



ELSEVIER

Signal Processing 78 (1999) 231–252

**SIGNAL
PROCESSING**

www.elsevier.nl/locate/sigpro

Issues in vision modeling for perceptual video quality assessment

Stefan Winkler*

Laboratoire de Traitement des Signaux, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland

Received 25 February 1998; received in revised form 16 April 1999

Abstract

Lossy compression algorithms used in digital video systems produce artifacts whose visibility strongly depends on the actual image content. Simple error measures such as RMSE or PSNR, albeit popular, ignore this important fact and are only a mediocre predictor of perceived quality. Many applications require more reliable assessment methods. This paper discusses issues in vision modeling for perceptual video quality assessment (PVQA). Its purpose is not to describe a particular model or system, but rather to summarize and to provide pointers to up-to-date knowledge of important characteristics of the human visual system, to explain how these characteristics may be incorporated in vision models for PVQA, to give a brief overview of the state-of-the-art and current efforts in this field, and to outline directions for future research. © 1999 Elsevier Science B.V. All rights reserved.

Zusammenfassung

Verlustbehaftete Kompressionsalgorithmen, wie sie in digitalen Video-Systemen verwendet werden, erzeugen Artefakte, deren Sichtbarkeit stark vom Bildinhalt abhängt. Einfache Fehlermetriken wie der mittlere quadratische Fehler oder der Signal-Rausch-Abstand sind zwar weitverbreitet, doch sie ignorieren diese wichtige Tatsache und können folglich nur ein mittelmäßiger Indikator für die wahrgenommene Qualität sein. Viele Anwendungen benötigen jedoch zuverlässigere Beurteilungsmethoden. Dieser Artikel behandelt Aspekte der Modellierung des visuellen Systems für wahrnehmungsbasierte Videoqualitätsbeurteilung (PVQA). Das Ziel ist weniger, ein spezielles Modell oder System zu beschreiben, sondern vielmehr den gegenwärtigen Wissensstand über wichtige Charakteristiken des menschlichen visuellen Systems zusammenzufassen, die Integration dieser Charakteristiken in Modelle zur Videoqualitätsbeurteilung zu erläutern, einen Überblick über den Stand der Technik und über aktuelle Arbeiten zu geben, sowie mögliche Richtungen für zukünftige Forschungsarbeit aufzuzeigen. © 1999 Elsevier Science B.V. All rights reserved.

Résumé

Les algorithmes de compression avec perte utilisés dans les systèmes vidéo numériques produisent des artefacts dont la visibilité dépend fortement du contenu des images traitées. Les mesures d'erreurs les plus simples, telles que l'erreur quadratique moyenne ou le rapport signal sur bruit, ignorent cette caractéristique et ne sont donc qu'un prédicteur médiocre de la qualité perçue. Beaucoup d'applications nécessitent des techniques de mesure plus fiables. Cet article présente divers aspects de la modélisation du système visuel dans le cadre de l'évaluation de qualité des séquences vidéo (PVQA). L'objectif n'est pas de décrire un modèle particulier mais plutôt d'exhiber et résumer l'état de nos connaissances

* Tel.: + 41-21-693-4622; fax: + 41-21-693-7600.

E-mail address: stefan.winkler@epfl.ch (S. Winkler)

des caractéristiques importantes du système visuel humain, d'expliquer comment celles-ci peuvent être incorporées dans les modèles de vision appliqués à l'évaluation de qualité des séquences video, de donner un aperçu de l'état de la recherche dans ces domaines et enfin de proposer de possibles directions d'investigation. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Human visual system; Vision models; Video quality assessment

1. Introduction

The advent of digital video systems has exposed the limitations of the techniques traditionally used for video quality measurement. For conventional analog video systems there are well-established performance standards. They rely on particular test signals and measurement procedures to determine parameters such as differential gain, differential phase or waveform distortion, which can be related to perceived quality with relatively high accuracy [143]. While these parameters are still useful today, their connection with perceived quality has become much more tenuous; because of compression, digital video systems exhibit artifacts fundamentally different from analog video systems – examples include blockiness, blurring, ringing, color bleeding, and motion compensation mismatches [8,140]. The amount and visibility of these distortions strongly depends on the actual image content. Therefore, traditional signal quality measurements are inadequate for the evaluation of these compression artifacts.

Given these limitations, the designers of compression algorithms have had to resort to subjective viewing tests in order to obtain reliable ratings for the quality of compressed images or video [50]. While these tests – if executed properly – are the closest we can get to the “truth” about perceived quality, they are complex, time-consuming and consequently expensive. Hence, they are often highly impractical or not feasible at all.

Looking for faster alternatives, researchers have turned to simple error measures such as root mean squared error (RMSE) or peak signal-to-noise-ratio (PSNR), suggesting that they would be equally valid. However, these simple error measures operate solely on a pixel-by-pixel basis and neglect the important influence of image con-

tent and viewing conditions on the actual visibility of artifacts. Therefore, they cannot correlate well with perceived quality, and many experiments confirm this low correlation – see for example [67,81].

These problems necessitate methods of objective video quality assessment. As a matter of fact, there is a broad range of applications for objective video quality assessment systems, including:

- evaluation, test, and comparison of video codecs;
- online quality monitoring and control;
- end-to-end testing of video transmission/communication systems;
- perceptual video compression;
- perceptual video restoration.

Coupled with appropriate video segmentation, the quality of specific features (e.g. contours or textures) or specific compression artifacts (e.g. blockiness) can be evaluated separately and eventually used to tune certain parts of the encoder [111]. In a similar fashion, the quality of motion rendition can be assessed [18].

In order to be able to replace subjective rating experiments, the ideal objective quality assessment system should rate video impairments just like a human being. Considering the variety of compression algorithms available and the rapid change of technology in this field, a quality metric that is independent of the particular algorithm is preferable in order to avoid early obsolescence. Metrics based on models of the human visual system are one way to achieve this technology independence, because they are the most general and potentially the most accurate ones. However, the human visual system is extremely complex, and many of its properties are not well understood even today. Evidently, these uncertainties about the actual processing of visual information in the human brain complicate the design of vision models and explain

many of the differences between existing PVQA systems.

While systems for the quality assessment of still images are already too numerous to mention here (see e.g. [2] for a review), their extension to moving pictures has not received much attention until recently. Lukas and Budrikis [68] were the first to propose a comprehensive metric based on a spatio-temporal model of the human visual system in 1982. Other models and metrics followed now and then [37,75,117], but only in the past few years has there been an increasing interest in PVQA, as the rising number of publications shows [14,46,62,67,108,110,112,118,122,127,132,138].

This paper is structured as follows. First, it discusses what quality constitutes to the human observer and how it can be measured. Then, it describes the human visual system, from the optics of the eye up to the neurons in the primary visual cortex and higher-level cognition. In parallel, it outlines in every section how each component or phenomenon may be incorporated in a vision model for PVQA, and which restrictions apply. The paper concludes with a section about the validation and evaluation of PVQA systems.

2. Quality factors

In order to be able to design a reliable PVQA system, it is necessary to understand what “quality” means to the viewer. Viewers’ enjoyment when watching a video depends on many factors. One of the most important is of course program content and material. Provided the content itself is at least “watchable”, video and sound quality play a prominent role. Research has shown that video quality depends on viewing distance, display size, resolution, brightness, contrast, sharpness, colorfulness, naturalness and other factors [3,56,74,93]. To make things worse, there is a difference between fidelity (the accurate reproduction on the display) and perceived quality – for instance, subjects prefer slightly more colorful images despite realizing that they look somewhat unnatural [26,139]. The accompanying sound has also been shown to influence perceived video quality: subjective quality ratings are generally higher when the test scenes are

accompanied by a good quality sound program, which apparently lowers the viewers’ ability to detect video impairments [90].

It is helpful for the following sections to relate the definitions of some of these factors to vision modeling and the human visual system. For instance, in the video community it is very popular to specify viewing distance in terms of display size, i.e. in multiples of screen height. There are two reasons for this: first, it was assumed for quite some time that the ratio of preferred viewing distance to screen height is constant [69]. More recent experiments with larger displays have shown, however, that this is not the case. While the preferred viewing distance is indeed around 6 or 7 screen heights for smaller displays, it approaches 3 to 4 screen heights with increasing display size [9,69]. Incidentally, typical home viewing distances are far from ideal in this respect [7]. The second reason was the implicit assumption about a certain display resolution (a certain number of scan lines), which is usually fixed for a given television standard. In the context of vision modeling, the size and resolution of the image projected onto the retina are more adequate specifications. The size is measured in degrees of visual angle α , and the resolution or maximum spatial frequency f_{\max} is measured in cycles per degree of visual angle (cpd). For a given screen height H , viewing distance D and number of scan lines L , these two units are computed as follows:

$$\alpha = 2 \arctan(H/2D),$$

$$f_{\max} = L/2\alpha \text{ [cpd]}.$$

Table 1 gives examples for the size and resolution of the image that some television systems and

Table 1
Approximate angular sizes and resolutions of TV systems at viewing distances of three and six times screen height

System	D	Vertical	Horizontal	f_{\max}
16:9 HDTV	$6H$	9.5°	17°	60 cpd
16:9 HDTV	$3H$	19°	34°	30 cpd
4:3 PAL	$6H$	9.5°	13°	30 cpd
4:3 NTSC	$6H$	9.5°	13°	25 cpd
4:3 PAL	$3H$	19°	25°	15 cpd

viewing setups produce on the retina. It is instructive to compare these figures to the corresponding “specifications” of the human visual system mentioned in later sections.

3. Subjective quality assessment

Subjective quality ratings form the benchmark for objective metrics. However, different applications require different testing procedures. Formal subjective testing is defined in ITU-R (formerly CCIR) Recommendation 500 [50], which suggests standard viewing conditions, criteria for observer and test scene selection, assessment procedures, and analysis methods. Outline three of the more commonly used procedures here.

- Double stimulus continuous quality scale (DSCQS). The presentation sequence for a DSCQS trial is illustrated in Fig. 1. Viewers are shown multiple sequence pairs consisting of a “reference” and a “test” sequence, which are rather short (typically 10 seconds). The reference and test sequence are presented twice in alternating fashion, with the order of the two chosen randomly for each trial. Subjects are not informed which is the reference and which is the test sequence. They rate each of the two separately on a continuous quality scale ranging from “bad” to “excellent” as shown in Fig. 2. Analysis is based on the difference in rating for each pair, which is often calculated from an equivalent numerical scale from 0 to 100.
- Double stimulus impairment scale (DSIS). The presentation sequence for a DSIS trial is illustrated in Fig. 3. As opposed to the DSCQS method, the reference is always shown before the test sequence, and neither is repeated. Subjects rate the amount of impairment in the test sequence on a discrete five-level scale ranging from “very annoying” to “imperceptible” as shown in Fig. 4.
- Single stimulus continuous quality evaluation (SSCQE) [77]. Instead of seeing separate short sequence pairs, viewers watch a program of typically 20–30 minutes duration which has been processed by the system under test; the reference is not shown. Using a slider, the subjects con-

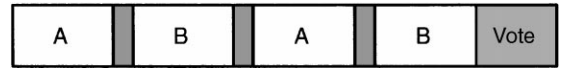


Fig. 1. Presentation sequence for the DSCQS method.

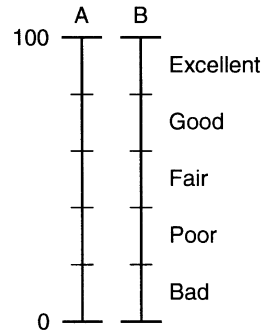


Fig. 2. Rating scale for the DSCQS method.



Fig. 3. Presentation sequence for the DSIS method.

- Imperceptible
- Perceptible but not annoying
- Slightly annoying
- Annoying
- Very annoying

Fig. 4. Rating scale for the DSIS method.

tinuously rate the instantaneously perceived quality on the DSCQS scale from “bad” to “excellent” (Fig. 2).

The above-mentioned methods generally have different applications. DSCQS is the preferred method when the quality of test and reference

sequence are similar, because it is quite sensitive to small differences in quality. The DSIS method is better suited for evaluating clearly visible impairments such as artifacts caused by transmission errors, for example. Both DSCQS and DSIS method share a common drawback, however: changes in scene complexity or statistical multiplexing in the transmission system can produce substantial quality variations that are not evenly distributed over time; severe degradations may appear only once every few minutes. The standard DSCQS and DSIS methods with their one-time rating are not suited to the evaluation of such long sequences because of the recency phenomenon, a bias in the ratings toward the final 10–20 seconds due to limitations of human working memory [6]. Furthermore, it has been argued that the presentation of a reference and the repetition of the sequences in the DSCQS method puts the subjects in a situation too removed from the home-viewing environment by allowing them to become familiar with the material under investigation [63]. SSCQE has been designed with these problems in mind, as it relates well to the time-varying quality of today's compressed digital video systems [78]. On the other hand, program content tends to have a significant influence on SSCQE scores. Also, SSCQE scores of different tests are harder to compare because of the lack of a reference.

4. Optics of the eye

The optics of the eye constitute the first processing stage that visual information passes on its way through the human visual system. Attempts to make general statements about the eye's optical characteristics are complicated by the fact that there are considerable variations of its properties between individuals. Furthermore, its components undergo continuous changes throughout life. In general, however, the parameters of an individual healthy eye are correlated in such a way that the eye can produce a sharp image of a distant object on the retina [16].

To determine the quality of the optics of the eye, the reflection of a visual stimulus projected onto the retina can be measured [13]. The retinal image

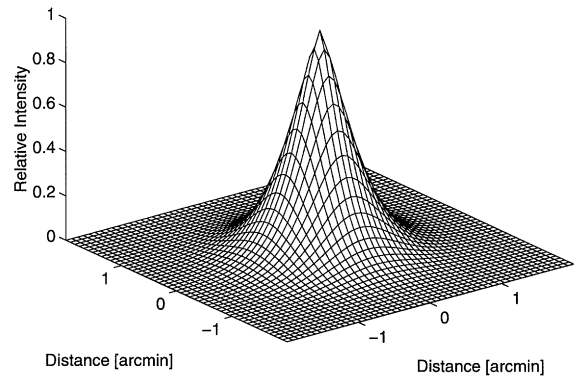


Fig. 5. Point spread function of the human eye as a function of visual angle [128].

turns out to be a distorted version of the input, the most noticeable distortion being blurring. To quantify the amount of blurring, a point or a thin line is used as the input image, and the resulting retinal image is called the *point spread function* or *line spread function* of the eye; its Fourier transform is the modulation transfer function. The amount of blurring depends on the pupil size: for small pupil diameters up to 3–4 mm, the optical blurring is close to the diffraction limit; as the pupil diameter increases (for lower ambient light intensities), the width of the point spread function increases as well, because the distortions due to cornea and lens imperfections become large compared to diffraction effects [13,94]. The optical quality also deteriorates with increasing distance from the optical axis and the fovea [61].

Westheimer [128] proposed a simple formula to approximate the foveal point spread function of the human eye when in good focus with a pupil diameter of 3 mm:

$$\text{PSF}(\alpha) = 0.952e^{-2.59|\alpha|^{1.36}} + 0.048e^{-2.43|\alpha|^{1.74}},$$

α being in minutes of arc. This function is illustrated in Fig. 5. For more elaborate expressions with parameters for pupil size [94], age and pigmentation [49], the reader is referred to the literature.

The point spread function also changes with wavelength. By accommodation, the eye can place any wavelength into good focus, but it is impossible to focus all wavelengths simultaneously. This effect is called *chromatic aberration*. It can be quantified

by determining the modulation transfer function of the human eye for different wavelengths. This is shown in Fig. 6 for a human eye model with a pupil diameter of 3 mm and in focus at 580 nm [73]. It is evident that the retinal image contains only poor spatial detail at wavelengths far from the in-focus wavelength (note the sharp cutoff going down to a few cycles per degree at short wavelengths). This tendency toward monochromacy becomes even more pronounced with increasing pupil diameter.

As far as modeling is concerned, some PVQA systems incorporate the point spread function to blur the input prior to all other processing, but none of the models I know take the effects of chromatic aberration into account explicitly. It can be argued that the blurring can be considered at a later stage by appropriate modeling of contrast sensitivity (see also Section 7), but this approach ignores many fine details of the shape of the modulation transfer function and its variation with wavelength (cf. Fig. 6).

5. Photoreceptor mosaic

Through the optics of the eye, the visual input is projected onto the retina, the neural tissue at the back of the eye composed of the photoreceptor mosaic. The photoreceptors sample the image and convert the information into signals that can be interpreted by the brain. There are two different types of photoreceptors, *rods* and *cones*. Rods are responsible for vision at low light levels, i.e. under *scotopic* conditions. In general, they can be neglected for the applications considered in this paper, because TV displays operate at much higher light levels.

Cones are responsible for vision at these higher light levels, i.e. under *photopic* conditions. They are concentrated in the fovea, the region of highest visual acuity, which covers approximately two degrees of visual angle on the retina. As a matter of fact, there are three types of cones, L-cones, M-cones, and S-cones, sensitive to long, medium and short wavelengths, respectively. They form the basis of color perception (see Section 6). Estimates of their spectral sensitivities are shown in Fig. 7 [104]. Note that these measurements were made

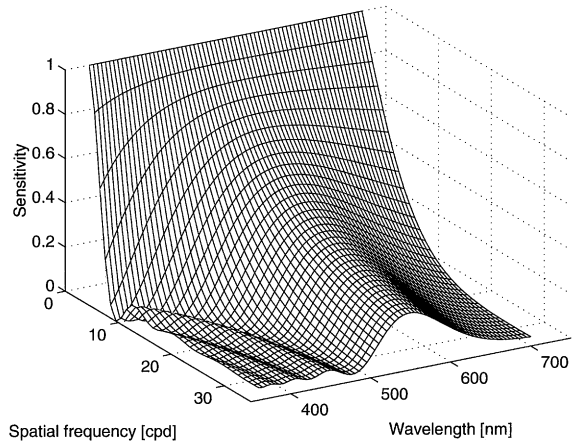


Fig. 6. Variation of the modulation transfer function of a human eye model with wavelength [73].

with a light source at the cornea, and are thus influenced by the optical system of the eye as described above.

The density of photoreceptors varies greatly across the retina [1]. L- and M-cones dominate overall; in the central fovea, they form a tightly packed mosaic reaching a density of up to 300,000 mm^2 [19]. At a size of approximately 0.5 minutes of visual angle, the maximum frequency of around 60 cpd attained here is high enough to capture all of the spatial variation after the blurring by the eye's optics. S-cones are much more sparse and account for less than 10% of the total number of cones. They are spaced approximately 10 minutes apart on average, resulting in a maximum frequency of only 3 cpd [20]. This is probably rooted in the strong defocus of short-wavelength light by the eye's optics (see Fig. 6).

In fact, many PVQA systems neglect eccentricity and off-axis effects and concentrate their modeling efforts on the fovea. This is often justified with the fact that the eyes are directed in such a way that the current region of attention is brought into focus there. It also significantly simplifies modeling, because the optical and retinal properties are relatively uniform across the fovea. However, it must not be forgotten that its diameter of two degrees is rather small compared to the size of a TV display projection on the retina (cf. Table 1).

6. Color perception

As is evident from Fig. 7, there is a significant overlap between L- and M-cone sensitivities. In order to improve the efficiency of the visual encoding, the L-, M-, and S-cone absorption rates are decorrelated very early in the visual system by forming new signals.

Hering [44] was the first to point out that some pairs of hues can coexist in a single color sensation (e.g. a reddish yellow is perceived as orange), while others cannot (we never perceive a reddish green, for instance). This led him to the conclusion that the sensations of red and green as well as blue and yellow are encoded in separate visual pathways, which is commonly referred to as the theory of *opponent colors*. Both psychological and physiological experiments in the 1950s yielded further evidence to support this theory [48,52]. The principal components of opponent-colors space are black-white (B-W), red-green (R-G) and blue-yellow (B-Y). The precise color directions of these components are still subject to debate, however. As an example, the spectral sensitivities of the opponent colors space derived by Poirson and Wandell [85,86] are shown in Fig. 8. As can be seen, the B-W channel, which encodes luminance information, is determined mainly by medium to long

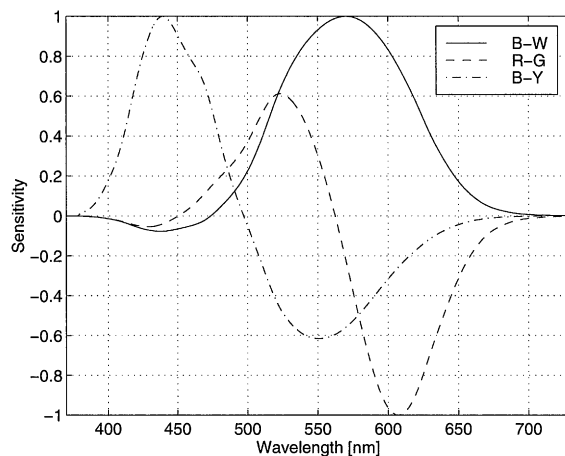


Fig. 8. Normalized spectral sensitivities of the three components black-white (solid), red-green (dashed), and blue-yellow (dot-dashed) of the opponent-colors space derived by Poirson and Wandell [85,86].

wavelengths. The R-G channel discriminates between medium and long wavelengths, while the B-Y channel discriminates between short and medium wavelengths. Because most of the psychophysical experiments for chromatic contrast sensitivity and chromatic masking (see Sections 7 and 9) are based on opponent-colors stimuli, vision models working in opponent-colors space have the advantage of their channels being adapted to these stimuli, which facilitates model design and analysis [110,120].

Alternatively, models employing CIE $L^*u^*v^*$ [67] or a modified CIE $L^*a^*b^*$ [141,142] color space instead of or in combination with an opponent-colors space have been proposed for PVQA systems. The roots of CIE $L^*u^*v^*$ can be traced back to color television studies, while CIE $L^*a^*b^*$ comes from the textile industry [47]. Both CIE $L^*u^*v^*$ and CIE $L^*a^*b^*$ color spaces (see Appendix A for transformation formulas) were defined with a perceptually uniform measure for color differences in mind: the Euclidean distance between color coordinates in these spaces provides an approximation to the perceived difference [134]. This can be advantageous for PVQA systems because they try to determine the amount of this perceived difference between reference and test sequences.

It is interesting to note that in a comparison between a luminance-only PVQA system and its

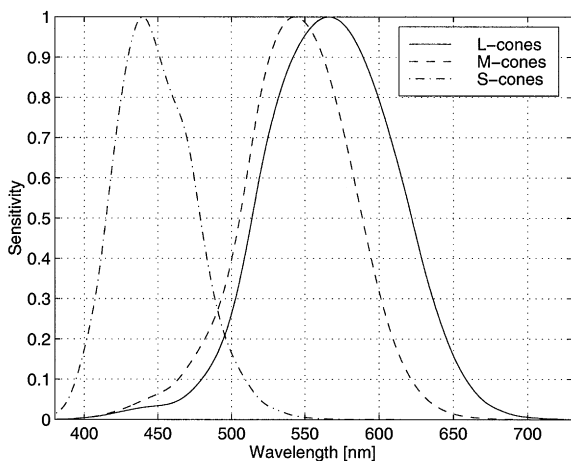


Fig. 7. Normalized spectral sensitivities of the three cone types: L-cones (solid), M-cones (dashed), and S-cones (dot-dashed) [104].

full-color extension, the results differed only slightly [111]. This is to be expected since many encoders distribute the distortions more or less equally between chromatic and achromatic channels. Future tests will have to show how the significant increase in complexity and computational load for color PVQA systems can be balanced against quality rating accuracy.

6.1. Component video

The color spaces used in many standards for coding visual information, including PAL, NTSC, JPEG, MPEG and others, already take into account certain properties of the human visual system; the above-mentioned theory of opponent colors and the fact that acuity for color information is lower than for luminance prompted the use of color difference components instead of color primaries for coding. Furthermore, the human visual system has a nonlinear, roughly logarithmic response to intensity. Therefore, a compressive nonlinearity is applied before coding [87].

It so happens that conventional television cathode ray tube (CRT) displays also have a nonlinear relationship between signal voltage or frame buffer values x and displayed intensity I [10]. This relationship can be slightly different for each of the three color primaries, but it can be approximated quite well by a function such as

$$I(x) = (\alpha x + \beta)^\gamma. \quad (1)$$

The exponent γ usually varies between 2.2 and 2.5; α and β can be adjusted with the picture/contrast and black level/brightness controls. Applying the inverse of this function to intensity values is referred to as *gamma correction*. Coincidentally, the lightness sensitivity of human vision is very nearly the inverse of Eq. (1) [87]. Therefore, coding visual information in the gamma-corrected domain is not only more meaningful perceptually, but also automatically compensates for CRT nonlinearities.

ITU-R Recommendation 601 [51] is the international standard for studio-quality component digital video. It defines a $Y'C'_B C'_R$ color space, where Y' encodes luminance, C'_B the difference between blue primary and luminance, and C'_R the difference between red primary and luminance (the

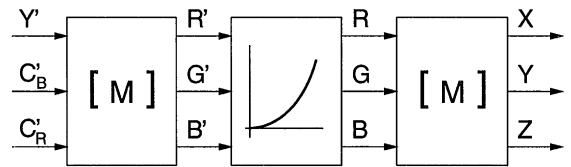


Fig. 9. Conversion from video $Y'C'_B C'_R$ components to CIE XYZ tristimulus values [88].

prime is used here to emphasize the nonlinear nature of these quantities). Because the $Y'C'_B C'_R$ space assumes a particular display device, or to be more exact, a particular spectral power distribution of the light emitted from the display, CIE XYZ tristimulus values serve as a reference for conversions from $Y'C'_B C'_R$ to the color spaces discussed above. Conversion from $Y'C'_B C'_R$ to CIE XYZ requires two linear transformations and gamma-correction as illustrated in Fig. 9; the corresponding formulas are given in Appendix B.

7. Contrast sensitivity

Contrast is a measure of the relative variation of luminance. Unfortunately, a common definition of contrast suitable for all stimuli does not exist. In the case of a periodic pattern of symmetrical deviations ranging from L_{\min} to L_{\max} , Michelson contrast [76] is generally used:

$$C_M = \frac{L_{\max} - L_{\min}}{L_{\max} + L_{\min}}.$$

When the pattern consists of a single increment or decrement ΔL to an otherwise uniform background luminance L , Weber contrast is often used [80]:

$$C_w = \frac{\Delta L}{L}.$$

These two definitions are by no means equivalent and do not even share a common range of values: Michelson contrast can range from 0 to 1, whereas Weber contrast can range from -1 to ∞ . To make things worse, neither of the two is appropriate for measuring contrast in complex images, because a few very bright or very dark points would

determine the contrast of the whole image. Furthermore, human contrast sensitivity varies with the adaptation level associated with the local average luminance. In order to address these issues, Peli [82] proposed a local band-limited contrast measure

$$C_i(x,y) = \frac{BP_i(x,y)}{LP_i(x,y)},$$

where $BP_i(x,y)$ is the bandpass image of band i , and $LP_i(x,y)$ contains the energy below band i . Modifications of this local band-limited contrast definition have been used successfully in vision models for PVQA [21,66] and are in good agreement with psychophysical contrast-masking experiments with Gabor patches [83]. Nevertheless, more experiments are necessary before one can conclude that the definite measure for contrast in complex images has been found [133].

Sensitivity to contrast depends on the color as well as the spatial and temporal frequency of the stimuli. Contrast sensitivity functions (CSFs) are generally used to quantify these dependencies. Contrast sensitivity is defined as the inverse of the contrast threshold, i.e. the minimum contrast necessary for an observer to detect the target.

Spatio-temporal CSF approximations are shown in Fig. 10. Achromatic contrast sensitivity is generally higher than chromatic, especially for high spatio-temporal frequencies. The full range of colors is perceived only at low frequencies. As spatio-temporal frequencies increase, blue-yellow sensitivity declines first. At even higher frequencies, red-green sensitivity diminishes as well, and perception becomes achromatic. On the other hand, albeit to a lesser extent, achromatic sensitivity decreases at low spatio-temporal frequencies, whereas chromatic sensitivity does not. However, this apparent attenuation of sensitivity towards low frequencies may be attributed to implicit masking, i.e. masking by the spectrum of the window within which the test gratings are presented [137].

There has been some debate about the space-time separability of the spatio-temporal CSF. This property is of interest in vision modeling because a CSF that could be expressed as a product of spatial and temporal components would simplify

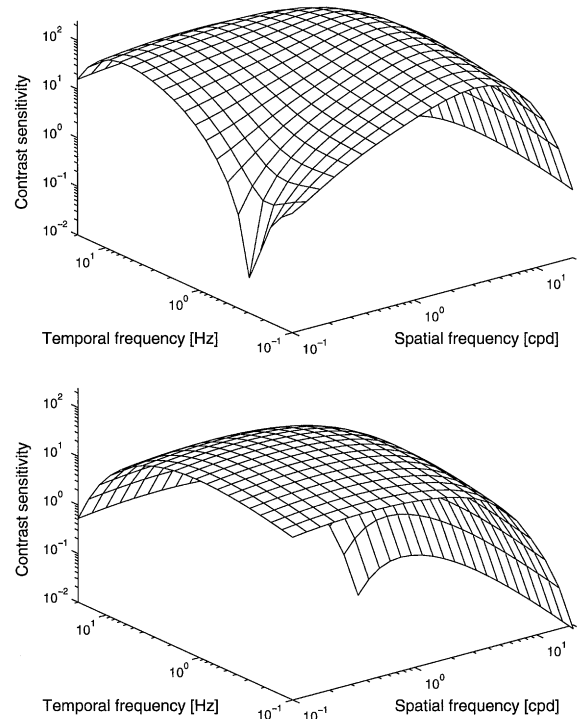


Fig. 10. Spatio-temporal contrast sensitivity functions of the B-W channel (top) and the R-G channel (bottom) according to [12,53–55]. The CSF of the B-Y channel (not shown) is very similar in shape to the CSF of the R-G channel.

modeling. Early studies concluded that the spatio-temporal CSF was not space-time separable at lower frequencies [59,91]. Kelly [53] measured contrast sensitivity under stabilized conditions (i.e. the stimuli were stabilized on the retina by compensating for the observers' eye movements). He fit an analytic function to his measurements [54], which is technically the CSF for traveling waves. Through variable substitution, it can be rewritten in terms of spatial frequency f_s and temporal frequency f_t of the test stimuli to yield a very close approximation of the spatio-temporal CSF for counterphase flicker:

$$\text{CSF}(f_s, f_t) = 4\pi^2 f_s f_t e^{-4\pi(f_t + 2f_s)/45.9} \times (6.1 + 7.3|\log(f_t/3f_s)|^3). \quad (2)$$

Burbeck and Kelly [12,55] found that this CSF and also its chromatic counterparts can be approximated by linear combinations of two space-time

separable components termed excitatory and inhibitory CSFs. The resulting approximations of the CSFs of the B-W and R-G channels are shown in Fig. 10. The CSF of the B-Y channel is very similar in shape to the CSF of the R-G channel; the B-Y sensitivity is somewhat lower overall, and its high-frequency decline sets in earlier.

Yang and Makous [136] measured the spatio-temporal CSF for both in-phase and conventional counterphase modulation. Their results suggest that the underlying filters are indeed spatio-temporally separable and have the shape of lowpass exponentials. The spatio-temporal interactions observed for counterphase modulation may be explained as a product of masking by the zero-frequency component of the gratings.

Recently, Daly [22] addressed the important issue of unconstrained eye movements for CSF models in PVQA systems. In particular, he showed how to include natural drift, smooth pursuit and saccadic eye movements in Kelly's formulation of the stabilized spatio-temporal CSF given by Eq. (2) using a model for eye velocity. The effect on the shape of the CSF is substantial and can best be described as a stretch along the temporal frequency axis. Westen et al. [127] incorporated eye movements into their model by a similar motion compensation of the CSF. They proposed a method for the estimation of smooth-pursuit eye movements under the worst-case assumption that the observer is capable of tracking all the objects in the scene.

Existing PVQA systems are based on a variety of different contrast sensitivity measurements, mostly from the early studies. Basically, there are two possibilities for how to incorporate contrast sensitivity into a vision model: The first is to set the gain of each filter in the bank of a multi-channel implementation (see next section) such that the ensemble approximates the empirical CSF. The second is to pre-filter the B-W, R-G and B-Y channels with the respective contrast sensitivity functions and to calibrate the following stages of the model in such a way that no further variations in contrast sensitivity are introduced. Both approaches have been used in the PVQA systems proposed so far. While the first method is more efficient in the implementation, the second facilitates a more accurate approximation of the shape of the CSF.

8. Multi-resolution architecture

Early vision models [68,72,95] were based on single-resolution theory and provided a first insight into simple visual phenomena. However, they are unable to cope with more complex patterns and were soon challenged by empirical data from masking and pattern adaptation experiments. These data can be explained quite successfully by a multi-resolution theory of vision, which employs a whole set of different filters instead of just one.

8.1. Spatial mechanisms

A large number of neurons in the primary visual cortex, the so-called simple cells, have receptive fields composed of several parallel elongated excitatory and inhibitory regions as illustrated in Fig. 11 [23,116]. Hence, they can be characterized by a particular radial spatial frequency, defined by the distance between adjacent maxima or minima of the response function, and by an orientation, corresponding to the angle perpendicular to the "bars". Serving as an oriented bandpass filter, the neuron will respond to a certain range of spatial frequencies and orientations about its center

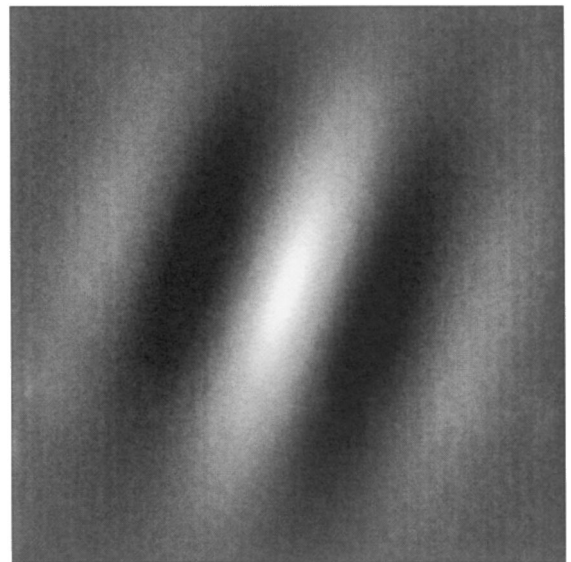


Fig. 11. Idealized receptive field of a neuron in the primary visual cortex.

values. There is still a lot of discussion about the exact tuning shape and bandwidth, and different experiments have led to different results. For the achromatic visual pathways, most of the studies give estimates of approximately 1 to 2 octaves for the spatial frequency bandwidth and 20 to 60 degrees for the orientation bandwidth, varying with spatial frequency [27,28,84]. These results are confirmed by psychophysical evidence from studies of discrimination and interaction phenomena [80]. Interestingly, these cell properties can also be related with and even derived from the statistics of natural images [32,113].

Fewer empirical data is available for the chromatic pathways. They probably have similar spatial frequency bandwidths [64,65,125], whereas their orientation bandwidths have recently been found to be significantly larger, ranging from 60 to 130 degrees [114].

Given these bandwidths, and considering the decrease in contrast sensitivity at high spatial frequencies (see previous section), the spatial frequency plane for the achromatic channel can be covered by 4–6 spatial frequency-selective and 4–8 orientation-selective mechanisms. Further reducing orientation selectivity can affect modeling accuracy, as was reported in a comparison of two models with 3 and 6 orientation-selective mechanisms, respectively [109].

Taking into account the larger orientation bandwidths of the chromatic channels, 2–3 orientation-selective mechanisms may suffice there. Chromatic sensitivity remains high down to very low spatial frequencies, which necessitates a low-pass mechanism and possibly additional spatial frequency-selective mechanisms at this end. For reasons of implementation simplicity, it may be advantageous to use the same decomposition for chromatic and achromatic channels nonetheless. An example of a partitioning of the spatial frequency plane used in a PVQA system by the author [131,132] is shown in Fig. 12.

8.2. Temporal mechanisms

Temporal mechanisms have been studied as well, but there is less agreement about their characteristics than for spatial mechanisms. While some

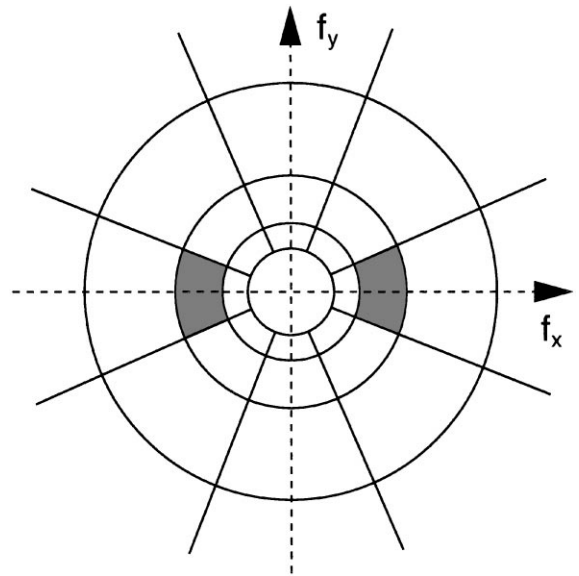


Fig. 12. Idealized illustration of a possible partitioning of the spatial frequency plane as used in [131,132]. The actual transitions between the bands are gradual. Three spatial frequency levels with four orientations plus one (isotropic) low-pass filter are shown. The shaded region indicates the spectral support of a single channel.

studies concluded that there are a large number of narrowly tuned mechanisms [60], it is now believed that there is just one lowpass and one bandpass mechanism [36,45,115], which are generally referred to as sustained and transient channel, respectively. An actual third mechanism was proposed [45,71], but has been called in question by later studies [36,40]. Physiological experiments confirm these findings to the extent that lowpass and bandpass mechanisms have been discovered [34], but neurons with bandpass properties exhibit a wide range of peak frequencies. Recent results also indicate that the peak frequency and bandwidth of the mechanisms change considerably with stimulus energy [35].

In a recent study, Fredericksen and Hess [35,36] model temporal mechanisms with derivatives of the impulse response function

$$h(t) = e^{-(\ln(t/\tau)/\sigma)^2}.$$

They achieve a very good fit to their data using only this function and its second derivative,

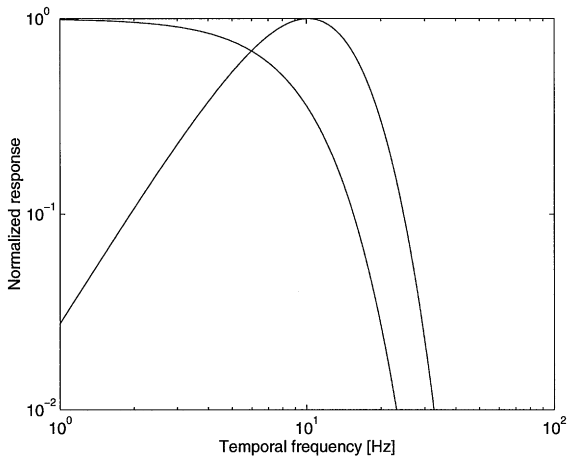


Fig. 13. Frequency responses of sustained (lowpass) and transient (bandpass) mechanisms of vision based on a model by Fredericksen and Hess [35,36].

corresponding to one sustained and one transient mechanism, respectively. For a typical choice of parameters $\tau = 0.16$ and $\sigma = 0.2$, the resulting frequency responses of their filters are shown in Fig. 13.

8.3. Filter design

The above findings can be incorporated into a PVQA system via a filter bank. The fundamental requirements for its design include joint localization in space, spatial frequency and orientation in order to model the frequency- and orientation-selectivity of channels in the human visual system. For implementation efficiency, a pyramid structure with self-similar filters and dyadic subsampling is favorable. Invertibility is an advantage in applications where perfect reconstruction from the channels is required; a filter set summing to 1 is also desirable because it allows the CSF to be modeled independently of the decomposition filters.

As noted earlier, certain receptive fields in the human visual system are tuned in spatial frequency and orientation; in fact, their profile (cf. Fig. 11) resembles two-dimensional Gabor functions [23,116]. Consequently, it was posited that cortical filters act to minimize simultaneously the joint product of standard deviation of spatial and

spatial-frequency sensitivities in accordance with the uncertainty principle from Fourier analysis [24]. Therefore, the Gabor transform may be considered an obvious implementation choice. However, this argument is based on a particular definition of uncertainty involving second-order moments, which may not be appropriate for the visual system [105]. Furthermore, only complex-valued Gabor functions have this property [57,135]. From a practical point of view, the Gabor transform is also difficult to reconstruct, hence other approaches have been investigated and have gained popularity.

Pyramid structures have been proposed for many image processing applications. They seek to reduce the number of pixels by repeated lowpass filtering and subsampling, which reduces the amount of computation. The cortex transform introduced by Watson [116] and later modified and used for quality assessment by Daly [21] is an example. It is appealing because of its flexibility: radial frequency selectivity and orientation selectivity are modeled separately, frequency and orientation bandwidth can be adjusted within a broad range, and the transform is easily invertible. Simoncelli et al. [99,100] proposed the steerable pyramid, which is attractive because of its “shiftability” property: it is translation- and rotation-invariant, self-inverting, essentially aliasing-free, and can be designed for any number of orientation bands.

The disadvantage of all these decompositions is that they are overcomplete. This is generally less of a concern for PVQA, but it is naturally undesirable for perceptual coding applications. Discrete wavelet transforms have proven highly efficient for coding applications because of their orthogonal basis functions. In contrast to the decompositions mentioned above, they are critically sampled, i.e. the number of transform coefficients is equal to the number of samples in the input signal. However, the amount of aliasing they introduce in the subbands as well as their behavior for translated or dilated input signals make them less useful for vision modeling [100]. The quadrature mirror filter (QMF) transform on a hexagonal grid [98] was used for perceptual distortion measurement by Teo and Heeger [109]. However, the orientation

bandwidth of these filters turned out to be too broad (nearly 60 degrees), which affected the fit of the model to psychophysical data.

The design of the temporal filter bank is governed by different criteria. In certain applications of PVQA systems such as monitoring and control, a low delay is important. This fact together with limitations of memory and computing power favor time-domain implementations of the temporal filters over frequency-domain decompositions. A trade-off has to be found between an acceptable delay and the accuracy with which the temporal mechanisms ought to be approximated. Recursive infinite impulse response (IIR) filters fare better in this respect than (nonrecursive) finite impulse response (FIR) filters; IIR filters can achieve a close approximation with delays of only a few frames [62,132], while FIR filters may introduce delays of a few dozen frames. However, the latter are easier to design and generally have better phase characteristics.

8.4. Spatio-temporal considerations

How can the interactions between spatial and temporal channels indicated by contrast sensitivity measurements be incorporated into the channels of a multi-resolution architecture? There are two hypotheses on this matter: The sensitivity-scaling hypothesis states that the temporal filters have a peak sensitivity that is independent of spatial frequency, only the filter gain changes; therefore, a filter bank that is separable in space and time can be used [129]. The covariation hypothesis states that the ensemble of filters exhibits a spatio-temporal covariation; in this case the filter bank cannot be separable.

The influence of these hypotheses on the performance of vision models has not yet been investigated. It is evident that the sensitivity-scaling paradigm permits an easier implementation. The sequence can be filtered first in the temporal domain, and afterwards the different temporal channels undergo separate spatial decompositions. PVQA system designers often choose this approach for reasons of simplicity [111].

The covariation paradigm offers a potentially more accurate modeling of vision mechanisms, but

requires a more elaborate spatio-temporal decomposition, as indicated in [117]. For this approach, the above-mentioned filter structures could be extended to the time dimension.

9. Masking

Masking is a very important phenomenon in vision in general and in PVQA in particular as it describes interactions between stimuli. Masking occurs when a stimulus that is visible by itself cannot be detected due to the presence of another. Sometimes the opposite effect, facilitation, occurs: a stimulus that is not visible by itself can be detected due to the presence of another. Within the framework of quality assessment it is helpful to think of the distortion or coding noise being masked (or facilitated) by the original image or sequence acting as background. Masking explains why similar coding artifacts are disturbing in certain regions of an image while they are hardly noticeable elsewhere.

9.1. Spatial masking

Many vision models are limited to intra-channel masking, assuming that masking occurs only between stimuli located in the same channel. However, more recent psychophysical experiments suggest that masking also occurs between channels of different orientation [33], between channels of different spatial frequency, and between chrominance and luminance channels [17,64,106].

Models have been proposed which explain a wide variety of empirical contrast masking data within a process of contrast gain control. These models were inspired by analyses of the responses of single neurons in the visual cortex of the cat [5,42,43], where contrast gain control serves as a mechanism to keep neural responses within the permissible dynamic range while at the same time retaining global pattern information.

Contrast gain control can be modeled by an excitatory nonlinearity that is inhibited divisively by a pool of responses from other neurons. Masking occurs through the inhibitory effect of the normalizing pool [33,109]. Watson and Solomon

[119] recently presented an elegant generalization of these models, which permits an easy integration of many kinds of channel interactions and spatial pooling: Let $a = a(t, c, f, \theta, x, y)$ be a coefficient of the perceptual decomposition in temporal channel t , color channel c , frequency band f , orientation band θ , at location x, y . Then the corresponding sensor output $s = s(t, c, f, \theta, x, y)$ can be computed as

$$s = k \frac{a^p}{\sigma^2 + a^q * h}. \quad (3)$$

The excitatory path in the numerator consists of a simple power-law nonlinearity with exponent p . The inhibitory path in the denominator controls the gain of the excitatory path. It also includes a nonlinearity with a possibly different exponent q . Additionally, filter responses are pooled over different channels in the inhibitory path by virtue of a convolution with the pooling function $h = h(t, c, f, \theta, x, y)$, for example a Gaussian kernel [119]. In its most general form, this pooling operation may combine coefficients from the dimensions of time, color, frequency, orientation, space, and phase. The saturation constant σ is added to prevent division by zero; k is used to adjust the overall gain of the mechanism. Introduced for luminance images, this contrast gain control model has been used successfully with color images and video by the author [131,132].

In Teo and Heeger's implementation [109], which is based on a direct model of neural cell responses [43], pooling is limited to orientation, and the exponents of both the excitatory and inhibitory nonlinearity are fixed at $p = q = 2$ so as to be able to work with local energy measures. However, this procedure rapidly saturates the sensor outputs, which is why they have to use multiple contrast bands (i.e. several different k s and σ s) for all coefficients in order to cover the full range of contrasts. Watson and Solomon [119] showed that the same effect can be achieved with a single contrast band when $p > q$. This reduces the number of model parameters considerably and simplifies the fitting process.

Although implemented one way or another in most PVQA systems, contrast masking is not the only conceivable masking mechanism and cannot

explain all masking data. The models described above are based on experiments with simple stimuli such as sinusoidal gratings and Gabor patches. With complex stimuli as are found in real scenes, the distortions can be more noise-like, and masking can become much larger [11,30]. Entropy masking has been proposed as a bridge between contrast masking and noise masking, when the distortion is deterministic but unfamiliar [121], which may be a good model for quality assessment by inexperienced viewers. A discussion and comparison of several different models for spatial masking can be found in [58].

9.2. Temporal masking

Temporal masking is an elevation of visibility thresholds due to temporal discontinuities in intensity, for example scene cuts. Within the framework of television, it was first studied by Seyler and Budrikis [96,97], who concluded that the threshold elevation may last up to a few hundred milliseconds after a transition from dark to bright or from bright to dark. More recently, Tam et al. [107] investigated the visibility of MPEG-2 coding artifacts after a scene cut and found significant visual masking effects only in the first subsequent frame. Carney et al. [15] noticed a strong dependence on stimulus polarity, with the masking effect being much more pronounced when target and masker match in polarity. They also found masking to be greatest for local spatial configurations.

Interestingly, temporal masking can occur not only after a discontinuity ("forward masking"), but also before. This "backward masking" may be explained as the result of the variation in the latency of the neural signals in the visual system as a function of their intensity [4]. The opposite of temporal masking, temporal facilitation, can occur at low-contrast discontinuities [38].

So far, the above-mentioned temporal effects have received much less attention in the video coding community than their spatial counterparts. In principle, temporal masking can be taken into account with a contrast gain control model as in Eq. (3) by adding a time-dependency to the pooling function h , as demonstrated by Girod [38]. Watson [118] recently outlined a PVQA system that

models forward masking effects by means of a masking sequence produced by passing the reference through a lowpass filter.

10. Pattern adaptation

Pattern adaptation adjusts the contrast sensitivity of the visual system in response to the prevalent stimulation patterns. For example, adaptation to patterns of a certain frequency can lead to a noticeable decrease of contrast sensitivity around this frequency [39,101,130]. Together with masking, adaptation was one of the major incentives for developing a multi-resolution theory of vision. However, pattern adaptation has a distinct temporal component to it and is not automatically taken into account by a multi-resolution representation of the input; rather, it needs to be incorporated explicitly by adapting the pertinent model parameters. Ross and Speed [92] presented a single-mechanism model that accounts for both pattern adaptation and masking effects of simple stimuli, but PVQA systems have largely ignored this phenomenon.

An interesting study in this respect was carried out by Webster and Miyahara [123]. They used natural images of outdoor scenes (both distant views and close-ups) as adapting stimuli. It was found that exposure to this kind of stimuli induces pronounced changes in contrast sensitivity. The effects can be characterized by selective losses in sensitivity at lower to medium spatial frequencies. This is consistent with the characteristic amplitude spectra of natural images, which decrease with frequency roughly as $1/f$. This is a typical situation when viewing video, and the CSF of the vision model may need to be adjusted so as to take this phenomenon into account.

Likewise, Webster and Mollon [124] examined how color sensitivity and appearance might be influenced by adaptation to the color distributions of images. They found that natural scenes exhibit a limited range of chromatic distributions, so that the range of adaptation states is normally limited as well. However, the variability is large enough so that different adaptation effects may occur for individual scenes and for different viewing conditions.

11. Pooling

The processes described so far take place before or in the primary visual cortex, also referred to as area V1. It is believed that the information represented there in various channels is integrated in the subsequent brain areas, beginning with area V2. This process can be simulated by gathering the data from these channels according to rules of probability or vector summation, also known as pooling. However, little is known about the nature of the actual integration in the brain. As a matter of fact, there is no firm experimental evidence that the mathematical assumptions and equations presented below are a good description of mechanism pooling in the human visual system [36,89].

If there are a number of independent “reasons” i for an observer noticing the presence of a distortion, each having probability P_i respectively, the overall probability P of the observer noticing the presence of the distortion is

$$P = 1 - \prod_i (1 - P_i). \quad (4)$$

This is the probability summation rule. The dependence of P_i on the distortion strength x_i can be described by the psychometric function

$$P_i = 1 - e^{-x_i^{\beta}}. \quad (5)$$

This is one version of a distribution function studied by Weibull [126] and first applied to vision by Quick [89]. β determines the slope of the function. Under the homogeneity assumption that all β_i are equal [79], Eqs. (4) and (5) can be combined to yield

$$P = 1 - e^{-\sum x_i^{\beta}}.$$

The exponent in the above equation is in itself an indicator of the visibility of distortions. Therefore, models may postulate a combination of mechanism responses before producing an estimate of detection probability. Vector summation (also called Minkowski summation) achieves this:

$$x = \sqrt[\beta]{\sum x_i^{\beta}}.$$

Different exponents β have been found to yield good results for different experiments and implementations. $\beta = 2$ was used e.g. in [109,131]; this case corresponds to the ideal observer formalism under independent Gaussian noise, which assumes that the observer has complete knowledge of the stimuli and uses a matched filter for detection. In a study of subjective experiments with coding artifacts, $\beta \approx 2$ was found to give good results [25]. Intuitively, a few high distortions may draw the viewer's attention more than many lower ones. This behavior can be emphasized with higher exponents, which have been used in several other vision models, for example $\beta = 4$ [110,112]. The best fit of a contrast gain control model to masking data was achieved with $\beta \approx 5$ [119].

In any case, the pooling operation need not be carried out over all pixels in the entire sequence or frame. In order to take into account the focus of attention of observers, for example, pooling can be carried out separately for spatio-temporal blocks of the sequence that cover roughly 100 milliseconds and two degrees of visual angle each [112]. Alternatively, the distortion can be computed locally for every pixel, yielding a perceptual distortion map for better visualization of the temporal and spatial distribution of distortions. For demonstration, I encoded the Basketball scene with the MPEG-2 encoder of the MPEG Software Simulation Group at 3 Mbit/s. Fig. 14 shows a sample frame from the sequence and the corresponding distortion map produced by the author's PVQA system [132], which includes temporal aspects of the distortions as well. Such a distortion map can help the expert to locate and identify problems in the processing chain or shortcomings of an encoder, for example. This can be more useful than a global measure in many PVQA applications.

12. Cognitive processes

While the previous sections were concerned mostly with lower-level near-threshold aspects of vision, the cognitive behavior of humans when watching a video cannot be ignored in advanced PVQA systems. However, cognitive behavior may differ greatly between individuals and situations,

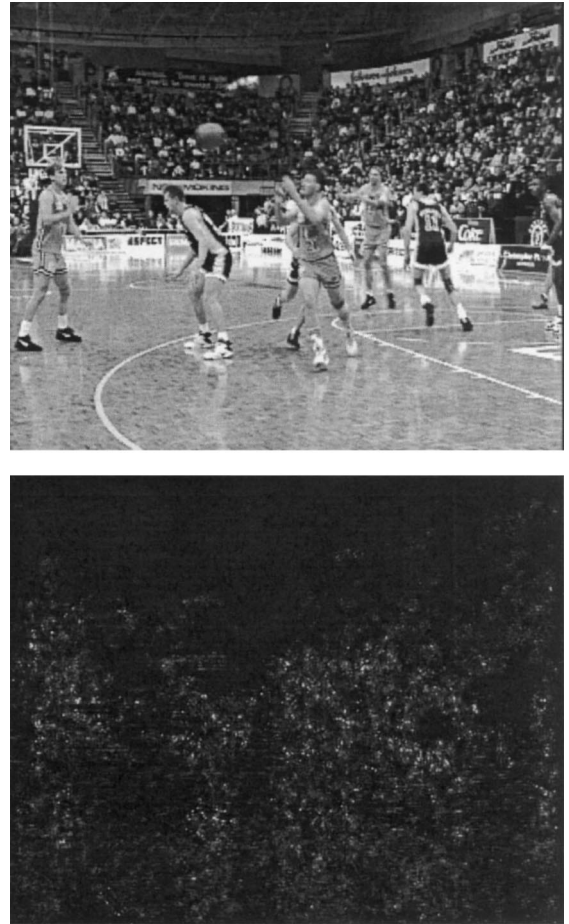


Fig. 14. Sample frame from MPEG-encoded Basketball sequence (top). The distortion map (bottom) contains spatial as well as temporal aspects of impairment visibility [132].

which makes it very difficult to generalize. Nevertheless, I want to point out two important components, the shift of the focus of attention and the tracking of moving objects, which are not unrelated.

When viewing a video, we focus our gaze on particular areas. Studies have shown that the direction of gaze during viewing is not completely idiosyncratic to individual viewers. Instead, a significant number of viewers will focus on the same regions of a scene [31,102,103]. Naturally, this focus of attention is highly scene-dependent. Maeder et al. [70] proposed constructing an

importance map for the sequence as a prediction for the focus of attention, taking into account perceptual factors such as edge strength, texture energy, contrast, color variation, homogeneity, etc.

In a similar manner, viewers may also track specific moving objects in a scene. In fact, motion tends to attract the viewers' attention. Now, the spatial acuity of the human visual system depends on the velocity of the image on the retina: as the retinal image velocity increases, spatial acuity decreases. The visual system addresses this problem by tracking moving objects with smooth-pursuit eye movements, which minimizes retinal image velocity and keeps the object of interest on the fovea. Smooth pursuit works well even for high velocities, but it is impeded by large accelerations and unpredictable motion [29,41]. On the other hand, tracking a particular movement will reduce the spatial acuity for the background and objects moving in different directions or at different velocities. An appropriate adjustment of the spatio-temporal CSF as described in Section 7 to account for some of these sensitivity changes can be considered as a first step in modeling such phenomena [22,127].

13. Evaluation of PVQA systems

Some authors have demonstrated the performance of their video quality metrics by computing the correlation of their system's ratings with subjective ratings of a set of sequences. However, subjectively rated sequences are hardly available in the public domain; the sequences and subjective ratings used in these demonstrations have been mostly proprietary, making it difficult to compare metrics with each other.

In 1997, the Video Quality Experts Group (VQEG) was formed with the objective to collect reliable subjective ratings for a well-defined set of sequences and to evaluate the performance of different video quality assessment systems with respect to these sequences. The goal of this effort is to recommend the video quality assessment system(s) whose quality predictions are in best agreement with subjective ratings. The emphasis of the first phase of VQEG is on distribution-class video, i.e.

mainly MPEG-2 encoded sequences with different profiles, levels and other parameter variations, the bit rates ranging from 768 kbit/s to 50 Mbit/s. In total, 16 conditions and 20 scenes of 8 seconds each were selected and encoded (the scenes were disclosed to the proponents only after the submission deadline). Ten different PVQA systems were submitted, and their output for each of the 16×20 sequences will be recorded. In parallel, DSCQS subjective ratings for all sequences will be obtained by several testing labs. The metrics' predictions will then be compared to the subjective ratings by means of statistical data analysis methods; performance criteria include prediction accuracy and consistency. The participating ITU study groups, ITU-T SG 9, ITU-T SG 12, and ITU-R SG 11, will base their recommendations on the results of this evaluation. An important measure of acceptability will be a comparison of metric prediction errors to rating differences between groups of subjective viewers. As this paper is being published, first results of this effort should become available.¹

14. Conclusions

I have discussed some of the issues in applying vision models to perceptual video quality assessment. Several models have already been proposed and implemented, and the results are quite promising. Nevertheless, some issues regarding the inner workings of the human visual system itself have not yet been resolved satisfactorily and are still under investigation; for others it is not clear how to best incorporate them into a vision model. We are still a long way from having developed or even designed the "perfect" PVQA system that could replace subjective tests. Research in this area is vivid, however, and with the VQEG effort as the first major undertaking to compare and analyze the performance of objective video quality metrics, we are taking another important step in this direction.

¹ Consult the official VQEG web site <http://www.crc.ca/vqeg> for more information and the current status of this effort.

Appendix A

Conversion from CIE 1931 XYZ tristimulus values to CIE $L^*a^*b^*$ and CIE $L^*u^*v^*$ color spaces is defined as follows [134]. The conversions make use of the function

$$f(x) = \begin{cases} x^{1/3}, & \text{if } x > 0.008856, \\ 7.787x + 16/116, & \text{otherwise.} \end{cases}$$

Both CIE $L^*a^*b^*$ and CIE $L^*u^*v^*$ space share a common lightness component L^* :

$$L^* = 116f(Y/Y_0) - 16.$$

The 0-subscript refers to the corresponding unit for the reference white being used. The two chromaticity coordinates u^* and v^* in CIE $L^*u^*v^*$ space are computed as follows:

$$u^* = 13L^*(u' - u'_0), \quad u' = \frac{4X}{X + 15Y + 3Z},$$

$$v^* = 13L^*(v' - v'_0), \quad v' = \frac{9Y}{X + 15Y + 3Z},$$

and the CIE $L^*u^*v^*$ color difference is given by

$$\Delta E_{uv}^* = \sqrt{(\Delta L^*)^2 + (\Delta u^*)^2 + (\Delta v^*)^2}.$$

The two chromaticity coordinates a^* and b^* in CIE $L^*a^*b^*$ space are computed as follows:

$$a^* = 500[f(X/X_0) - f(Y/Y_0)],$$

$$b^* = 200[f(Y/Y_0) - f(Z/Z_0)],$$

and the CIE $L^*a^*b^*$ color difference is given by

$$\Delta E_{ab}^* = \sqrt{(\Delta L^*)^2 + (\Delta a^*)^2 + (\Delta b^*)^2}.$$

By definition, $L^* = 100$, $u^* = v^* = 0$, and $a^* = b^* = 0$ for the reference white.

Appendix B

$Y'C'_B C'_R$ color space is defined in ITU-R Recommendation 601 [51]. Conversion from $Y'C'_B C'_R$ to standard CIE 1931 XYZ tristimulus values requires three steps as illustrated in Fig. 9. $Y'C'_B C'_R$ coding uses 8 bits for each component: Y' is coded with an offset of 16 and an amplitude range of 219,

while C'_B and C'_R are coded with an offset of 128 and an amplitude range of ± 112 . The extremes of the coding range are reserved for synchronization and signal processing headroom, which requires clipping prior to conversion. Nonlinear $R'G'B'$ values in the range $[0,1]$ are then computed from $Y'C'_B C'_R$ as follows:

$$\begin{bmatrix} R' \\ G' \\ B' \end{bmatrix} = \frac{1}{219} \begin{bmatrix} 1 & 0 & 1.3707 \\ 1 & -0.3365 & -0.6982 \\ 1 & 1.7324 & 0 \end{bmatrix} \times \left(\begin{bmatrix} Y' \\ C'_B \\ C'_R \end{bmatrix} - \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} \right).$$

Gamma correction as in Eq. (1) has to be applied to R' , G' , and B' in order to obtain linear RGB values. For displays with standard phosphors, these linear RGB values can then be converted to CIE XYZ tristimulus values as follows:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = \begin{bmatrix} 0.4306 & 0.3415 & 0.1784 \\ 0.2220 & 0.7067 & 0.0713 \\ 0.0202 & 0.1295 & 0.9394 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix}.$$

References

- [1] P.K. Ahnelt, The photoreceptor mosaic, *Eye* 12 (3B) (1998) 531–540.
- [2] A.J. Ahumada Jr., Computational image quality metrics: A review, *SID Symposium Digest* 24 (1993) 305–308.
- [3] A.J. Ahumada Jr., C.H. Null, Image quality: A multi-dimensional problem, in: A.B. Watson (Ed.), *Digital Images and Human Vision*, MIT Press, 1993, pp. 141–148.
- [4] A.J. Ahumada Jr., B.L. Beard, R. Eriksson, Spatio-temporal discrimination model predicts temporal masking function, in: *Proc. SPIE*, Vol. 3299, San Jose, CA, 1998, pp. 120–127.
- [5] D.G. Albrecht, W.S. Geisler, Motion selectivity and the contrast-response function of simple cells in the visual cortex, *Vis. Neurosci.* 7 (1991) 531–546.
- [6] R. Aldridge, J. Davidoff, M. Ghanbari, D. Hands, D. Pearson, Recency effect in the subjective assessment of digitally-coded television pictures, in: *Proc. IPA*, Edinburgh, UK, 1995, pp. 336–339.
- [7] T. Alpert, The influence of the home viewing environment on the measurement of quality of service of digital TV broadcasting, in: *MOSAIC Handbook*, 1996, pp. 159–163.

- [8] American National Standards Institute, Digital transport of video teleconferencing/video telephony signals – performance, terms, definitions, and examples, ANSI T1.801.02, 1995.
- [9] M. Ardito, M. Gunetti, M. Visca, Preferred viewing distance and display parameters, in: *MOSAIC Handbook*, 1996, pp. 165–181.
- [10] R.S. Berns, R.J. Motta, M.E. Gorzynski, CRT colorimetry. Part I: Theory and practice, *Color Res. and Appl.* 18 (5) (1989) 299–314.
- [11] K.T. Blackwell, The effect of white and filtered noise on contrast detection thresholds, *Vision Res.* 38 (2) (1998) 267–280.
- [12] C.A. Burbeck, D.H. Kelly, Spatiotemporal characteristics of visual mechanisms: Excitatory-inhibitory model, *J. Opt. Soc. Amer.* 70 (9) (1980) 1121–1126.
- [13] F.W. Campbell, R.W. Gubisch, Optical quality of the human eye, *J. Physiol.* 186 (1996) 558–578.
- [14] T. Carney, Mindseye: A visual programming and modeling environment for imaging science, in: *Proc. SPIE*, Vol. 3299, San Jose, CA, 1998, pp. 48–58.
- [15] T. Carney, S.A. Klein, Q. Hu, Visual masking near spatiotemporal edges, in: *Proc. SPIE*, Vol. 2657, San Jose, CA, 1996, pp. 393–402.
- [16] W.N. Charman, Optics of the eye, in: M. Bass et al. (Eds.), *Handbook of Optics: Fundamentals, Techniques, and Design*, Vol. 1, Chapter 24, McGraw-Hill, 2nd edn., 1995.
- [17] G.R. Cole, C.F.S. III, R.E. Kronauer, Visual interactions with luminance and chromatic stimuli, *J. Opt. Soc. Amer.* A 7 (1) (1990) 128–140.
- [18] D. Costantini, C.J. van den Branden Lambrecht, G.L. Sicuranza, M. Kunt, Motion rendition quality metric for MPEG coded video, in: *Proc. ICIP*, Vol. 1, Lausanne, Switzerland, 1996, pp. 889–892.
- [19] C.A. Curcio, K.R. Sloan, R.E. Kalina, A.E. Hendrickson, Human photoreceptor topography, *J. Comp. Neurol.* 292 (1990) 497–523.
- [20] C.A. Curcio, K.A. Allen, K.R. Sloan, C.L. Lerea, J.B. Hurley, J.B. Klock, A.H. Milam, Distribution and morphology of human cone photoreceptors stained with anti-blue opsin, *J. Comp. Neurol.* 312 (1991) 610–624.
- [21] S. Daly, The visible differences predictor: An algorithm for the assessment of image fidelity, in: A.B. Watson (Ed.), *Digital Images and Human Vision*, MIT Press, 1993, pp. 179–206.
- [22] S. Daly, Engineering observations from spatiovelocity and spatiotemporal visual models, in: *Proc. SPIE*, Vol. 3299, San Jose, CA, 1998, pp. 180–191.
- [23] J.G. Daugman, Two-dimensional spectral analysis of cortical receptive field profiles, *Vision Res.* 20 (10) (1980) 847–856.
- [24] J.G. Daugman, Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters, *J. Opt. Soc. Amer.* A 2 (7) (1985) 1160–1169.
- [25] H. de Ridder, Minkowski-metrics as a combination rule for digital-image-coding impairments, in: *Proc. SPIE*, Vol. 1666, San Jose, CA, 1992, pp. 16–26.
- [26] H. de Ridder, F.J.J. Blommaert, E.A. Fedorovskaya, Naturalness and image quality: Chroma and hue variation in color images of natural scenes, in: *Proc. SPIE*, Vol. 2411, San Jose, CA, 1995, pp. 51–61.
- [27] R.L. De Valois, E.W. Yund, N. Hepler, The orientation and direction selectivity of cells in macaque visual cortex, *Vision Res.* 22 (5) (1982) 531–544.
- [28] R.L. De Valois, D.G. Albrecht, L.G. Thorell, Spatial frequency selectivity of cells in macaque visual cortex, *Vision Res.* 22 (5) (1992) 545–559.
- [29] M.P. Eckert, G. Buchsbaum, The significance of eye movements and image acceleration for coding television image sequences, in: A.B. Watson (Ed.), *Digital Images and Human Vision*, MIT Press, 1993, pp. 89–98.
- [30] M.P. Eckstein, A.J. Ahumada, Jr., A.B. Watson, Visual signal detection in structured backgrounds. II. Effects of contrast gain control, background variations, and white noise, *J. Opt. Soc. Amer.* A 14 (9) (1997) 2406–2419.
- [31] C. Endo, T. Asada, H. Haneishi, Y. Miyake, Analysis of the eye movements and its applications to image evaluation, in: *Proc. Color Imaging Conference*, Scottsdale, AZ, 1994, pp. 153–155.
- [32] D.J. Field, Relations between the statistics of natural images and the response properties of cortical cells, *J. Opt. Soc. Amer.* A 4 (12) (1987) 2379–2394.
- [33] J.M. Foley, Human luminance pattern-vision mechanisms: Masking experiments require a new model, *J. Opt. Soc. Amer.* A 11 (6) (1994) 1710–1719.
- [34] K.H. Foster, J.P. Gaska, M. Nagler, D.A. Pollen, Spatial and temporal frequency selectivity of neurons in visual cortical areas V1 and V2 of the macaque monkey, *J. Physiol.* 365 (1985) 331–363.
- [35] R.E. Fredericksen, R.F. Hess, Temporal detection in human vision: Dependence on stimulus energy, *J. Opt. Soc. Amer.* A 14 (10) (1997) 2557–2569.
- [36] R.E. Fredericksen, R.F. Hess, Estimating multiple temporal mechanisms in human vision, *Vision Res.* 38 (7) (1998) 1023–1040.
- [37] B. Girod, Ein Modell der menschlichen visuellen Wahrnehmung zur Irrelevanzreduktion von Fernseh luminanzsignalen, PhD thesis, Universität Hannover, Germany, 1988, Published as VDI Fortschritt-Berichte, Reihe 10, Nr. 84.
- [38] B. Girod, The information theoretical significance of spatial and temporal masking in video signals, in: *Proc. SPIE*, Vol. 1077, Los Angeles, CA, 1989, pp. 178–187.
- [39] M.W. Greenlee, J.P. Thomas, Effect of pattern adaptation on spatial frequency discrimination, *J. Opt. Soc. Amer.* A 9 (6) (1992) 857–862.
- [40] S.T. Hammett, A.T. Smith, Two temporal channels or three? A re-evaluation, *Vision Res.* 32 (2) (1992) 285–291.
- [41] P.J. Hearty, Achieving and confirming optimum image quality, in: A.B. Watson (Ed.), *Digital Images and Human Vision*, MIT Press, 1993, pp. 149–162.

- [42] D.J. Heeger, Half-squaring in responses of cat striate cells, *Vis. Neurosci.* 9 (1992) 427–443.
- [43] D.J. Heeger, Normalization of cell responses in cat striate cortex, *Vis. Neurosci.* 9 (1992) 181–197.
- [44] E. Hering, *Zur Lehre vom Lichtsinne*, Carl Gerolds & Sohn, Vienna, Austria, 1878.
- [45] R.F. Hess, R.J. Snowden, Temporal properties of human visual filters: Number, shapes and spatial covariation, *Vision Res.* 32 (1) (1992) 47–59.
- [46] Y. Horita, M. Katayam, T. Murai, M. Miyahara, Objective picture quality scale for video coding, in: *Proc. ICIP*, Vol. 3, Lausanne, Switzerland, 1996, pp. 319–322.
- [47] R.W.G. Hunt, *The Reproduction of Colour*, Fountain Press, 5th edn., 1995.
- [48] L.M. Hurvich, D. Jameson, An opponent-process theory of color vision, *Psych. Rev.* 64 (1957) 384–404.
- [49] J.K. Ijspeert, T.J.T.P. van den Berg, H. Spekreijse, An improved mathematical description of the foveal visual point spread function with parameters for age, pupil size and pigmentation, *Vision Res.* 33 (1) (1993) 15–20.
- [50] ITU-R Recommendation BT.500-7: Methodology for the subjective assessment of the quality of television pictures, ITU, Geneva, Switzerland, 1995.
- [51] ITU-R Recommendation BT.601-5, Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios, ITU, Geneva, Switzerland, 1995.
- [52] D. Jameson, L.M. Hurvich, Some quantitative aspects of an opponent-colors theory. I. Chromatic responses and spectral saturation, *J. Opt. Soc. Amer.* 45 (7) (1955) 546–552.
- [53] D.H. Kelly, Motion and vision. I. Stabilized images of stationary gratings, *J. Opt. Soc. Amer.* 69 (9) (1979) 1266–1274.
- [54] D.H. Kelly, Motion and vision. II. Stabilized spatio-temporal threshold surface, *J. Opt. Soc. Amer.* 69 (10) (1979) 1340–1349.
- [55] D.H. Kelly, Spatiotemporal variation of chromatic and achromatic contrast thresholds, *J. Opt. Soc. Amer.* 73 (6) (1983) 742–750.
- [56] S.A. Klein, Image quality and image compression: A psychophysicist's viewpoint, in: A.B. Watson (Ed.), *Digital Images and Human Vision*, MIT Press, 1993, pp. 73–88.
- [57] S.A. Klein, B. Beutner, Minimizing and maximizing the joint space-spatial frequency uncertainty of Gabor-like functions: Comment, *J. Opt. Soc. Amer.* A 9 (2) (1992) 337–340.
- [58] S.A. Klein, T. Carney, L. Barghout-Stein, C.W. Tyler, Seven models of masking, in: *Proc. SPIE*, Vol. 3016, San Jose, CA, 1997, pp. 13–24.
- [59] J.J. Koenderink, A.J. van Doorn, Spatiotemporal contrast detection threshold surface is bimodal, *Opt. Letters* 4 (1) (1979) 32–34.
- [60] S.R. Lehky, Temporal properties of visual channels measured by masking, *J. Opt. Soc. Amer.* A 2 (8) (1985) 1260–1272.
- [61] J. Liang, G. Westheimer, Optical performances of human eyes derived from double-pass measurements, *J. Opt. Soc. Amer.* A 12 (7) (1995) 1411–1416.
- [62] P.J. Lindh, C.J. van den Branden Lambrecht, Efficient spatio-temporal decomposition for perceptual processing of video sequences, in: *Proc. ICIP*, Vol. 3, Lausanne, Switzerland, 1996, pp. 331–334.
- [63] N. Lodge, An introduction to advanced subjective assessment methods and the work of the MOSAIC consortium, in: *MOSAIC Handbook*, 1996, pp. 63–78.
- [64] M.A. Losada, K.T. Mullen, The spatial tuning of chromatic mechanisms identified by simultaneous masking, *Vision Res.* 34 (3) (1994) 331–341.
- [65] M.A. Losada, K.T. Mullen, Color and luminance spatial tuning estimated by noise masking in the absence of off-frequency looking, *J. Opt. Soc. Amer.* A 12 (2) (1995) 250–260.
- [66] J. Lubin, A visual discrimination model for imaging system design and evaluation, in: E. Peli (Ed.), *Vision Models for Target Detection and Recognition*, World Scientific Publishing, 1995, pp. 245–283.
- [67] J. Lubin, D. Fibush, Sarnoff JND vision model, T1A1.5 Working Group Document #97-612, T1 Standards Committee, 1997.
- [68] F.X.J. Lukas, Z.L. Budrikis, Picture quality prediction based on a visual model, *IEEE Trans. Comm.* 30 (7) (1982) 1679–1692.
- [69] A.M. Lund, The influence of video image size and resolution on viewing-distance preferences, *SMPTE J.* 102 (5) (1993) 407–415.
- [70] A. Maeder, J. Diederich, E. Niebur, Limiting human perception for image sequences, in: *Proc. SPIE*, Vol. 2657, San Jose, CA, 1996, pp. 330–337.
- [71] M.B. Mandler, W. Makous, A three-channel model of temporal frequency perception, *Vision Res.* 24 (12) (1984) 1881–1887.
- [72] J.L. Mannos, D.J. Sakrison, The effects of a visual fidelity criterion on the encoding of images, *IEEE Trans. Inform. Theory* 20 (4) (1974) 525–536.
- [73] D.H. Marimont, B.A. Wandell, Matching color images: The effects of axial chromatic aberration, *J. Opt. Soc. Amer.* A 11 (12) (1994) 3113–3122.
- [74] J.-B. Martens, V. Kayargadde, Image quality prediction in a multidimensional perceptual space, in: *Proc. ICIP*, Vol. 1, Lausanne, Switzerland, 1996, pp. 877–880.
- [75] T. Matsui, S. Hirahara, A new human vision system model for spatio-temporal image signals, in: *Proc. SPIE*, Vol. 1453, San Jose, CA, 1991, pp. 282–289.
- [76] A.A. Michelson, *Studies in Optics*, University of Chicago Press, 1927.
- [77] MOSAIC, A New Single Stimulus Quality Assessment Methodology, RACE R2111, January 1996.
- [78] MOSAIC Final Project Report, Methods for Optimisation and Subjective Assessment in Image Communications, RACE R2111, July 1996.
- [79] J. Nachmias, On the psychometric function for contrast detection, *Vision Res.* 21 (1981) 215–223.

- [80] L.A. Olzak, J.P. Thomas, Seeing spatial patterns, in: K.R. Boff, L. Kaufman, J.P. Thomas (Eds.), *Handbook of Perception and Human Performance*, Vol. 1, Chapter 7, John Wiley & Sons, 1986.
- [81] S. Pefferkorn, J.-L. Blin, Perceptual quality metric of color quantization errors on still images, in: *Proc. SPIE*, Vol. 3299, San Jose, CA, 1998, pp. 210–220.
- [82] E. Peli, Contrast in complex images, *J. Opt. Soc. Amer. A* 7 (10) (1990) 2032–2040.
- [83] E. Peli, In search of a contrast metric: Matching the perceived contrast of Gabor patches at different phases and bandwidths, *Vision Res.* 37 (23) (1997) 3217–3224.
- [84] G.C. Phillips, H.R. Wilson, Orientation bandwidth of spatial mechanisms measured by masking, *J. Opt. Soc. Amer. A* 1 (2) (1984) 226–232.
- [85] A.B. Poirson, B.A. Wandell, Appearance of colored patterns: Pattern-color separability, *J. Opt. Soc. Amer. A* 10 (12) (1993) 2458–2470.
- [86] A.B. Poirson, B.A. Wandell, Pattern-color separable pathways predict sensitivity to simple colored patterns, *Vision Res.* 36 (4) (1996) 515–526.
- [87] C. Poynton, The rehabilitation of gamma, in: *Proc. SPIE*, Vol. 3299, San Jose, CA, 1998, pp. 232–249.
- [88] C.A. Poynton, *A Technical Introduction to Digital Video*, John Wiley & Sons, 1996.
- [89] R.F. Quick Jr., A vector-magnitude model of contrast detection, *Kybernetik* 16 (1974) 65–67.
- [90] S. Rihs, The influence of audio on perceived picture quality and subjective audio-video delay tolerance, in: *MOSAIC Handbook*, 1996, pp. 183–187.
- [91] J.G. Robson, Spatial and temporal contrast-sensitivity functions of the visual system, *J. Opt. Soc. Amer.* 56 (1966) 1141–1142.
- [92] J. Ross, H.D. Speed, Contrast adaptation and contrast masking in human vision, *Proc. R. Soc. Lond. B* 246 (1991) 61–70.
- [93] J.A.J. Roufs, Perceptual image quality: Concept and measurement, *Philips J. Res.* 47 (1) (1992) 35–62.
- [94] J. Rovamo, H. Kukkonen, J. Mustonen, Foveal optical modulation transfer function of the human eye at various pupil sizes, *J. Opt. Soc. Amer. A* 15 (9) (1998) 2504–2513.
- [95] O.H. Schade, Optical and photoelectric analog of the eye, *J. Opt. Soc. Amer.* 46 (9) (1956) 721–739.
- [96] A.J. Seyler, Z.L. Budrikis, Measurements of temporal adaptation to spatial detail vision, *Nature* 184 (1959) 1215–1217.
- [97] A.J. Seyler, Z.L. Budrikis, Detail perception after scene changes in television image presentations, *IEEE Trans. Inform. Theory* 11 (1) (1965) 31–43.
- [98] E.P. Simoncelli, E.H. Adelson, Non-separable extensions of quadrature mirror filters to multiple dimensions, *Proc. IEEE* 78 (4) (1990) 652–664.
- [99] E.P. Simoncelli, W.T. Freeman, The steerable pyramid: A flexible architecture for multi-scale derivative computation, in: *Proc. ICIP*, Washington, DC, 1995, pp. 444–447.
- [100] E.P. Simoncelli, W.T. Freeman, E.H. Adelson, D.J. Heeger, Shiftable multi-scale transforms, *IEEE Trans. Inform. Theory* 38 (2) (1992) 587–607.
- [101] R.J. Snowden, S.T. Hammitt, Spatial frequency adaptation: Threshold elevation and perceived contrast, *Vision Res.* 36 (12) (1996) 1797–1809.
- [102] L.B. Stelmach, W.J. Tam, Processing image sequences based on eye movements, in: *Proc. SPIE*, Vol. 2179, San Jose, CA, 1994, pp. 90–98.
- [103] L.B. Stelmach, W.J. Tam, P.J. Hearty, Static and dynamic spatial resolution in image coding: An investigation of eye movements, in: *Proc. SPIE*, Vol. 1453, San Jose, CA, 1991, pp. 147–152.
- [104] A. Stockman, D.I.A. MacLeod, N.E. Johnson, Spectral sensitivities of the human cones, *J. Opt. Soc. Amer. A* 10 (12) (1993) 2491–2521.
- [105] D.G. Stork, H.R. Wilson, Do Gabor functions provide appropriate descriptions of visual cortical receptive fields?, *J. Opt. Soc. Amer. A* 7 (8) (1990) 1362–1373.
- [106] E. Switkes, A. Bradley, K.K. De Valois, Contrast dependence and mechanisms of masking interactions among chromatic and luminance gratings, *J. Opt. Soc. Amer. A* 5 (7) (1988) 1149–1162.
- [107] W.J. Tam, L.B. Stelmach, L. Wang, D. Lauzon, P. Gray, Visual masking at video scene cuts, in: *Proc. SPIE*, Vol. 2411, San Jose, CA, 1995, pp. 111–119.
- [108] K.T. Tan, M. Ghanbari, D.E. Pearson, An objective measurement tool for MPEG video quality, *Signal Processing* 70 (3) (1998) 279–294.
- [109] P.C. Teo, D.J. Heeger, Perceptual image distortion, in: *Proc. SPIE*, Vol. 2179, San Jose, CA, 1994, pp. 127–141.
- [110] C.J. van den Branden Lambrecht, Color moving pictures quality metric, in: *Proc. ICIP*, Vol. 1, Lausanne, Switzerland, 1996, pp. 885–888.
- [111] C.J. van den Branden Lambrecht, Perceptual models and architectures for video coding applications, PhD thesis, École Polytechnique Fédérale de Lausanne, Switzerland, 1996.
- [112] C.J. van den Branden Lambrecht, O. Verscheure, Perceptual quality measure using a spatio-temporal model of the human visual system, in: *Proc. SPIE*, Vol. 2668, San Jose, CA, 1996, pp. 450–461.
- [113] J.H. van Hateren, A. van der Schaaf, Independent component filters of natural images compared with simple cells in primary visual cortex, *Proc. R. Soc. Lond. B* 265 (1998) 1–8.
- [114] R.L.P. Vimal, Orientation tuning of the spatial-frequency mechanisms of the red-green channel, *J. Opt. Soc. Amer. A* 14 (10) (1997) 2622–2632.
- [115] A.B. Watson, Temporal sensitivity, in: K.R. Boff, L. Kaufman, J.P. Thomas (Eds.), *Handbook of Perception and Human Performance*, Vol. 1, Chapter 6, John Wiley & Sons, 1986.
- [116] A.B. Watson, The cortex transform: Rapid computation of simulated neural images, *Computer Vision, Graphics, and Image Processing* 39 (3) (1987) 311–327.

- [117] A.B. Watson, Perceptual-components architecture for digital video, *J. Opt. Soc. Amer. A* 7 (10) (1990) 1943–1954.
- [118] A.B. Watson, Toward a perceptual video quality metric, in: *Proc. SPIE*, Vol. 3299, San Jose, CA, 1998, pp. 139–147.
- [119] A.B. Watson, J.A. Solomon, Model of visual contrast gain control and pattern masking, *J. Opt. Soc. Amer. A* 14 (9) (1997) 2379–2391.
- [120] A.B. Watson, C.L.M. Tiana, Color motion video coded by perceptual components, *SID Symposium Digest* 23 (1992) 314–317.
- [121] A.B. Watson, R. Borthwick, M. Taylor, Image quality and entropy masking, in: *Proc. SPIE*, Vol. 3016, San Jose, CA, 1997, pp. 2–12.
- [122] A.A. Webster, C.T. Jones, M.H. Pinson, S.D. Voran, S. Wolf, An objective video quality assessment system based on human perception, in: *Proc. SPIE*, Vol. 1913, San Jose, CA, 1993, pp. 15–26.
- [123] M.A. Webster, E. Miyahara, Contrast adaptation and the spatial structure of natural images, *J. Opt. Soc. Amer. A* 14 (9) (1997) 2355–2366.
- [124] M.A. Webster, J.D. Mollon, Adaptation and the color statistics of natural images, *Vision Res.* 37 (23) (1997) 3283–3298.
- [125] M.A. Webster, K.K. De Valois, E. Switkes, Orientation and spatial-frequency discrimination for luminance and chromatic gratings, *J. Opt. Soc. Amer. A* 7 (6) (1990) 1034–1049.
- [126] W. Weibull, A statistical distribution function of wide applicability, *J. Appl. Mech.* 18 (1951) 292–297.
- [127] S.J.P. Westen, R.L. Lagendijk, J. Biemond, Spatio-temporal model of human vision for digital video compression, in: *Proc. SPIE*, Vol. 3016, San Jose, CA, 1997, pp. 260–268.
- [128] G. Westheimer, The eye as an optical instrument, in: K.R. Boff, L. Kaufman, J.P. Thomas (Eds.), *Handbook of Perception and Human Performance*, Vol. 1, Chapter 4, John Wiley & Sons, 1986.
- [129] H.R. Wilson, Spatiotemporal characterization of a transient mechanism in the human visual system, *Vision Res.* 20 (1980) 443–452.
- [130] H.R. Wilson, R. Humanski, Spatial frequency adaptation and contrast gain control, *Vision Res.* 33 (8) (1993) 1133–1149.
- [131] S. Winkler, A perceptual distortion metric for digital color images, in: *Proc. ICIP*, Vol. 3, Chicago, IL, 1998, pp. 399–403.
- [132] S. Winkler, A perceptual distortion metric for digital color video, in: *Proc. SPIE*, Vol. 3644, San Jose, CA, 1999.
- [133] S. Winkler, P. Vanderghenst, Band-limited local contrast, in: *Proc. ICIP*, Kyoto, Japan, 1999, to appear.
- [134] G. Wyszecki, W.S. Stiles, *Color Science: Concepts and Methods, Quantitative Data and Formulae*, John Wiley & Sons, 2nd edition, 1982.
- [135] J. Yang, Do Gabor functions provide appropriate descriptions of visual cortical receptive fields? comment, *J. Opt. Soc. Amer. A* 9 (2) (1992) 334–336.
- [136] J. Yang, W. Makous, Spatiotemporal separability in contrast sensitivity, *Vision Res.* 34 (19) (1994) 2569–2576.
- [137] J. Yang, W. Makous, Implicit masking constrained by spatial inhomogeneities, *Vision Res.* 37 (14) (1997) 1917–1927.
- [138] E.M. Yeh, A.C. Kokaram, N.G. Kingsbury, A perceptual distortion measure for edge-like artifacts in image sequences, in: *Proc. SPIE*, Vol. 3299, San Jose, CA, 1998, pp. 160–172.
- [139] S.N. Yendrikhovskij, F.J.J. Blommaert, H. de Ridder, Perceptually optimal color reproduction, in: *Proc. SPIE*, Vol. 3299, San Jose, CA, 1998, pp. 274–281.
- [140] M. Yuen, H.R. Wu, A survey of hybrid MC/DPCM/DCT video coding distortions, *Signal Processing* 70 (3) (1998) 247–278.
- [141] X. Zhang, B.A. Wandell, A spatial extension of CIELAB to predict the discriminability of colored patterns, *SID Symposium Digest* 27 (1996) 731–735.
- [142] X. Zhang, B.A. Wandell, Color image fidelity metrics evaluated using image distortion maps, *Signal Processing* 70 (3) (1998) 201–214.
- [143] W.Y. Zou, Performance evaluation: From NTSC to digitally compressed video, *SMPTE J.* 103 (12) (1994) 795–800.