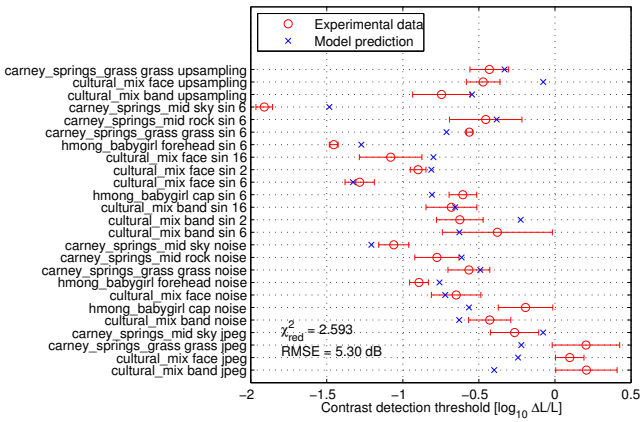


Figure 12: *Visual model predictions for the* **complex images** *data set.*

Threshold versus intensity curve. One of the most classical measurements of the threshold variation with adapting luminance was performed by Blackwell and his colleagues [Blackwell 1946]. In a laboratory built for the purpose of these measurements, over 400,000 observations were recorded and manually analyzed to determine detection thresholds for circular disks of different sizes (from 0.06 to 2° diameter) shown on a uniform adapting field (from 10^{-5} to $10^{3.5}$ cd/m^2). This is an excellent validation set for our model because it both covers a large luminance range and was measured for circular stimuli, which is very different than the sine gratings on which the CSF is based. Figure 9 shows that our model accounts well for both luminance and size variations, which confirms a good choice of sensitivity curves $s_{A|R}$. Additionally, it indicates that a CSF-based model with probability summation well integrates complex stimuli data with multiple spatial frequencies. We also observed that the fit improved after introducing separate rod and cone pathways.

Color glare at scotopic levels. This data set validates metric predictions for rod-mediated vision in the presence of colored glare [Mantiuk et al. 2009], and was used to adjust the intra-ocular light scatter parameters (Equation 2). The stimuli are 1 cpd dark Gabor patches (0.002 cd/m^2) seen in the presence of a strong source of glare. The prototype display with individually controlled red, green and blue LED backlight was used to create narrow bandwidth glare light. The advantage of this is that the metric can be tested for spectral sensitivity of both rod and cone photoreceptors. The predictions compared to the experimental results shown in Figure 10 demonstrate that the metric correctly predicts the wavelength dependence on glare in the scotopic range.

Foley's masking measurements [Foley 1994] were used to calibrate the inter-channel masking mechanism as well as validate the choice of masking non-linearity. Figure 11 (left) shows that the metric follows the masking slopes for both in-channel masking, where test and mask sine gratings share the same orientation (0°), and inter-channel masking, where test and mask gratings differ in orientation by 90°. The predictions generated by the metric do not change the shape of the masking function with the making signal orientation, which is the limitation of the current masking model. We also do not model facilitation that is causing the 'dip' in the 0 deg curve. Georgeson and Georgeson [1987] found that the facilitation disappears with variation in phase or temporal offset. This effect is regarded as fragile and absent in complex images [Daly 1993].

CSF flattening is observed for contrasts above the detection threshold [Georgeson and Sullivan 1975] and is an essential characteristic for any visual model that needs to predict visibility in complex images. To capture this effect we measured the detection threshold for Gabor patches from 4 to 16 cpd superimposed on an actual image (portrait, see supplementary) in three different regions: the region with almost no masking (hair), with moderate masking (face) and with strong masking (band). As shown in Figure 11 (right) the sharp decrease in contrast sensitivity at high frequencies in uniform

region (hair) is even reversed for strong masking (band). The model correctly reverses the shape of the CSF, though the reversal is exaggerated for the high masking region. This data set captures a very important characteristic that is often missing in visual models. The visual model that relies on a CSF alone would make worse predictions for strongly masked regions than the visual model without any CSF weighting.

Complex images. To test the metric prediction in likely scenarios, we measured thresholds for the data set, in which several types of distortions (biliniear upsampling or blur, sinusoidal grating, noise and JPEG compression distortions) were superimposed in complex images. This is the most demanding data set and it produces the highest errors for our metric, as shown in Figure 12. This data set is especially problematic as it does not isolate the effects that would produce a systematic variation in the relevant parameter space (spatial frequency, masking signal, etc.), and thus does not allow to identify the potential cause for lower accuracy. We assume that the prediction error is the net result of all the mechanisms that we do not model exactly: inter-channel masking, spatial pooling, the signal-well-known effect [Smith Jr and Swift 1985], as well as a multitude of minor effects, which are difficult, if possible, to model. Despite these problems, the overall metric predictions follow most data points and demonstrate the good performance of the metric for complex images.

5.1 Comparison with visibility metrics

	VDP'93	HDR-VDP 1.7	**HDR-VDP-2**
ModelFest	4.32 *dB*, (1.4)	3.3 *dB*, (0.75)	2.78 *dB*, (0.6)
CSF lum. range	15 *dB*, (9.4)	7.53 *dB*, (2.4)	1.9 *dB*, (0.15)
Blackwell's t.v.i.	27.5 *dB*	41.2 *dB*	3.54 *dB*
Glare at scotopic lum.	15 *dB*, (2.2)	2.68 *dB*, (0.071)	1.07 *dB*, (0.011)
Foley's masking data	7.73 *dB*	7.07 *dB*	1.98 *dB*
CSF flattening	3.95 *dB*, (5.8)	5.97 *dB*, (13)	2.61 *dB*, (2.5)
Complex images	7.05 *dB*, (4.6)	7.01 *dB*, (4.5)	5.3 *dB*, (2.6)

Table 1: *Prediction error for the HDR-VDP-2 compared to the VDP [Daly 1993] and the HDR-VDP [Mantiuk et al. 2005]. The values represent the root-mean-square-error and the χ^2_{red} statistic in parenthesis (where information is available).*

We compare HDR-VDP-2 predictions with two other metrics: the Visual Difference Predictor (VDP), which we reimplemented based on the book chapter [Daly 1993] and after some correspondence with the author; and with the visual difference predictor for high dynamic range images (HDR-VDP 1.7), for which we use the publicly available C++ code from `http://hdrvdp.sourceforge.net/`. The same data sets are used for comparison as for the calibration and validation described in the previous section. The peak sensitivity parameter of each metric was adjusted individually for each data set. Additionally, we optimized the masking slope of both the VDP and the HDR-VDP for the *complex images* data set and found a value of 0.9 optimal for the VDP and the original masking slope 1.0 to be the best for the HDR-VDP.

The results of the comparisons are summarized in Table 1 (see supplementary for more details). The new CSF function, described in Section 4, improved the *ModelFest* predictions as compared to the VDP and the HDR-VDP. But the most noticeable improvement is for the *CSF for wide luminance range* and *Blackwell's t.v.i* data sets, which revealed problems with low luminance predictions in both the VDP and the HDR-VDP. The prediction errors are especially large for the VDP and a luminance lower than 0.1 cd/m^2, as this metric was not intended to work in that luminance range.

Both the HDR-VDP and the HDR-VDP-2 handle predictions for *glare at scotopic luminance* levels relatively well, though the new model predictions are better due to separate cone and rod pathways. The poor prediction for low luminance and the lack of glare model make the VDP unsuitable for this data set.

The improvement in masking predictions is most noticeable for *Foley's masking* data set, as both VDP and HDR-VDP do not predict inter-channel masking. Although the non-linearities in *VDP* compensate to some extent changes in the shape of CSF due to masking signal, the lack of *CSF flattening* is an issue in HDR-VDP. The new model also improves the predictions for the *complex images* data set containing a wide range of image distortions, though the improvement is not well quantified due to limited number of test images.

6 Predicting image quality

Sometimes it is more important to know how a visual difference affects overall image quality, than to know that such a difference exists. The subjective severity of visual difference is usually measured by quality metrics, which quantify the visual distortion with a single value of quality score. Such a quality score can be measured in subjective experiments in which a large number of observers rate or rank images [ITU-R-BT.500-11 2002]. The automatic (objective) metrics attempt to replace tedious experiments with computational algorithms.

The HDR-VDP-2 has been designed and calibrated to predict visibility rather than quality. However, in this section we demonstrate that the metric can be extended to match the performance of state-of-the-art quality metrics.

6.1 Pooling strategy

The goal of the majority of quality metrics is to perceptually linearize the differences between a pair of images, so that the magnitude of distortion corresponds to visibility rather than mathematical difference between pixel values. The HDR-VDP-2 achieves this goal when computing the threshold-normalized difference for each band $D[f,o]$ (Equation 11). However, this gives the difference value for each pixel in each spatially- and orientation-selective bands. The question is how to pool the information from all pixels and all bands to arrive at a single value predicting image quality.

To find the best pooling strategy, we tested over 20 different combinations of aggregating functions and compared the predictions against two image quality databases: LIVE [Sheikh et al. 2006] and TID2008 [Ponomarenko et al. 2009]. The aggregate functions included maximum value, percentiles (50, 75, 95) and a range of power means (normalized Minkowski summation) with the exponent ranging from 0.5 to 16. Each aggregating function was computed on the linear and logarithmic values. As the measure of prediction accuracy we selected Spearman's rank order correlation coefficient as it is not affected by non-linear mapping between subjective and objective scores. The pooling function that produced the strongest correlation with the quality databases was:

$$Q = \frac{1}{F \cdot O} \sum_{f=1}^{F} \sum_{o=1}^{O} w_f \log \left(\frac{1}{I} \sum_{i=1}^{I} D^2[f,o](i) + \varepsilon \right), \qquad (24)$$

where i is the pixel index, ε is a small constant (10^{-5}) added to avoid singularities when D is close to 0, and I is the total number of pixels. Slightly higher correlation was found for exponents greater than 2, but the difference was not significant enough to justify the use of a non-standard mean. The per-band weighting w_f was set to 1 to compare different aggregating functions, but then was optimized using the simulated annealing method to produce the highest

correlation with the LIVE database. The weights that maximize correlation are listed in the supplementary materials.

The objective quality predictions do not map directly to the subjective mean opinion scores (MOS) and there is a non-linear mapping function between subjective and objective predictions. Following ITU-R-BY.500.11 recommendations [2002], we fit a logistic function to account for such a mapping:

$$Q_{MOS} = \frac{100}{1+\exp(q_1\,(Q+q_2))}. \quad (25)$$

The LIVE database was used to fit the logistic function and to find the per-band weights w_f, while the TID2008 database was used only for testing. The TID2008 database contains a larger number of distortions and is a more reliable reference for testing quality metrics, but the distortions are mostly concentrated in the central part of the MOS scale with a lower number of images of perfect or very poor quality. This biases fitting results toward better predictions only in the central part of the MOS scale.

Since the HDR-VDP-2 operates on physical units rather than pixel values, in all experiments we converted LIVE and TID2008 database images to trichromatic XYZ values assuming a standard LCD display with CCFL backlight, sRGB color primaries, 2.2 gamma, 180 cd/m^2 peak luminance and 1 cd/m^2 black level.

6.2 Comparison with quality metrics

We compared the quality predictions of the HDR-VDP-2 with several state-of-the-art quality metrics, including Structural Similarity Index (SSIM) [Wang et al. 2004], its multi-scale extension (MS-SSIM) [Wang et al. 2003], mDCT-PSNR [Richter 2009] and still the most commonly used, the PSNR metric. The recent comprehensive comparison of quality metrics against the TID2008 database found the MS-SSIM to be the most accurate [Ponomarenko et al. 2009]; thus we compare our predictions with the best available method. Figure 13 shows the correlation between each metric and the subjective DMOS scores from the LIVE and TID2008 quality databases. The continuous lines correspond to the logistic mapping function from Equation 25. The HDR-VDP-2 has the highest Spearman's correlation coefficient for both databases, which means that it gives the most accurate ranking of images. The HDR-VDP-2 also ranks first in terms of RMSE for the LIVE database, but second for the TID2008 database, for which MS-SSIM produces smaller error. The correlation coefficients are also the highest for individual distortions (see supplementary), suggesting that our metric is good at ranking each distortion type individually, but is less successful at differentiating the quality between the distortions. The differences in the correlation coefficient and RMSE for the HDR-VDP-2 and MS-SSIM are statistically significant at $\alpha = 0.05$.

Unlike MS-SSIM, the HDR-VDP-2 can account for viewing conditions, such as display brightness or viewing distance. It can also measure quality for the scenes outside the luminance range of typical LCD or CRT displays (though such the predictions have not been validated for quality). Given that, the HDR-VDP-2 is a good alternative to MS-SSIM for all applications that require finer control of the viewing parameters.

7 Conclusions and future work

In this work we demonstrated that the new HDR-VDP-2 metric based on a well calibrated visual model can reliably predict visibility and quality differences between image pairs. The predictions for the tested data sets are improved or at least comparable to the state-of-the-art visibility (discrimination) and quality metrics. The underlying visual model is specialized to predict differences outside the luminance range of a typical display (1–100 cd/m^2), which is important for the new high dynamic range display technologies and the range of applications dealing with real-world lighting. This work also stresses the importance of validating visual models against experimental data.

The underlying visual model of the HDR-VDP-2 can be used in combination with higher level visual metrics, such as dynamic range independent metrics [Aydin et al. 2008; Aydin et al. 2010]. The detection component of these metrics, which relies on the HDR-VDP, can be easily replaced with the HDR-VDP-2 detection model.

The HDR-VDP-2 is a step towards a better visibility and quality predictor, but there is still room for improvement. The two major omissions are the modelling of color vision and temporal processing. Including the spatio-velocity and spatio-temporal components, as done in [Myszkowski et al. 1999; Aydin et al. 2010], could provide an extension to the temporal domain. The existing achromatic model would benefit from a better model of spatial integration, improved sensitivity characteristics for each photoreceptor type, and a refined masking model, which is calibrated to a more extensive data set. The metric could also consider a less conservative assumption when the distortion signal is not known exactly [Smith Jr and Swift 1985].

Acknowledgements

We wish to thank Scott Daly, Karol Myszkowski and the anonymous reviewers for their insightful comments. We also wish to thank all volunteers who participated in our experiments. This work was partly supported by the EPSRC research grant EP/I006575/1.

References

ARTAL, P., AND NAVARRO, R. 1994. Monochromatic modulation transfer function of the human eye for different pupil diameters: an analytical expression. *J. Opt. Soc. Am. A 11*, 1, 246–249.

AYDIN, T. O., MANTIUK, R., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2008. Dynamic range independent image quality assessment. *ACM Trans. on Graphics (SIGGRAPH'08) 27*, 3, 69.

AYDIN, T. O., ČADÍK, M., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2010. Video quality assessment for computer graphics applications. *ACM Trans. Graph. 29*, 161:1–161:12.

BARTEN, P. G. J. 1999. *Contrast sensitivity of the human eye and its effects on image quality*. SPIE Press.

BLACKWELL, H. 1946. Contrast thresholds of the human eye. *Journal of the Optical Society of America 36*, 11, 624–632.

CIE. 1951. In *CIE Proceedings*, vol. 1, 37.

DALY, S., CO, E., AND ROCHESTER, N. 1994. A visual model for optimizing the design of image processingalgorithms. In *Proc. of IEEE ICIP*, vol. 2, 16–20.

DALY, S. 1990. Application of a noise-adaptive contrast sensitivity function to image data compression. *Optical Engineering 29*, 08, 977–987.

DALY, S. 1993. *Digital Images and Human Vision*. MIT Press, ch. The Visible Differences Predictor: An Algorithm for the Assessment of Image Fidelity, 179–206.

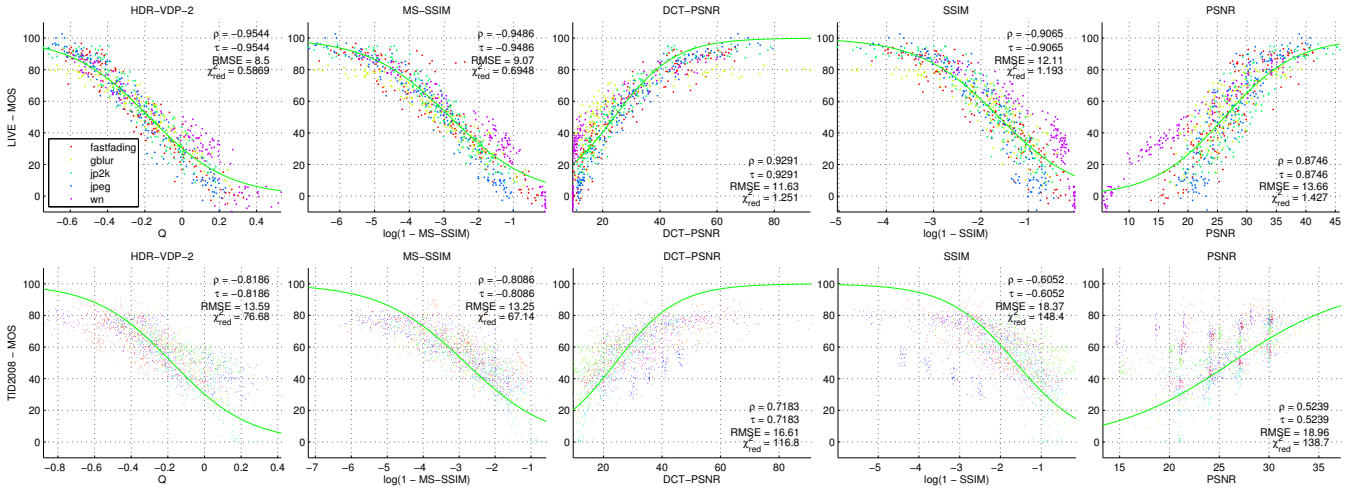


Figure 13: *Predictions of quality metrics for the LIVE (top) and TID2008 (bottom) databases. The prediction accuracy is reported as Spearman's ρ, Kendall's τ, root-mean-square-error (RMSE), and the reduced χ^2 statistics. The solid line is the best fit of the logistic function. Only the LIVE database was used for fitting. See supplementary for enlarged and more detailed plots.*

DE VALOIS, R., ALBRECHT, D., AND THORELL, L. 1982. Spatial frequency selectivity of cells in macaque visual cortex. *Vision Research 22*, 5, 545–559.

FERWERDA, J., PATTANAIK, S., SHIRLEY, P., AND GREENBERG, D. 1996. A model of visual adaptation for realistic image synthesis. In *Proc. of SIGGRAPH 96*, 249–258.

FOLEY, J. 1994. Human luminance pattern-vision mechanisms: masking experiments require a new model. *Journal of the Optical Society of America A 11*, 6, 1710–1719.

GEORGESON, M., AND GEORGESON, J. 1987. Facilitation and masking of briefly presented gratings: time-course and contrast dependence. *Vision Research 27*, 3, 369–379.

GEORGESON, M. A., AND SULLIVAN, G. D. 1975. Contrast constancy: deblurring in human vision by spatial frequency channels. *J. Physiol. 252*, 3 (Nov.), 627–656.

HE, S., AND MACLEOD, D. 1998. Contrast-modulation flicker: Dynamics and spatial resolution of the light adaptation process. *Vision Res 38*, 7, 985–1000.

HESS, R., SHARPE, L., AND NORDBY, K. 1990. *Night Vision: Basic, Clinical and Applied Aspects*. Cambridge University Press.

IJSPEERT, J., VAN DEN BERG, T., AND SPEKREIJSE, H. 1993. An improved mathematical description of the foveal visual point spread function with parameters for age, pupil size and pigmentation. *Vision research 33*, 1, 15–20.

ITU-R-BT.500-11, 2002. Methodology for the subjective assessment of the quality of television pictures.

LEE, B., DACEY, D., SMITH, V., AND POKORNY, J. 1999. Horizontal cells reveal cone type-specific adaptation in primate retina. *Proceedings of the National Academy of Sciences of the United States of America 96*, 25, 14611.

LOVELL, P., PÁRRAGA, C., TROSCIANKO, T., RIPAMONTI, C., AND TOLHURST, D. 2006. Evaluation of a multiscale color model for visual difference prediction. *ACM Transactions on Applied Perception (TAP) 3*, 3, 155–178.

LUBIN, J. 1995. *A visual discrimination model for imaging system design and evaluation*. World Scientific Publishing Company, 245.

LUKIN, A. 2009. Improved Visible Differences Predictor Using a Complex Cortex Transform. *International Conference on Computer Graphics and Vision (GraphiCon)*.

MACLEOD, D., WILLIAMS, D., AND MAKOUS, W. 1992. A visual nonlinearity fed by single cones. *Vision Res 32*, 2, 347–63.

MANTIUK, R., DALY, S., MYSZKOWSKI, K., AND SEIDEL, H. 2005. Predicting visible differences in high dynamic range images: model and its calibration. In *Proc. SPIE*, vol. 5666, 204–214.

MANTIUK, R., DALY, S., AND KEROFSKY, L. 2008. Display adaptive tone mapping. *ACM Transactions on Graphics (Proc. of SIGGRAPH) 27*, 3, 68.

MANTIUK, R., REMPEL, A. G., AND HEIDRICH, W. 2009. Display considerations for night and low-illumination viewing. In *Proc. of APGV '09*, 53–58.

MARIMONT, D., AND WANDELL, B. 1994. Matching color images: The effects of axial chromatic aberration. *Journal of the Optical Society of America A 11*, 12, 3113–3122.

MEESE, T., AND SUMMERS, R. 2007. Area summation in human vision at and above detection threshold. *Proceedings of the Royal Society B: Biological Sciences 274*, 2891–2900.

MYSZKOWSKI, K., ROKITA, P., AND TAWARA, T. 1999. Perceptually-informed accelerated rendering of high quality walkthrough sequences. *Rendering Techniques 99*, 5–18.

PATTANAIK, S. N., FERWERDA, J. A., FAIRCHILD, M. D., AND GREENBERG, D. P. 1998. A multiscale model of adaptation and spatial vision for realistic image display. In *Proc. of SIGGRAPH'98*, 287–298.

PATTANAIK, S., TUMBLIN, J., YEE, H., AND GREENBERG, D. 2000. Time-dependent visual adaptation for realistic image display. In *Proc. of SIGGRAPH'00*, 47–54.

PONOMARENKO, N., BATTISTI, F., EGIAZARIAN, K., ASTOLA, J., AND LUKIN, V. 2009. Metrics performance comparison for color image database. In *4th int. workshop on video processing and quality metrics for consumer electronics (QoMEX)*.

RAMANARAYANAN, G., FERWERDA, J., WALTER, B., AND BALA, K. 2007. Visual equivalence: towards a new standard for image fidelity. *ACM Trans. on Graphics (SIGGRAPH'07)*, 76.

RAMASUBRAMANIAN, M., PATTANAIK, S. N., AND GREENBERG, D. P. 1999. A perceptually based physical error metric for realistic image synthesis. In *Proc. of SIGGRAPH '99*, 73–82.

RICHTER, T. 2009. On the mDCT-PSNR image quality index. In *Quality of Multimedia Experience, 2009. QoMEx*, 53–58.

RITSCHEL, T., IHRKE, M., FRISVAD, J. R., COPPENS, J., MYSZKOWSKI, K., AND SEIDEL, H.-P. 2009. Temporal Glare: Real-Time Dynamic Simulation of the Scattering in the Human Eye. *Computer Graphics Forum 28*, 2, 183–192.

ROHALY, A., AHUMADA JR, A., AND WATSON, A. 1997. Object detection in natural backgrounds predicted by discrimination performance and models. *Vision Research 37*, 23, 3225–3235.

ROVAMO, J., KUKKONEN, H., AND MUSTONEN, J. 1998. Foveal optical modulation transfer function of the human eye at various pupil sizes. *Journal of the Optical Society of America A 15*, 9, 2504–2513.

SHEIKH, H., SABIR, M., AND BOVIK, A. 2006. A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms. *IEEE Transactions on Image Processing 15*, 11, 3440–3451.

SIMONCELLI, E., AND FREEMAN, W. 2002. The steerable pyramid: a flexible architecture for multi-scale derivative computation. In *Proceedings., International Conference on Image Processing*, IEEE Comput. Soc. Press, vol. 3, 444–447.

SMITH JR, R., AND SWIFT, D. 1985. Spatial-frequency masking and Birdsalls theorem. *Journal of the Optical Society of America A 2*, 9, 1593–1599.

STOCKMAN, A., AND SHARPE, L. 2000. The spectral sensitivities of the middle-and long-wavelength-sensitive cones derived from measurements in observers of known genotype. *Vision Res 40*, 13, 1711–1737.

STROMEYER, C. F., AND JULESZ, B. 1972. Spatial-Frequency Masking in Vision: Critical Bands and Spread of Masking. *Journal of the Optical Society of America 62*, 10 (Oct.), 1221.

VAN DEN BERG, T., IJSPEERT, J., AND DE WAARD, P. 1991. Dependence of intraocular straylight on pigmentation and light transmission through the ocular wall. *Vision Res 31*, 7-8, 1361–7.

VOS, J., AND VAN DEN BERG, T. 1999. Report on disability glare. *CIE Research Note 135*, 1.

WANG, Z., AND BOVIK, A. C. 2006. *Modern Image Quality Assessment*. Morgan & Claypool.

WANG, Z., SIMONCELLI, E., AND BOVIK, A. 2003. Multiscale structural similarity for image quality assessment. In *Asilomar Conference on Signals, Systems & Computers, 2003*, 1398–1402.

WANG, Z., BOVIK, A., SHEIKH, H., AND SIMONCELLI, E. 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing 13*, 4, 600–612.

WATSON, A., AND AHUMADA JR, A. 2005. A standard model for foveal detection of spatial contrast. *Journal of Vision 5*, 9, 717–740.

WATSON, A., AND PELLI, D. 1983. QUEST: A Bayesian adaptive psychometric method. *Perception & Psychophysics 33*, 2, 113–120.

WATSON, A., AND SOLOMON, J. 1997. Model of visual contrast gain control and pattern masking. *Journal of the Optical Society of America A 14*, 9, 2379–2391.

WATSON, A. 1987. The cortex transform: Rapid computation of simulated neural images. *Computer Vision, Graphics, and Image Processing 39*, 3, 311–327.

WHITAKER, D., STEEN, R., AND ELLIOTT, D. 1993. Light scatter in the normal young, elderly, and cataractous eye demonstrates little wavelength dependency. *Optometry and Vision Science 70*, 11, 963–968.