ELSEVIER

Signal Processing 70 (1998) 177–200

SIGNAL PROCESSING

Perceptual quality metrics applied to still image compression

Michael P. Eckert[a,*], Andrew P. Bradley[b]

[a] *Faculty of Engineering, University of Technology, Sydney, P.O. Box 123, Broadway, NSW 2007, Australia*
[b] *Canon Information Systems Research Australia, P.O. Box 313, North Ryde, NSW 2113, Australia*

Received 30 July 1998

Abstract

We present a review of perceptual image quality metrics and their application to still image compression. The review describes how image quality metrics can be used to guide an image compression scheme and outlines the advantages, disadvantages and limitations of a number of quality metrics. We examine a broad range of metrics ranging from simple mathematical measures to those which incorporate full perceptual models. We highlight some variation in the models for luminance adaptation and the contrast sensitivity function and discuss what appears to be a lack of a general consensus regarding the models which best describe contrast masking and error summation. We identify how the various perceptual components have been incorporated in quality metrics, and identify a number of psychophysical testing techniques that can be used to validate the metrics. We conclude by illustrating some of the issues discussed throughout the paper with a simple demonstration.

Zusammenfassung

Wir geben einen Überblick über Wahrnehmungsmodelle und ihre Anwendung auf Bildkompression. Der Überblick beschreibt wie Maße der Bildkompression eine Kompressionsmethode lenken können, und deutet Vor- und Nachteile sowie Einschränkungen mehrerer Bildqualitätmaße an. Wir überprüfen einen Umfang Kompressionsmaße, von einfachen mathematischen Maßen bishin zu vollständigen Wahrnehmungsmodellen. Wir heben eine beträchtliche Schwankung der vorgeschlagenen Helligkeits- bzw. Kontrastempfindlichkeitsfunktionen vor, und erörtern eine scheinbare Mangel an Übereinstimmung über die Grundsätze der Kontrastmaskierung und der Fehlersummierung. Wir erörtern dann die Gültigkeitsprüfung dieser Modelle, und besprechen mehrere psychophysische Testverfahren, die die vorgeschlagenen Modelle vergleichen können. Zum Schluß veranschaulichen wir die vorangegangene Diskussion mit einer einfachen Vorführung.

Résumé

Nous présentons un court état des qualités métriques de perception visuels et leur application sur la compression d'images fixes. Cet état de l'art décrit comment des mesures de la compression des images peuvent être utilisées pour guider la phase de compression et souligner les avantages, inconvenients et limitations de certaines mesures de la qualité de l'image. Nous examinons un vaste choix de mesures de complexité diverse allant de la simple mesure mathématique

* Corresponding author.

0165-1684/98/$ – see front matter
PII: S0165-1684(98)00124-8

à la perception complète du modèle. Nous mettons également en valeur les variations dans les fonctions existantes de luminescence et de sensibilité au contraste. Nous evoquons ensuite l'apparant manque de consensus général sur les principes sous-jacents du masquage du contraste et la sommation d'erreurs. Nous discutons ensuite de la manière dont ces modèles ont été validés et nous recommendons un nombre de techniques de tests psychologiques qui devraient être utilisées pour comparer ces modèles. Nous concluons en illustrant à l'aide d'une démonstration simple, certains des problèmes présentés dans l'article.

Keywords: Objective image quality; Perceptual model; Image compression; Quality metric

1. Introduction

After years of image compression research, one of the main problems hindering further development of compression schemes is the lack of a well-accepted metric for the prediction of image quality. The most commonly used metrics still remain simple, mathematically defined measures such as peak signal to noise ratio (PSNR) or mean squared error (MSE). When the quantisation is varied on a single image in a straightforward manner, such as by varying the scale factor in JPEG compression, these metrics do correlate with image quality. However, they often fail to predict image quality when different compression techniques are used. Even more importantly, the metrics do not accurately predict visual quality across a set of images with varying content such as edges, textured regions, and large luminance variations. In response to the failure of standard mathematical metrics, image quality metrics that incorporate perceptual factors to varying degrees have been proposed. The number and variety of these perceptual metrics described in the literature is stunning. Some only use the contrast sensitivity function to weight the importance of spatial frequencies before computing errors, while others are complex multiple frequency channel models replete with non-linearities. Usually, perceptual metrics are reported to provide more consistent estimates of image quality than mathematically defined metrics when artefacts are near the visual threshold. However, the implementation of the metrics is often so complex, and the psychophysical testing required to validate them so time consuming, that a comprehensive validation and direct comparison of performance between metrics is rarely performed. There are, however, some exceptions to this [8,31,57,72]. When tests are performed, they often illustrate disappointing inaccuracies in the predictions of the perceptual quality metrics, albeit when artefacts are significantly above the visual threshold. Our knowledge of visual factors continues to progress, though areas of controversy which affect observer quality ratings need further investigation, particularly contrast masking, summation of distortion artefacts, search strategies, and attention. Because of rapid progress in recent years, one can expect that a model that provides acceptable performance over a wide range of image quality will soon be demonstrated.

There are a number of notable reviews of image quality metrics. In particular, Ahumada [2] provides a succinct summary of perceptual metrics applied to image quality research. Eskicioglu [27] surveys a number of quality metrics, but concentrates primarily on mathematically oriented metrics. Jayant et al. [37] describe how perceptual characteristics have been applied to signal compression, but concentrate on audio rather than image compression. Daly [17] provides a useful discussion of a number of visual factors which should be incorporated in a perceptual metric designed to predict image quality.

The purpose of this paper is to discuss image quality metrics as applied to image compression. We review how the metrics are used in image compression schemes, discuss some of the visual factors which are incorporated in the metrics, describe a number of quality metrics commonly used by image compression researchers, and discuss the difficulties associated with validating them in psychophysical experiments. We conclude with a simple demonstration which illustrates the difficulty that metrics have in predicting image quality.

2. The utility of image quality metrics

One of the primary uses of an image quality metric is to accurately measure the visual quality of the compressed image during the compression process without feedback from the user. Without a metric to assess quality, it is often left to a human observer to manually adjust the quantisation level, usually by multiplying the quantisation matrix with a scale factor, until an acceptable level of visual quality is reached. Ideally, the quality metric should be able to predict the visually lossless compression point as well as provide a perceptually meaningful scale when distortions are significantly above the visual threshold. Such a metric should be validated using psychophysical experiments, a task which can be surprisingly difficult (see Section 5).

Selectable quality image compression can be implemented by feeding the original image and the compressed image into a quality computation, as illustrated in Fig. 1. The output of the perceptual model guides the quantisation until the desired level of perceptual quality is achieved. This process can be implemented using any quality metric, though the efficiency of computation is rather low when each iteration requires the application of quantisation, the inverse transform, and computation of the metric. Computational efficiency becomes a significant issue when using a perceptual metric, because the number of computations will be significantly greater than most mathematical metrics.

The issue of computational efficiency has been addressed by Safranek and Johnston [73] and Watson [89,90], both of whom implemented a perceptual model in the transform domain as illustrated in Fig. 2. This approach is more efficient as the computations are performed in the linear transform domain and the transform need only be applied once. The approach has been modified to work with other linear transforms such as the lapped orthogonal transform (LOT) and the wavelet transform [26]. The primary limitation of this approach is that linear transforms used in compression schemes have deficiencies with respect to the spatial frequency characteristics of human visual channels which can lead to mis-predictions of image quality.

An image quality metric should be able to characterise spatial variations in quality across an image. This is because the visibility of artefacts is highly dependent upon local image content. For this reason, many perceptual metrics provide a 2-D quality map, assigning a level of perceived distortion to each location in an image. The ability to predict local image quality in a 2-D map significantly improves spatially adaptive compression techniques because it allows one to spread quantisation noise over an image in a perceptually uniform manner. This is particularly appropriate for any image that varies in spatial content, such as an image which contains texture, sharp edges, smooth areas, or large intensity differences. Rosenholtz and Watson [69], Tran and Safranek [83], and

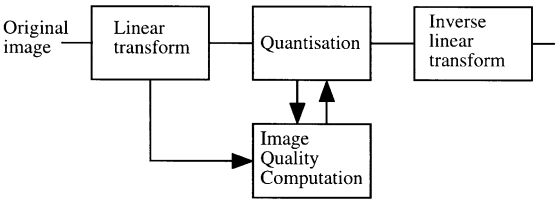

Fig. 2. Optimising quantisation when the quality metric is defined in the transform domain.

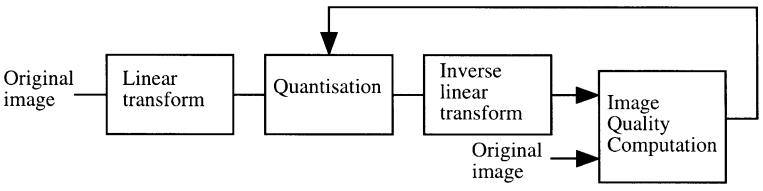

Fig. 1. Optimising quantisation using a quality metric defined in the space domain.

Hontsch and Karam [34,35] describe examples of perceptually guided adaptive compression schemes.

2.1. Some difficulties associated with the design and validation of an image quality metric

While a 2-D quality map is useful as part of an adaptive compression process, a selectable quality compression scheme should only require the user to specify a single quality number for the entire image. As a result, there is a need to collapse the 2-D quality map to produce a single number that reflects overall image quality. Several approaches have been proposed to do this, ranging from taking an average of local error measurements [13], performing a nonlinear summation of errors over an entire image and across all frequency bands [89,95], to specifying quality in terms of the worst quality region in the image [17,26,47]. All of these 2-D to 1-D reductions are based on reasonable arguments in terms of how an observer will rate image quality, but are somewhat artificial unless the observer is instructed to rate image quality in a corresponding fashion in validation experiments. After all, an observer is easily able to see the differences in quality across the image and can change the rating technique depending on what is required in the psychophysical experiment. As a result, the predictive ability of a quality metric, when expressed as a single number for an entire image, is closely tied to the psychophysical methods used to validate the metric.

Care must be taken in using and validating a quality metric at high compression ratios, when compression artefacts are significantly above the threshold of visibility. In this situation, it is possible to undertake the laborious process of a full multidimensional scaling analysis to assess the quality dimensions for suprathreshold compression artefacts [4,49], but the quality dimensions obtained from this analysis will depend on the compression artefacts present in the image set used in the analysis, i.e., a multidimensional scaling analysis performed only with a set of JPEG compressed images would not produce a quality dimension which includes wavelet ringing artefacts. This limits the generality of any model developed using this technique. Furthermore, the objectionability of different types of artefacts will depend on the personal preference of the observer [4,49]. As an example, a blocking artefact from JPEG compression may be more acceptable to one group of observers (such as employees of companies who sell JPEG compression hardware), than it would be to a different group of observers (such employees of companies who sell wavelet compression hardware). One can ignore the differences between observers, and form a model based on the preferences of an "average" observer, in which case there will always be a residual variability in the prediction of image quality. Attempts to deal with preferences for suprathreshold artefacts include directly incorporating them into the model [39,51], or by ignoring them. Most mathematical metrics take the latter approach, as do perceptual metrics based on threshold visual factors [17,47,89]. For the case of perceptual metrics, a metric designed using threshold visual factors, and validated using threshold visual experiments, can be expected to provide accurate predictions of quality at distortion ranges near the visual threshold, but will not always provide good predictions of quality when distortions reach levels where observer preferences become a significant factor. Unfortunately, there is no consensus regarding the distortion levels at which observer preferences begin to play a significant role.

3. Visual factors used in perceptual image quality metrics

There are a number of well accepted perceptual factors which influence the visibility of distortions in an image. Admittedly, even a perceptual attribute as simple as the shape of the contrast threshold curve will change depending on the stimulus configuration used to measure it, but the importance of each of the characteristics listed in this section is well recognised, even if the particular implementation may vary between perceptual metrics. Our knowledge of perceptual factors is patchy, with well accepted models for visual factors such as contrast sensitivity functions and luminance adaptation, and incomplete models for perceptual

factors such as contrast masking and error summation. The purpose of this section is to discuss a few of the primary perceptual factors used in perceptual quality metrics, and to concentrate on the differences in interpretation and implementation of these factors.

3.1. Contrast sensitivity functions

The contrast threshold function (CTF), or its inverse, the contrast sensitivity function (CSF), is the most widely used perceptual attribute for both simple and complex image quality metrics. The CTF defines the contrast at which frequency components become just visible. A reasonable interpretation is that the CTF specifies the internal noise levels across spatial frequencies, thus identifying the relative amount of quantisation that can be applied near the visibility threshold for the same perceptual error. Models of contrast sensitivity curves abound in the literature, and it is only necessary for the designer of a perceptual quality metric to pick one of them [1,9,14,67,94].

An excellent summary of contrast sensitivity curves for various stimulus configurations is provided by Peli [59], who measured contrast sensitivity functions for sine waves with different apertures and Gabor patches of various bandwidths using a variety of temporal windows. The curves illustrate that differences in temporal presentation, stimulus, and stimulus aperture can significantly change the shape of the CSF. In particular, the assumption that the CSF is band-pass with spatial frequency is usually obtained only for sinusoids within a fixed spatial aperture. The CSF measured for Gabor patches with an octave frequency bandwidth is typically more of a low-pass function of spatial frequency [59,88]. As examples, Fig. 3 illustrates the CSF for block DCT basis functions and wavelet basis functions for a background luminance of 20 cd/m^2 [1,62,94]. Note the significant difference in shape at low frequencies. The question is whether to use the low frequency attenuation which arises in contrast sensitivity experiments conducted under fixed aperture conditions. When used in a perceptually optimised compression scheme, this low frequency attenuation implies that quantisation can be increased for low spatial frequencies relative to mid range spatial frequencies. However, increasing the viewing distance will shift the low frequency artefacts to spatial frequencies for which the observer has greater sensitivity. Thus, an artefact which is invisible for a given viewing distance can become visible when viewing distance is increased. Ahumada and Peterson [3] suggest that the best approach, in the context of image compression, is to assume a low-pass function (illustrated by dashed lines) to ensure that quantisation artefacts will become less visible for increasing viewing distances.

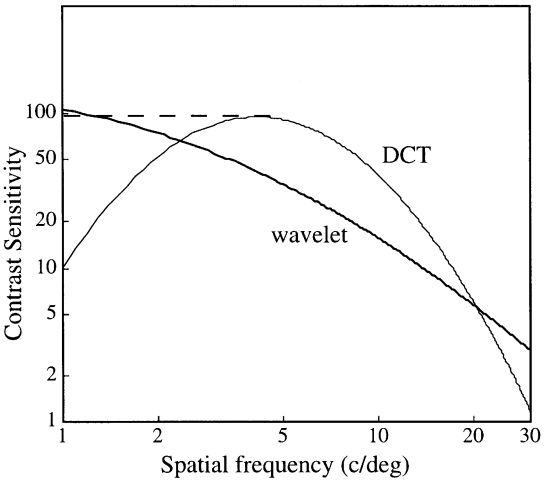

Fig. 3. The CSF for the block DCT basis functions was generated from the model of Ahumada and Peterson [1] with a 20 cd/m^2 background luminance. Ahumada and Peterson [3] suggest modifying the curves to be a low pass function of spatial frequency (dashed lines) to ensure that low frequency artefacts will not be more visible as viewing distance increases. The CSF for wavelet basis functions was generated from the model of Watson et al. [94]. This paper defined contrast threshold in terms of grey levels, so luminance contrast was estimated by assuming that each grey level step is approximately 0.321 cd/m^2 on the 20 cd/m^2 background luminance used in the experiments.

3.2. Luminance adaptation

The second most commonly used perceptual attribute is luminance adaptation. It is well known that sensitivity to intensity differences is dependent

on the local luminance in regions of the image. The basic model for this dependence is the Weber–Fechner law, which states that sensitivity to luminance differences in a stimulus is proportional to the mean luminance of the stimulus (contrast threshold remains constant for increasing luminance levels). The Weber–Fechner regime holds for background luminance levels above approximately 10 cd/m^2 [36]. Below this level the contrast threshold increases as luminance decreases. The importance of luminance masking for compression purposes is that as local luminance increases, an increased level of quantisation can be tolerated. In the Weber–Fechner regime, quantisation of frequency components can be approximately doubled for every doubling of the background luminance for the same perceptual error.

The more interesting aspect of the luminance adaptation is how it is incorporated in the various models. Luminance adaptation can be implemented either in the spatial domain [9,17] or in the frequency domain [1,47,58,89]. In the spatial domain, luminance adaptation is modelled by sending the image through a compressive point non-linearity, typically using a logarithmic, cube root, or square root function before applying the linear transform. Daly [17] comments that using a logarithmic nonlinearity overestimates visual sensitivity in low intensity regions and that a cube root nonlinearity is a better model. A frequency domain implementation of luminance masking is obtained by dividing the AC coefficients by an estimate of the local luminance. As an example, Peli [58] and Lubin [47] implement luminance masking by dividing the energy in a frequency band by an estimate of local luminance obtained from low pass filtered version of the image. Similarly, Watson [89] scales the frequency coefficients in the block DCT by an estimate of local luminance obtained from the DC coefficient in each block. In general, a spatial domain or frequency domain implementation of luminance masking will make only a small impact in predictions of the models, except in cases where image contrast is large [43], or spatial localisation of luminance masking becomes significant.

It is well known that luminance masking is a spatially localised phenomenon, but the effect of variations between local luminance levels and global (mean) luminance has not been extensively investigated. Most measurements of contrast sensitivity are performed with the background luminance equal to the mean luminance of the stimulus, but in complex images there are often large differences in local luminance. One of the few experiments that allowed for the background luminance to be different than the local luminance around the stimulus is the contrast threshold curve model of Rogers and Carel [67], also reported in [12,14]. Their results suggest that when local luminance is significantly less than the surrounding luminance in the image, the standard model of luminance masking overpredicts human visual sensitivity. Fig. 4 shows the amplitude and contrast threshold for a sinusoid as the local luminance changes. The top graph illustrates the changes in the amplitude threshold when the average background luminance remains fixed at 50 cd/m^2 ($L_t = 50$) and the local luminance in the region of the stimulus varies over a wide range. The bottom graph illustrates the amplitude threshold when local luminance varies with the average luminance ($L_t = L_b$). Note the large increase in the contrast threshold for luminance levels below 10 cd/m^2, indicating that the curve exits the Weber–Fechner region below this point.

An issue which seems to be occasionally forgotten among engineers, though not visual scientists, is that most psychophysical results are defined in terms of image luminance (with SI units of candelas per meter^2), not an arbitrary grey level representation. Because CRT monitors have an expansive nonlinear mapping between grey levels and luminance, images are usually represented in gamma compressed form, i.e., $g = 255(L/L_{\max})^{0.45}$, where L is displayed luminance, $L_{\max}$ is the maximum luminance on the monitor, and g is the grey level. One could define a quality metric in the gamma domain, but this makes a priori assumptions regarding both the display and image representation. This assumed relationship does not hold for many applications, and image grey levels may be in a gamma domain, density domain, reflectance domain, or film domain with a specific H&D curve. Printers often operate in a density (log) domain. Scanners, such as a laser scanner for X-ray film (LUMISCAN 100, Lumysis Corp Sunnyvale, CA), may operate in either a density domain, or in

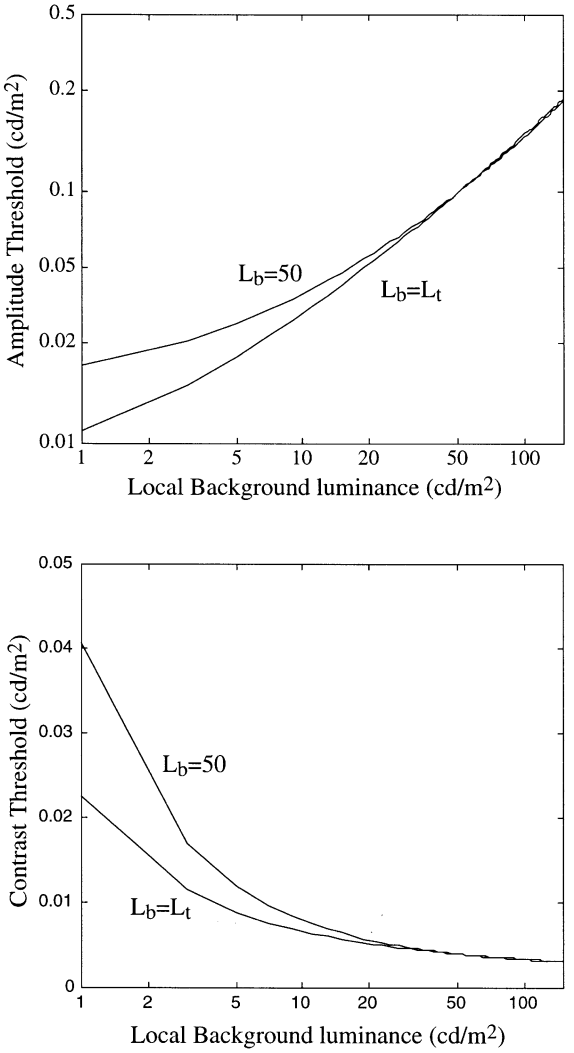

Fig. 4. Amplitude (top figure) and contrast (bottom figure) thresholds for a sinusoid under two cases from the contrast threshold model of Rogers and Carel [67]. The $L_b = L_t$ curves illustrate the case when the average luminance around the target is the same as the background luminance for the entire image. The $L_b = 50$ curves illustrate the case when the background luminance is constant at $50\,\mathrm{cd/m^2}$ and the local average luminance varies. Luminance masking in most images will have a curve more analogous to the top curve since the local luminance varies significantly across most images.

a linear domain if the scanner uses CCD elements. LCD displays, which are becoming increasingly common, have a linear mapping between grey level and luminance. Perceptually linearised displays, proposed for use in medical imaging applications [12] have a well defined relationship between grey level and luminance which is significantly different than for most CRT monitors. Because the relationship between grey levels and luminance cannot be guaranteed, we argue that a perceptual quality metric which is to be used for a broad number of applications should accept a luminance image as the input and make explicit the grey level to luminance mapping for specific applications.

3.3. Linear transforms: their use in image compression and perceptual quality metrics

Most psychophysical evidence suggests that human vision consists of a number of parallel visual channels [56]. These channels are selective to spatial frequency with approximately an octave bandwidth and to orientation with a sensitivity between 15 and 60 degrees, depending on the stimulus used in the experiment. A number of requirements and desirable properties for linear transforms used to model the frequency selective nature of human vision have been described in both Watson [87] and Daly [17]. In summary, they include frequency and orientation selectivity, linear and/or quadrature phase, minimum overlap between adjacent channels (minimal aliasing), unity frequency response, shift invariance, and scale variancy (small spatial extent at high frequencies). In addition, a number of mathematical properties such as invertability and orthogonality are advantageous. Most linear transforms in common use meet some, but not all, of these properties. For example, the wavelet transform and other quadrature mirror filters only have three orientation sensitive channels (0, 90 and 45/135 degrees) and when they are critically sampled, they are not shift invariant. Gabor transforms, on the other hand, are not easily invertible and do not normally have unity frequency response. Block transforms, such as the DCT, are not scale variant, i.e., high frequencies have large spatial extent, and are also not selective to diagonal frequencies. Relatively few transforms meet all of these requirements and the transforms that are commonly used in image compression usually have more than one failing. The first transform which did meet all of

these requirements was the cortex transform [87] which was used, and subsequently modified, by Daly [17] in his perceptual metric. More recently, the shiftable pyramid [76], which uses steerable filters [30] meets all of the above requirements in addition to being computationally efficient and self inverting.

There are slight differences in the implementation of the linear transforms in the perceptual models. Daly [17] applies a series of filter stages, namely a contrast sensitivity filter followed by a frequency and orientation selective filter bank. This approach allows one to easily combine existing contrast sensitivity models with different filter banks. Watson [89], Safranek and Johnston [73], and Chou and Li [16] simply weight each frequency band by the contrast threshold of the basis functions in the transform. The latter approach is more computationally efficient as well as providing a method of direct psychophysical validation of the first stages of the perceptual model, but requires a significant amount of psychophysical testing to identify the visual system weighting for each filter band.

3.4. Masking: contrast masking, noise masking, and mutual masking

Contrast or pattern masking is a phenomenon whereby a signal can be masked, i.e., its visibility reduced, by the presence of another signal. In the context of compression, we are interested in the ability of image content to mask quantisation noise. For an image signal to maximally mask a noise signal, both signals must occur in approximately the same spatial location, be of approximately the same spatial frequency, and have their spatial frequencies in approximately the same orientation. Both psychophysical and physiological experiments of visual masking led to the development of perceptual processing models as parallel, octave bandwidth, and orientation selective visual channels. Some of this work is detailed in Sakrison [75] and Mostafavi and Sakrison [52].

Taking advantage of masking in a perceptual metric is fraught with difficulty, and incorrect predictions of contrast masking are likely to be a major reason why perceptual metrics fail. The reason for this is that masking results obtained in experiments are highly dependent on masker and target stimulus used. Masking thresholds will vary depending upon whether the masker/target has narrow or broad bandwidth, the target/masker phase, target/masker orientation, and the familiarity of the target/masker to the observer. The accepted model of contrast masking [44,77], estimates the degree to which a sinusoidal target is masked by the presence of sinusoidal maskers. This model has since been modified by Foley [28], Foley and Boynton [29], Teo and Heeger [81], and Watson and Solomon [91]. The modifications suggest that masking depends not only on the energy within a band, but also the energy in bands at other orientations. These models correctly predict the threshold elevation for contrast detection or contrast discrimination for a signal, typically a sinusoid or Gabor patch, in the presence of a masker, a sinusoid or Gabor patch of a different frequency, phase and contrast.

Another body of research examined the masking of a signal (typically a sinusoid) with additive broadband noise. This research showed that the elevation of contrast discrimination was proportional to the added noise energy [45,52,60]. Noise masking experiments predict significantly larger amounts of masking than found for contrast masking of narrowband signals. The situation is made even murkier when considering the results of Swift and Smith [80], who showed that if an observer was given enough time to become familiar with a noise mask, the contrast masking elevation for noise masks reduced to that of a sinusoid. This issue was recently reinvestigated by Watson et al. [92], who used maskers such as a white noise masker, band-pass noise masker, cosine masker, and even an image as a masker. They also found that learning could bring the noise masking effect down to the same level as the cosine masker, essentially confirming the results of Swift and Smith. These results suggest that the degree of masking experienced in local regions in an image will depend on the familiarity of the image to the observer. When an image is first shown to an observer, simple image structures such as edges or curves will have only a small degree of masking compared to textured regions, even if the energy content is

similar, because the edge region is simpler and the observer typically has prior information about what an edge looks like. A texture is less predictable and more difficult to learn, and thus one would expect significant masking in this region. However, the masking in a textured region will lessen as an observer becomes familiar with the image through repeated observations. This presents a problem to the designer of a perceptual metric, because one cannot predict the familiarity of an image to the observer, and thus the amount of contrast masking. Daly [17] settled on an interesting compromise, setting the contrast masking parameter for the base-band (low frequencies) to that for a sinusoid masker and for middle and high frequencies to that of a noise like masker. Daly's rationale is that an area in an image which contains energy at mid and high frequencies will be less familiar or predictable, and thus more difficult to "learn" in the context of the Swift and Smith [80] and Watson [92] results, so the image content is better modelled as a noise like masker. However, for specific instances of image regions consisting of only low frequencies, these regions are quite predictable and easily "learned", even without prior familiarity with the specific image, so masking is better modelled using data based on sinusoidal maskers. The learning effect also has significant consequences for the psychophysical validation experiments. As an image is repeatedly displayed during an experiment, the observer will become more familiar with the image, and the level of masking will reduce as the number of presentations increase. This means that caution must be used when evaluating the just noticeable difference compression point using repeated displays of a single image.

Fig. 5 provides a simple illustration of contrast masking. In the figure, a Gabor stimulus is added to a constant background, a white noise background, a $1/f$ noise background, and a "mountain image" background. The backgrounds provide differing amounts of masking: The stimulus is well masked by the $1/f$ noise background, somewhat masked by the white noise background and the "mountain image" background, and not at all masked by the constant background. The local variance in the vicinity of the distortion is the same for all backgrounds except the constant background, and the "in band" noise is the same for the "mountain image" and $1/f$ noise image, but significantly lower for the white noise background. The differences in masking between the white noise background and $1/f$ noise background can be explained in terms of the differing amounts of "in band" noise. However, the higher visibility of the Gabor stimulus in the "mountain image" background compared to the $1/f$ noise image is difficult to explain, except for the fact that the "mountain image" has a more familiar or predictable structure, and thus there is less masking than would be anticipated from the level of "in band" noise.

Another aspect of masking which needs further investigation is the spatial extent of masking. The standard approach is to model masking as a spatially localised phenomenon, so that the response of a filter at a location in space is masked only by the response of itself and other filters at the same location. However, recent evidence suggests that masking effects are not completely localised, and may extend up to eight times the wavelength of the centre frequency band [24]. This would significantly change the masking models as presently implemented in most perceptual metrics.

Daly [17] describes the phenomenon of "mutual masking", which essentially states that the level of masking should take into account both the original image and the compressed image. For example, if the original image contains a textured area, one might assume that we could expect a significant masking to occur in that area. However, if quantisation in the compression scheme significantly reduces the contrast of the texture, as often happens in quantisation schemes which possess a "dead zone" for the frequency coefficients, then this assumption would be incorrect. In this case, little or no visual masking will occur, and the observer will report differences in image quality. This reduction or removal of contrast in textured regions is often the first visible sign of compression artefacts in X-ray bone radiographs which contain large regions of low contrast texture (mottle). In addition, high frequency compression artefacts may be introduced into an area which contains little or no energy at high frequencies, in which case the artefact will be highly visible because of the lack of masking at high frequencies. A common example is

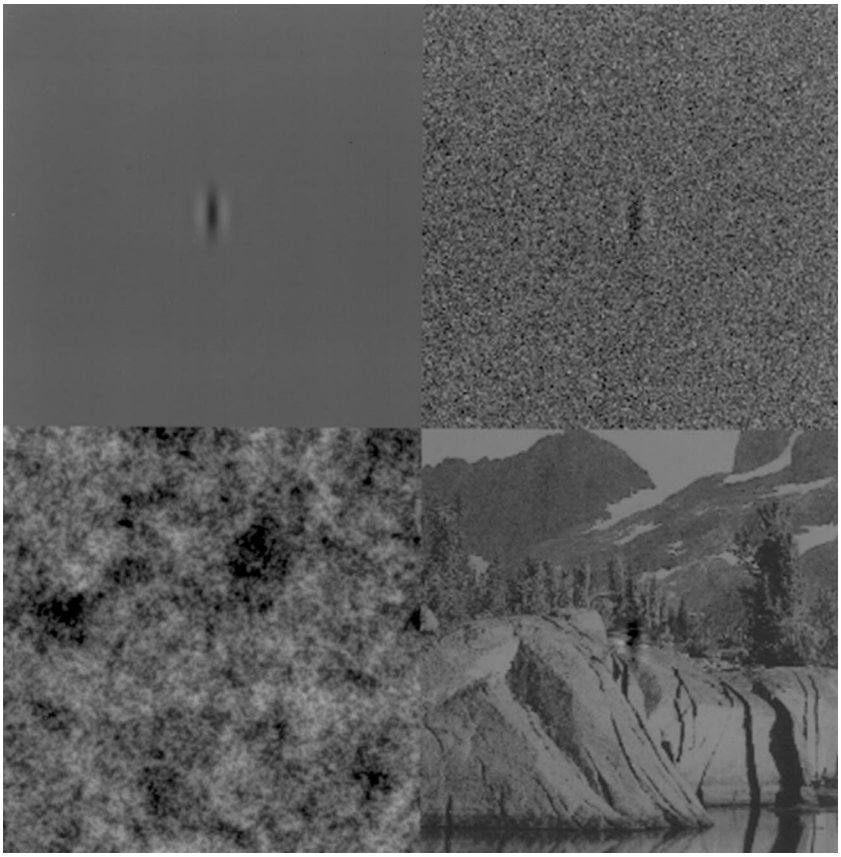

Fig. 5. Example of a Gabor stimulus obscured by various maskers. Top left is uniform background, top right is random white noise, bottom left is noise with a $1/f$ magnitude spectrum, and bottom right is an image. With the exception of the constant background, all the maskers contain approximately the same variance.

JPEG block artefacts appearing in regions containing a smooth luminance gradient. This situation can be handled by using both the original and compressed image to compute the level of masking and taking the minimum of the two. The application of mutual masking in a perceptual metric is essential for compression when artefacts are significantly above the visual threshold, but will have less effect for compression at or near the visual threshold.

3.5. Summation of errors

The standard model of visual processing assumes that the image is filtered using a bank of parallel band-pass channels (filters) to produce a filtered image for every channel. As an example, if one assumes the filter bank consists of one low-pass and three band-pass frequency bands at four orientations, then there will be a total of thirteen filters across all orientations and scales. Using this filter bank in a perceptual metric results in a bank of thirteen visible distortion maps. Not surprisingly, this is a glut of information, even if the channels are critically sampled, and some assumptions must be made about how these distortions can be combined into a single map, and then perhaps into a single number.

The logical way of reducing the dimensionality is to sum the responses in the frequency bands across frequency or space or both. Most approaches to

reducing the dimensionality sum errors across frequency bands to obtain a 2-D visible difference map, and if required, then sum across space to obtain a single number [13,17,47]. The exception to this is Watson [89] who applies summation across space as the first step followed by summation across frequency bands. Preferably, the summation rule should be analogous to the summation rule used in vision. Probability summation, a non-linear (exclusive or) summation rule, is the most well accepted basis for summation of signal energy across frequency channels and across spatially distributed channels, though this has only been verified when signals are near threshold. An approximation to probability summation, as well as energy summation, is the Minkowski metric [65],

$$M = \left(\sum_i |s_i|^\beta \right)^{1/\beta},$$

where s_i is the response of a single frequency band at a specific location in space, the index i refers to channels distributed across space, frequency or both, and β is the summation parameter. Energy summation is modelled with $\beta = 2$, probability summation is well modelled with $\beta \approx 3.5$, and a MAX operator is obtained for $\beta = \infty$.

3.5.1. Summation across frequency bands

The psychophysical evidence suggests that summation across frequency bands is best modelled as probability summation, approximated using the Minkowski metric with $\beta \approx 3.5$ [85]. Recent work with DCT basis functions confirms summation across frequency channels, but found that the summation rule is better modelled with a summation parameter of $\beta = 2.4$, with a range between 2 and 3 [63]. Daly [17] directly implements probability summation, which is essentially equivalent to using a Minkowski metric with $\beta = 3.5$, Lubin [47] uses a Minkowski metric with $\beta = 2.4$, and Watson [89,90] selects the maximum value across all frequency bands which is essentially $\beta = \infty$. Needless to say, this is an area which requires further research, particularly regarding the summation of broadband and suprathreshold signals.

3.5.2. Summation across space

The model for the summation of errors across space is as murky as the model for summation across frequencies. The evidence that it exists is clear [66], but the spatial extent of the summation and summation parameter seems to depend on both eccentricity dependent changes in the CSF as well as whether the signal is coherent or noncoherent [52]. As an example, for spatial summation of DCT coefficients, Peterson et al. [64] found that the summation parameter changed significantly as the spatial extent increased. This dependence of summation on eccentricity agrees with the qualitative observation that image distortions are not perceived as being significantly worse when an image is increased to very large sizes, and that closely spaced errors are more perceptible than spatially distributed errors. When a single number is required as an output, both Lubin [47] and Daly [17,20] implement summation across space as a MAX operator ($\beta = \infty$). Watson [89,90] sums across space using a Minkowski metric with $\beta = 3.5$. In a comparison of various metrics applied to target detection problems [68], it was found that $\beta = 4$ provided the best prediction of psychophysical results.

As mentioned in Section 2.1, the technique for summation of errors across space may have to be modified when artefacts are significantly above the visual threshold. In this regime, the observer will be able to easily identify artefacts at different locations in the image, and the summation rule used by the observer will depend on the instructions given for the psychophysical experiment. As an example, the observer can be instructed to rate image quality based strictly on the location in the image with the largest distortion ($\beta = \infty$), or the instructions could be to rate the "average" level of quality over the entire image ($\beta = 1$ to 3).

4. Image quality metrics

In this section, we place the various image quality metrics into a number of broad classes. First, there are the mathematical metrics which measure quality in terms of relatively simple mathematical functions, usually with point processing. Second,

there are models which incorporate simple perceptual characteristics, such as the contrast sensitivity function and luminance adaptation. Third, there are the models which incorporate perceptual characteristics which include suprathreshold preferences of observers, but also use measurements of image characteristics, such as texture content, edge content, smoothness, spectral slope, etc. Finally, there are perceptual metrics which attempt to model early visual processing in as complete a manner as possible and thus provide a perceptually meaningful measure of image quality near the visual threshold.

4.1. Mathematical metrics

A number of mathematically defined metrics have been used in the literature, including signal to noise ratio (SNR), peak signal to noise ratio (PSNR), mean absolute error (MAE), mean squared error (MSE), local mean squared error, and distortion contrast [23,31,32,71]. These metrics perform well when using images with constraints on the image content or for particular stimulus configurations. As an example, Limb [46] showed that mean squared error works well for predicting error visibility in smooth areas. However, extensive evaluation of these metrics has shown that they do not work well across images which contain significantly different content [31]. In the context of a compression scheme that attempts to implement selectable quality compression, this means that there will be significant quality variations from image to image when compressed to the same predicted level of quality, e.g., the same PSNR or MSE.

The primary advantage of mathematical quality metrics is their ease of use. They do not require any information about viewing conditions, do not adapt to local image content, and the computations are simple. Interestingly, it has been argued that mathematical metrics are superior to perceptual metrics because they do not depend on viewing conditions or image content. It is not clear why this is argued as being desirable since viewing conditions and image content clearly play a major role in human perception of image quality. The fact that a metric is defined without consideration for visual factors does not make visual factors insignificant, rather it simply ignores them, and relies on luck to provide correlation with perceived quality. One apparent argument used against mathematical metrics is that they typically provide a single number for the entire image, and thus cannot reflect spatial variations in image quality. However, any of the metrics described above could be easily modified to operate on local regions in a sliding window and thus provide such a spatially varying quality map. An example of this is the local mean squared error metric used by Girod [32].

4.2. Metrics which incorporate the CSF and luminance adaptation

One of the first attempts to incorporate visual characteristics into an image quality metric is that of Mannos and Sakrison [48], who used two well-known aspects of visual perception, namely luminance adaptation and the CSF. However, the model has only a single visual channel, does not incorporate contrast masking, and so would fail to accurately predict quality in many circumstances. This metric is equivalent to a weighted mean squared error metric.

More recent metrics include those proposed by Nill [53,54] and Saghri [74]. These metrics include a compressive nonlinearity for luminance adaptation, filter the image according to the contrast sensitivity function, and then calculate a difference metric between the original and compressed images. Ahumada [5] also incorporates estimates of image contrast in an attempt to capture some of the properties of contrast masking.

Peli [58] describes a metric which incorporates CSF weighting as well as localised luminance masking, but no contrast masking within the frequency bands. Zetsche and Hauske [96] incorporate CSF weighting, and use a ratio of Gaussian filter bank to implement luminance and contrast masking. Both of these metrics include a multiple frequency band decomposition with no orientation selectivity.

Recent metrics include attempts to incorporate visual factors into wavelet compression schemes

[6,10,42,55]. These schemes weight the wavelet bands by the contrast sensitivity function, and implement luminance adaptation by providing the metric with a gamma compressed version of the image (though this is not explicitly stated in the references).

4.3. Metrics which incorporate observer preferences for suprathreshold artefacts

There are a number of quality metrics which incorporate threshold perceptual factors, somewhat ad hoc mathematical measures, and penalties for specific suprathreshold artefacts. Mathematical measures include "correlation quality" or "local image activity". Suprathreshold compression artefacts which invoke penalties include blocking artefacts of JPEG compression or ringing artefacts around edges [11,39]. As an example of such a metric, we can consider the picture quality scale (PQS) developed by Miyahara et al. [51]. They incorporate luminance adaptation, the CSF, block artefacts, edge artefacts, and local summation of errors in a 5×5 pixel window. Each factor is weighted in a linear regression model to obtain a single quality number for the image. The spatial masking factors allow the model to predict differences in observer judgements of quality for smooth regions, edge regions and textured regions. Unlike many of the metrics which only incorporate threshold level perceptual factors, these metrics attempt to explicitly model the objectionability of suprathreshold compression artefacts. However, none of the above metrics were implemented using a formalised multidimensional scaling analysis [49]. While there is wide variability in the implementation of masking for these types of models, as would be expected in attempts to model observer preferences, there is still significant agreement on threshold visual factors, namely the CSF and luminance adaptation.

4.4. Threshold perceptual metrics

The development of computational models culminated in the early 1990s with two computational models of early vision based on psychophysical and physiological evidence. The models are able to predict threshold visibility levels of distortions for both simple and complex stimuli. Complete descriptions of these models are provided in Daly [17] and Lubin [47]. These two papers contain in-depth descriptions of the development of general purpose perceptual metrics and the psychophysical evidence on which they are based. As far as the authors are aware, no other models exist which purport to provide as consistent a prediction of the visibility of distortions near the visual threshold as the models described in these two papers.

Lubin and Daly's models are quite similar and include modelling the visual system as a series of linear and non-linear stages, including a filter model for the CSF, octave bandwidth frequency decomposition in oriented bands, luminance masking, contrast masking, a distance computation, and summation of errors across frequency bands to produce a visible difference map. A block diagram of the steps in the Daly perceptual model is provided in Fig. 6. The result of both models is a visible difference map which specifies the probability of seeing a difference between the two images at each pixel location. The Lubin model has been validated against psychophysical experiments, and has been shown to predict psychometric curves for simple contrast discrimination experiments and edge sharpness discrimination. Daly [18,19] reports extensive validation using known psychophysical results.

It is interesting to examine the aspects of visual processing which these models do not include. Specifically, neither the Daly nor the Lubin model include masking across orientations. They only implement contrast masking using energy within a band at a single orientation, which is in agreement with the known literature at the time the models were developed. As mentioned previously, recent developments suggest that significant masking can occur across orientations. Second, the models predate the work of D'zmura and Singer [24], so neither model extends masking effects using spatially adjacent responses. Finally, neither model incorporates summation of errors across space when computing the visible difference map, preferring to produce a visible difference map by summing errors across frequency channels.

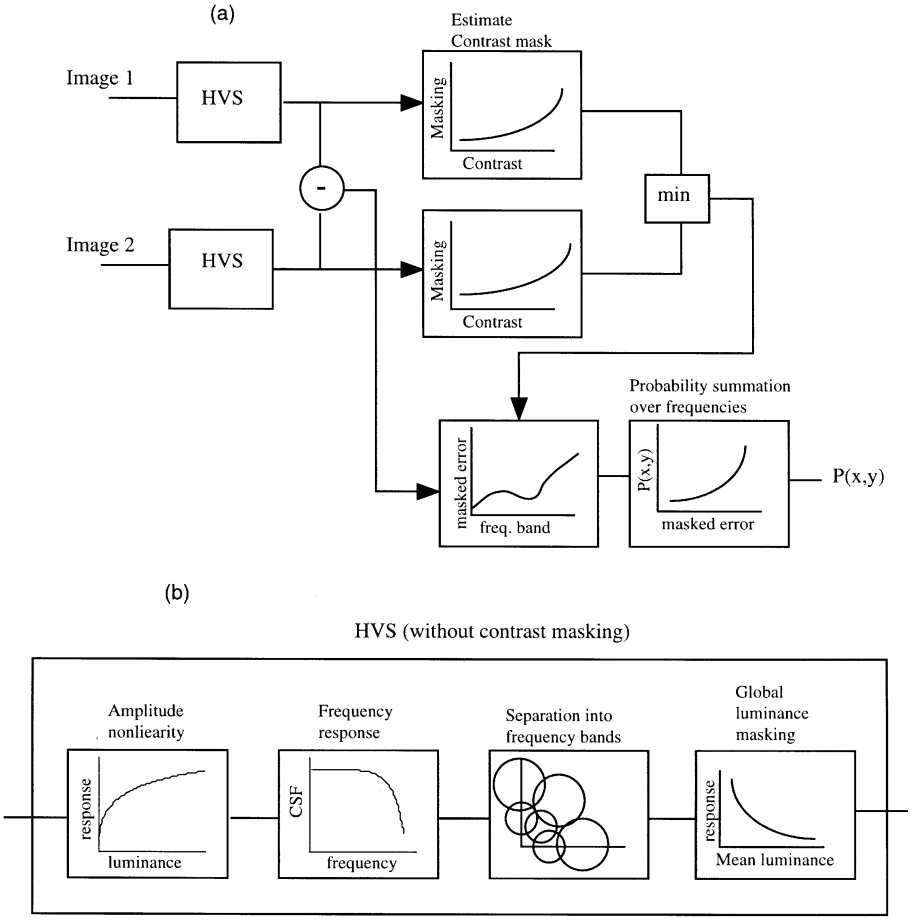

Fig. 6. Sketch of Daly's [17] visible difference predictor. Image 1 and Image 2 are the original and distorted images (in either order). (a). This diagram illustrates the steps involved in computing the visible difference map, $P(x,y)$, which specifies the visibility of errors at every point in the image. The bottom figures illustrates human visual system modelling including local luminance masking, the CSF, the division into multiple frequency bands, and global luminance masking.

One of the more recent, and complete, perceptual metrics is described by Westen et al. [95]. The metric incorporates all of the threshold visual factors discussed earlier, namely weighting by the CSF, luminance adaptation in a manner similar to Peli [58], decomposition into multiple, oriented frequency bands, contrast masking in a manner similar to Daly [17], and summation of errors using the Minkowski metric, first over frequency bands of the same orientation, then over frequency bands over different orientations, and then over space. A significant difference between this model and previous models is that it allows a different summation exponent for different orientations, which has some validity when considering recent contrast masking experiments.

The DCTune algorithm of Watson [89,90] incorporates a perceptual metric, though it attempts only to predict the quality of JPEG compressed images. The Watson model contains similar elements to the Lubin and Daly models, with the exception that it uses block DCT basis functions as its frequency decomposition rather than more physiologically realistic filters. One notable difference between the DCTune algorithm and most other perceptual metrics is that it first implements

summation of errors across space rather than summation of errors across frequency. The output of the model is not a visible difference map, but rather an 8×8 matrix of values specifying the visibility of errors within each frequency band. The advantage of such an approach is that it allows independent modification of each element in a quantisation matrix, but at the cost of losing the ability to create a visible difference map. The reduction of image quality prediction to a single number is performed with a MAX operator (a Minkowski metric with $\beta = \infty$), which means that overall image quality is determined by identifying the frequency band with the most visible artefacts. Watson [89] implemented summation of errors across the entire image. Eckert [26] found that for very large images, this predicts that image quality is significantly worse than found in subjective experiments. This problem was avoided by implementing spatial summation of errors over a sliding window rather than the entire image.

The mismatch between the block DCT basis functions and spatial frequency characteristics of the channels in human vision can limit the ability of the DCTune perceptual model to predict certain types of artefacts. In particular, a block DCT based perceptual model will fail to predict ringing artefacts around edges because it integrates energy over the entire block, and thus is not sensitive to errors localised in sub-block regions. Neither will the block DCT based perceptual model directly predict the visibility of blocking artefacts at lower bit rates. This is due to the fact that the block boundaries in the compression scheme and perceptual model directly align, and thus these quantisation errors will be invisible to the perceptual model.

Eckert [26] proposed variants of the DCTune algorithm in which the linear filter bank has been replaced with the lapped orthogonal transform (LOT) and wavelet transform. The LOT has similar problems to block DCT associated with predicting artefacts near edges. In fact, this problem was exacerbated in the LOT because the basis functions for high spatial frequency components spread energy over a 16×16 pixel region. Note that for both the block DCT and LOT based perceptual model, the problem does not lie in the fact that compression with these techniques produce quantisation artefacts (such as ringing around edges), but rather that the perceptual model which uses these basis functions cannot predict the visibility of these particular artefacts.

Safranek and Johnston [73], Johnston and Safranek [38], and Chou [16] implement a perceptual model using subband coding filters. The model incorporates a multiple frequency channel decomposition using equal bandwidth subband filters, weighting of each subband according to the CSF, luminance adaptation, and a contrast masking component referred to as "texture masking". The use of equal bandwidth subband filters means that spatial localisation of high frequencies is not matched to that of the human visual system. Hontsch and Karam [34,35] implement a similar model in the context of a spatially adaptive compression scheme, but also include a component of interband contrast masking.

Bradley [13] produced a model similar to that of Daly's except that he experimented with both over-complete and critically sampled wavelet transforms. The aim of this work was to produce a simplified visual model based on the wavelet transform that could be applied directly to wavelet compression schemes. He found that the main problems with the wavelet transform were the amount of overlap between adjacent frequency channels (leading to aliasing in the subbands) and the shift invariance of the critically sampled wavelet transform. However, the over-complete wavelet transform was found to improve the reliability of predictions.

Barten [9] presented a metric designed to predict observer ratings for factors such as image size and image sharpness (spatial resolution). The metric includes a CSF which depends on image size and the average luminance. Sakrison [75] presented an early, but quite thorough, description of how a perceptual metric can be constructed with multiple frequency channels, luminance adaptation, contrast masking, and error summation. However, the model was neither implemented nor validated.

In summary, there are a number of similarities in all of the perceptual metrics referenced in Section 4.4: All of these metrics incorporate CSF weighting and luminance adaptation, in slightly different, but probably insignificant ways. All of these metrics implement a multiple-channel

frequency decomposition, with fixed or octave bandwidth filters, some with full orientation selectivity and others without orientation selectivity for the 45/135° diagonal. None of these metrics incorporate factors involving suprathreshold observer preferences. The primary differences lie in the different implementations of contrast masking and the different ways in which errors are summed across space and frequency. Not surprisingly, the variability of the implementations of masking and summation of errors reflects the fact that there is no consensus of opinion as to the best psychophysical model for these two perceptual factors.

5. Psychophysical validation

One of the problems associated with validating an image quality metric is that the "gold standard" is a human observer. This means that the accuracy and robustness of a perceptual metric is closely tied to the psychophysical experiments used to validate the metric. Unfortunately, there seem to be as many psychophysical techniques used to validate metrics as there are metrics. Here we describe a number of approaches that have been used to validate quality metrics in the literature and discuss some of the difficulties that may arise.

5.1. Assessing quality in images with suprathreshold compression artefacts

The simplest approach for evaluating the accuracy of a metric is to use a rating scale as suggested in CCIR Recommendation 500-3 [15]. This scale has a range of 1–5, with the adjective descriptions, *bad*, *poor*, *fair*, *good* and *excellent*. In a typical rating experiment an image from a set is displayed to an observer who rates the image using the scale. This continues until all images have been rated a number of times by different observers. These scores can be used to validate an image quality metrics in a number of ways [79]:

1. The raw scores can be correlated with the predictions of the image quality metric to evaluate its accuracy [51].
2. The raw scores can be converted to z-scores and correlated with the image quality metric. z-scores are used to account for biases between individuals, and are obtained by subtracting the mean score for each individual and dividing by the standard deviation of the scores computed over the set of images for an individual [84].
3. Thurstone scaling can be used to create an interval scale, so that the scale represents equal perceptual distances [82,84].
4. Multi-dimensional scaling can be performed to extract the different quality dimensions. This provides the quality dimensions and relative weighting of the dimensions for each observer [4,49].

The limitations of a rating scale assessment are that it will only characterise relatively large differences in image quality and may produce inconsistent results when evaluating an image set which contains different types of artefacts. Both van Dijk et al. [84] and de Ridder and Majoor [22] report acceptable results with rating experiments as long as the type of artefacts in the compressed images are similar. A common difficulty with a rating scale technique is a lack of consistency across laboratories, but Roufs [70] reports good consistency as long as care is taken in the design of the test conditions. Minor modifications to this technique include assigning ratings to sub-image regions rather than a single number for the entire image in order to account for quality variations across the image.

Paired comparison experiments are also used to assess images with suprathreshold compression artefacts [21,84]. A set of images, compressed at different bit rates, are compared to one another. A typical technique is to provide a comparison scale ranging from -3 to 3 for the judgements *much worse*, *worse*, *slightly worse*, *same*, *slightly better*, *better*, *much better*. Alternatively, two point scales can be used, such as identifying the image of higher quality [31], or instructions to judge the "dissimilarity" between images [50]. The observer compares two images from the set and makes a judgement until all images have been compared to one another. The advantage of this technique is that comparisons can be made using images with different types of artefacts and observers are forced to link judgements of quality for the different artefact types

[84]. The data from paired comparison experiments can be used in a number of ways. The simplest approach is to reduce the paired comparison results to a rank ordering of the images in order to correlate the ability of the perceptual model to predict the rank ordering of image quality [31]. Alternatively, multidimensional scaling can be applied to the results to evaluate the number of quality dimensions and weighting of each dimension by observers [4,40,41,49].

5.2. Assessing quality in images with threshold compression artefacts

Just noticeable difference (JND) testing is used to evaluate the ability of a metric to predict the visually lossless point between the compressed and original image. JND testing is particularly appropriate for perceptual metrics which incorporate only threshold visual factors, such as all of the metrics described in Section 4.4. An example of the JND approach is described by Watson et al. [93]. The image is displayed to an observer for a set period of time (such as one second) and the observer decides whether the image has been compressed. After a block of trials, in which the observer has been repetitively presented with both the original and a set of compressed images over a range of bit rates, the JND point is identified as the compression point at which the observer correctly identifies the compressed image 50% of the time (the exact percentage correct for identifying threshold is somewhat arbitrary). The advantage of JND experiments is that they should not be biased significantly by differences in the types of artefacts. Thus, the JND compression point can be used to evaluate a variety of compression techniques. As an example, Eckert [26] compares the compression ratio at the JND threshold for block DCT, LOT and wavelet compression techniques and observers reported no difficulties in performing the experiments.

Several aspects of JND experiments may affect the results. The display time, search strategies, and learning effects can play a major role in determining the JND point. In our experiments, we have found that placing a one second time limit on trials leads the observer to use the following strategy: During initial trials, the observer searches for the most visible artefacts. Prior familiarity with typical compression artefacts speeds this search. Once the location of the most visible artefact(s) have been located, the observer ceases to search, and concentrates on one or two locations for the rest of the trials. In this way, the observer essentially learns the characteristics of these regions as accurately as possible. As the number of trials increases, the JND point, which now depends only on these regions, very gradually decreases, reaching an asymptote after approximately two hundred trials, though we have not assessed exact number of trials needed before the asymptote is reached. All four observers the authors have used in JND experiments almost immediately developed this strategy during testing and all experienced a gradual decrease in the JND point with increased familiarity with the image. As an example of these factors playing a role in JND tests, Fuhrmann et al [31] showed significant reductions in the JND point when observers were given hints about where to search for artefacts in the images. The dependence of initial estimates of the JND point on search strategies is in general agreement with the observations of Sperling and Dosher [78] and the reduction in the JND point with learning is consistent with the results of Swift and Smith [80] and Watson [92].

The importance of prior information regarding image content has also been observed by Good et al. [33], who report significantly different visibility thresholds for two observer groups during JND experiments using compressed mammograms. Basically, experts in image compression could identify compression artefacts in mammograms much more easily than radiologists who were familiar with the image content but not familiar with compression artefacts.

As an example of the application of psychophysical validation of quality metrics, Fuhrmann et al. [31] evaluated a set of mathematically defined quality metrics using a JND experiment to judge the ability of the metrics to predict the visually lossless point and a paired comparison experiment to rank order the images when artefacts were suprathreshold. This combined approach is conceptually pleasing because an ideal quality metric

should not only provide good prediction of the JND point, but also be able to scale with suprathreshold artefacts in a manner which is consistent with paired comparison experiments.

Finally, we return to a point made at the beginning of this review, namely, that image quality varies across the image, yet all of psychophysical experimental techniques force the observer to collapse the local variations of quality into a single judgment. One way of handling this issue is to describe to the observer how the judgements should be made. For example, the observer could be instructed to judge image quality based on the worst artefact, an average of overall quality, only in smooth regions, only in edge regions, etc. When artefacts are suprathreshold, such instructions will change the observer's ratings. Consequently, the instructions given to the observer should reflect the technique used in the quality metric to sum errors across space and assign a quality rating to the image. Another way of handling this issue is to maintain a visible difference map and have the user specify image quality for different regions in the image. This has the advantage of providing a better match to the output of most perceptual metrics which is a 2-D quality map. Fuhrmann et al. [31] essentially used the latter approach and divided images into small sub images for the paired comparison experiments.

5.3. Technical issues associated with subjective experiments

There are a number of technical issues associated with image display and experimental design which should be controlled when performing experiments to validate perceptual metrics. Experiments performed on CRT monitors should have a calibrated gamma function and compensate for the spatial resolution limitations of the monitor [25,61]. The spatial resolution limitations can be handled by interpolating the images by a factor of two or more. Viewing distance, image size, image contrast, ambient illumination, and the average luminance of each image can influence perceptions of image quality and should be carefully controlled in subjective experiments [70].

6. A simple demonstration

In this section, we provide a simple demonstration to illustrate the difficulty experienced by image quality metrics in handling contrast masking. To do this, we have attempted to minimise the effect of other factors, such as luminance adaptation, frequency sensitivity and error summation. We have created two artificial images, one an "edge" image and the other a "texture" image. Both images have a 1/f power spectrum, the same background luminance level, and similar energy content in local regions, i.e., the average variance in an 8×8 pixel region around the edges in the edge image is the same as average variance in 8×8 pixel regions of the texture image. In addition, the image content is homogeneous as the edge image consists of only edges and the texture image consists of only texture. As a consequence, quantisation artefacts are well distributed throughout each image with no large, spatially localised, error peaks. The homogeneity of the errors minimises the dependence of the results on the summation parameter in the Minkowski metric.

The edge and the texture image were both transformed using Antonini/Daubechies 9/7 biorthogonal wavelets [7] and quantised with a uniform quantisation matrix. The synthesis filters were normalised to have unit energy and so the passband gain is approximately two, i.e., coefficient amplitudes scale by a factor of two with increasing levels of the wavelet pyramid. Therefore, uniform quantisation of the coefficients means that high spatial frequencies are quantised more harshly than lower frequencies. We varied the quantisation from two to ten and conducted JND experiments to identify the visually lossless compression point for both images. An estimate of the JND threshold for each observer was obtained from a block of sixty-four 2-alternative forced choice (2AFC) trials, where each block of trials was controlled by the QUEST adaptive staircase algorithm [86]. For each 2AFC trial the observer was presented with the original and a compressed image, and each was displayed for one second and in randomised order. The observer was then asked to select the compressed image and auditory feedback was given when the correct image was selected. The images were pixel

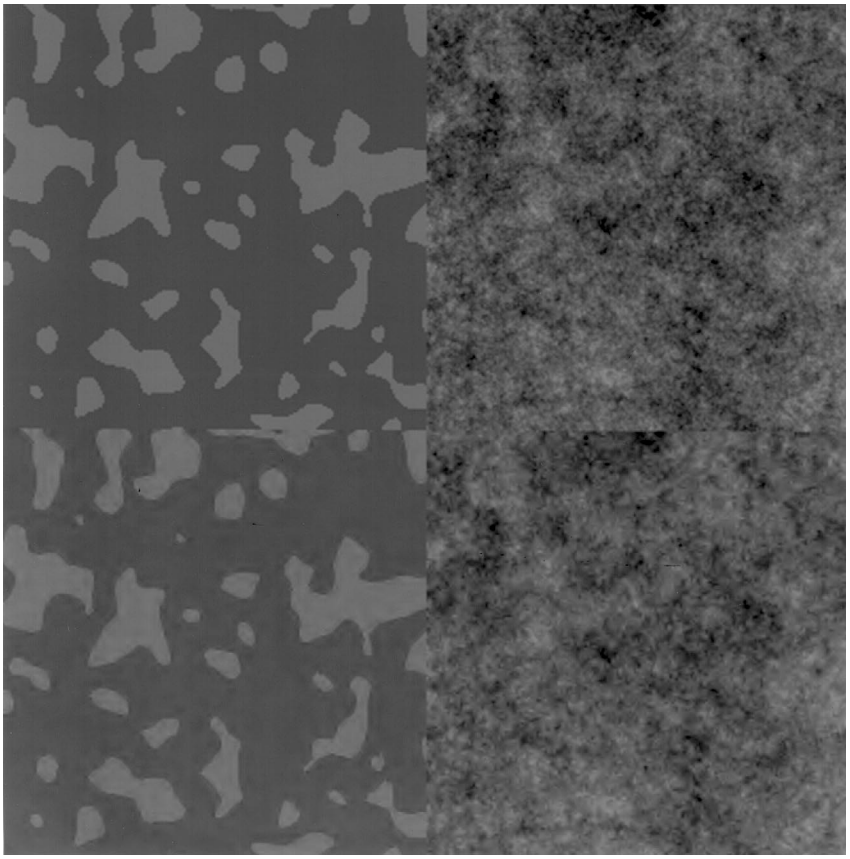

Fig. 7. A small segment of original and quantised edge and texture images. The quantisation factor is 15 in both cases, so the artifacts are significantly suprathreshold. The correct viewing distance is 10 times the width of the entire image.

doubled to avoid the resolution limitations of the CRT monitor and viewing distance was set so that there were 44 pixels/degree of visual angle. The JND point was selected as the 82% correct point from a block of trials and averaged across three observers.

We found that the JND quantisation factor, averaged over the three observers, was two for the edge image and six for the texture image. This means that compression artefacts that were equally visible in both the texture and edge images, i.e., at the JND point, were three times larger in amplitude in the texture image than in the edge image. This significant difference indicates that, near the visibility threshold, three times as much quantisation noise is masked in the texture image than in the edge image.

In an additional experiment, ten observers were presented with the original and compressed versions of the images, as illustrated in Fig. 7. The compressed images had quantisation factors of ten or twenty. For each quantisation factor the observers were asked to identify which compressed image was least similar to its original image. At a quantisation factor of ten, seven out of the ten observers stated that the texture image was least similar to the original (though most stated that the decision was difficult). At a quantisation factor of twenty all ten observers stated that the texture image was clearly more distorted.

Fig. 8 illustrates a number of image quality metrics applied to the texture and edge image as the uniform quantisation matrix is increased from two

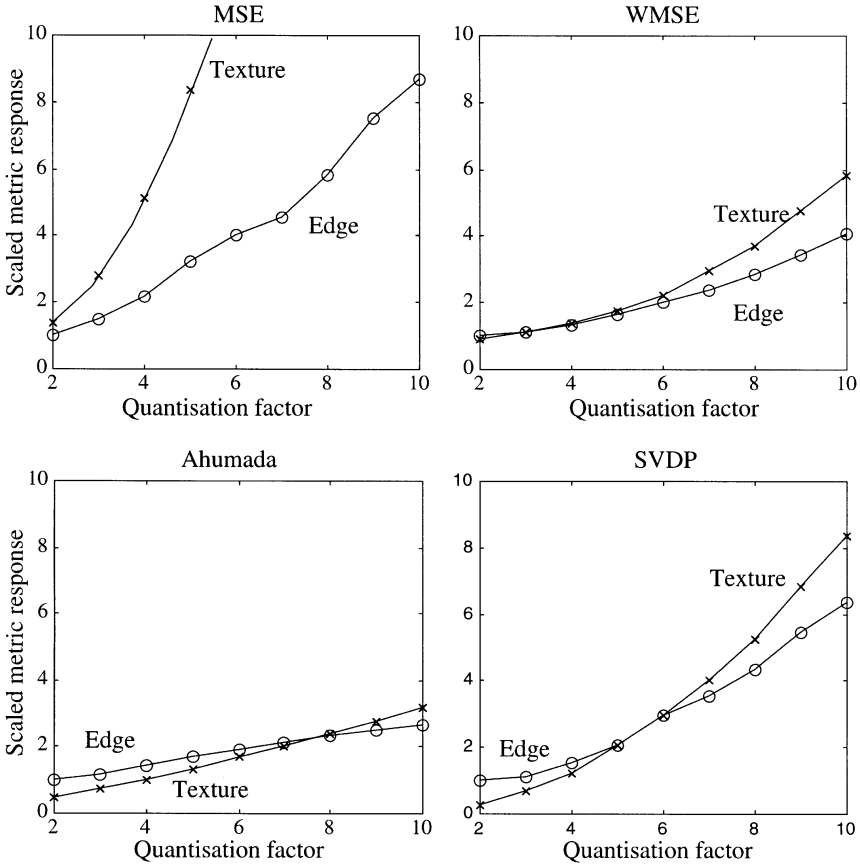

Fig. 8. Various quality metrics which have been used in the literature (higher numbers predict high levels of distortion). The top left curve is the mean squared error of grey levels. The top right curve is filtered mean squared error after applying Daly's luminance masking and CSF filtering. The bottom left curve [5] performs luminance masking, CSF filtering, and contrast masking on a 0.11 degree region. The bottom right curve is essentially Daly's visible difference predictor. The curves on each graph have been scaled by the value for the edge image at $Q = 2$.

to ten. The curves on each graph are normalised by the value of the metric on the edge image at a quantisation factor of two. The image quality metrics shown are:

- Mean square error (MSE).
- Weighted mean square error (WMSE), which incorporates luminance adaptation and frequency sensitivity.
- The Ahumada metric [5] which incorporates luminance adaptation, frequency sensitivity, and a contrast gain control mechanism that operates on a 0.11 degree window (approximately five times the width of the CSF filter) to produce a visible difference map. This metric has been shown to work almost as well as multiple channel models for the detection of signals buried in noise [68]. The visible difference map produced was then reduced to a single number using a Minkowski metric ($\beta = 4$) summed over the entire image.
- The SVDP metric is an implementation of the Daly [17] visible difference predictor using steerable filters [30]. The visible difference map was reduced to a single number, again using a Minkowski metric ($\beta = 4$) summed over the entire image.

A threshold perceptual metric should be able to predict the correct quality ranking near the visibility threshold, i.e., for the JND experiment, whereas a suprathreshold quality metric should be able to correctly predict image rankings at suprathreshold quantisation levels, i.e., image similarity at quantisation factors of ten and twenty. This means that the metric scores in Fig. 8 should predict a larger number (lower quality) for the edge image near the visibility threshold and a larger number for the texture image at quantisation factors of ten and twenty. Fig. 8 shows that the mean squared error (MSE) and the weighted mean squared error (WMSE) metric predict equivalent distortion near the visual threshold and larger amounts of distortion in the texture image for all quantisation levels. Consequently, both provide an incorrect prediction of quality near the visual threshold. Of the four metrics considered, only the two models which contained a contrast masking component, the Ahumada metric [5] and Daly's [17] visual model, correctly predicted rankings at threshold and suprathreshold levels.

Naturally, strong conclusions cannot be drawn from a simple two image demonstration, but it does illustrate that metrics such as MSE and weighted MSE have difficulty predicting visual quality near threshold when contrast masking plays a role. In addition, the demonstration illustrates some of the psychophysical techniques used to evaluate quality metrics near and above the visual threshold.

7. Conclusions

In this paper we have reviewed a number of image quality metrics, discussed how they incorporate visual factors in their design, and described common experimental techniques used to validate the metrics. We can summarise our observations as follows:

- Both simple and complex perceptual quality metrics incorporate luminance adaptation and frequency sensitivity. They are robust characteristics of visual processing, are reasonably well understood, and are easily incorporated into a quality metric.
- The lack of contrast masking in a quality metric is a significant reason for the failure of many quality metrics across a range of image content. There is still no consensus of opinion regarding models for contrast masking, particularly when compression artefacts are suprathreshold. Consequently, one can expect significant variability in the predictions of quality metrics until contrast masking models are developed which demonstrate robust performance across a variety of image content and compression artefacts.
- Summation of errors across frequency and across space is a well accepted component of a quality metric. Models for error summation near the visual threshold are well accepted, but summation at suprathreshold levels has not been well investigated. As a result, variability in the predictions of a quality metric may arise if the observer sums suprathreshold errors differently than is assumed by the quality metric.
- Search strategies and learning play a major role in the perception of threshold artefacts during JND experiments, and observer preferences and expectations play a role when comparing images when artefacts are suprathreshold. Therefore, the experimental technique used to validate a quality metric must be carefully selected. In particular, one must ensure that the type of experiments used to validate the metric match the quality range for which the metric was designed (threshold or suprathreshold artefacts).

We used a simple demonstration to illustrate the importance of including contrast masking into a perceptual quality metric. We showed how a number of simple metrics, such as mean squared error and weighted mean squared error, could not account for the differences in image content (edges and texture) and that metrics which contained contrast gain components provided the best prediction of the relative quality of the two images at both threshold and suprathreshold levels.

References

[1] A.J. Ahumada, H.A. Peterson, Luminance-Model based DCT quantization for color image compression, Proc. SPIE 1666 (1992) 365–374.

[2] A.J. Ahumada, Computational image quality metrics: A review, SID Digest of Technical Papers 24 (1993) 305–308.
[3] A.J. Ahumada, H.A. Peterson, A visual detection model for DCT coefficient quantization, in: AIAA Computing in Aerospace 9: A Collection of Technical Papers, San Diego, California, 19–21 October 1993, pp. 314–317.
[4] A.J. Ahumada, C.H. Null, Image quality: A multidimensional problem, in: A.B. Watson (Ed.), Digital Images and Human Vision, MIT Press, Cambridge, MA, 1993, pp. 141–148.
[5] A.J. Ahumada, Simplified vision models for image quality assessment, Society for Information Display, 1996.
[6] M.G. Albanesi, Wavelets and human visual perception in image compression, in: Proc. ICPR, IEEE, 1996, pp. 859–863.
[7] M. Antonini, M. Barlaud, P. Mathieu, I. Daubechies, Image coding using the wavelet transform, IEEE Trans. Image Processing 1 (1992) 205–220.
[8] N. Avadhanam, V. Algazi, Prediction and measurement of high quality in still-image coding, Proc. SPIE 2663 (1996) 100–109.
[9] P. Barten, Evaluation of subjective image quality with the square-root integral method, J. Opt. Soc. Amer. A 7 (10) (1990) 2024–2031.
[10] S. Bertoluzza, M.G. Albanesi, On the coupling of human visual system model and wavelet transform for image compression, Proc. SPIE 2303 (1994) 389–397.
[11] W. Bishtawi, W.E. Lynch, Objective measurement of image impairments blocking, blurring, and spatial edge noise, in: Canadian Conference on Electrical and Computer Engineering, IEEE, Vol. 1, 1995, pp. 156–159.
[12] H. Blume, S. Daly, E. Muka, Presentation of medical images on CRT displays: A renewed proposal for a display function standard, Proc. SPIE 1897 (1993) 213–231.
[13] A. Bradley, A wavelet visible difference predictor, in: Proc. Digital Images: Techniques and Applications (DICTA'97), Auckland, New Zealand, 1997, pp. 77–82.
[14] S.J. Briggs, Photometric technique for deriving a "Best Gamma" for Displays, Optical Engineering 20 (4) (1981) 651–657.
[15] CCIR, Method for the Subjective Assessment of the Quality of Television Pictures, Recommendation 500-3, in: Recommendations and Reports of the CCIR, International Telecommunication Union, Geneva, 1986.
[16] C.H. Chou, Y.C. Li, A perceptually tuned subband image coder based on the measure of just-noticeable distortion profile, IEEE Trans. Circuits and Systems for Video Technology 5 (6) (1995) 467–476.
[17] 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, Cambridge, MA, 1993, pp. 179–206.
[18] S. Daly, Quantitative performance assessment of an algorithm for the determination of image fidelity, SID Digest of Technical Papers, 1993, pp. 317–320.
[19] S. Daly, A visual model for optimizing the design of image processing algorithms, in: Proc. ICIP-94, IEEE Computer Society Press, 13–16 November 1994, pp. 16–20.
[20] S. Daly, Method and apparatus for determining visually perceptible differences between images, U.S. Patent: 5394483, February 1995.
[21] H.A. David, The Method of Paired Comparisons, Charles Griffin and Company Limited, 1969.
[22] H. de Ridder, G.M. Majoor, Numerical category scaling: An efficient method for assessing digital image coding impairments, Proc. SPIE 1249 (1990) 65–77.
[23] R. de Vore, B. Jawerth, B. Lucier, Image compression through wavelet transform coding, IEEE Trans. Inform Theory 38 (2) (1992) 719–746.
[24] M. Dzmura, B. Singer, Spatial Pooling of Contrast Gain Control, J. Opt. Soc. Amer. A 13 (11) (1996) 2135–2140.
[25] M.P. Eckert, D.P. Chakraborty, Video display quality control measurements for PACS, in: Proc. SPIE 2431 (1995) 328–340.
[26] M.P. Eckert, Lossy compression using wavelets, block DCT, and lapped orthogonal transforms optimized with a perceptual model, Proc. SPIE 3031 (1997) 339–351.
[27] A.M. Eskicioglu, P.S. Fisher, A survey of quality measures for grey scale image compression, in: Proc. NASA Space Earth Science Data Compression Workshop, 1993, pp. 49–61.
[28] J.M. Foley, Human luminance pattern mechanisms: Masking experiments require a new model, J. Opt. Soc. Amer. A 11 (6) (1994) 1710–1719.
[29] J.M. Foley, G.M. Boynton, A new model of human luminance pattern vision mechansims: Analysis of the effects of pattern orientation, spatial phase and temporal frequency, Computational Vision Based on Neurobiology, Proc. SPIE 2054 (1993) 32–42.
[30] W.T. Freeman, E.H. Adelson, The design and use of steerable filters, IEEE Trans. on Pattern Analysis and Machine Intelligence 13 (9) (1991) 891–906.
[31] D.R. Fuhrmann, J.A. Baro, J.R. Cox, Experimental evaluation of psychophysical distortion metrics for JPEG-encoded Images, J. Electronic Imaging 4 (4) (1995) 397–406.
[32] B. Girod, What's wrong with mean-square error, in: A.B. Watson (Ed.), Digital Images and Human Vision, MIT Press, Cambridge, MA, 1993, pp. 207–220.
[33] W.F. Good, G.S. Maitz, D. Gur, Joint photographic experts group (JPEG) compatible data compression of mammograms, J. Digital Imaging 7 (3) (1994) 123–132.
[34] I. Hontsch, L.J. Karam, APIC: Adaptive perceptual image coding based on subband decomposition with locally adaptive perceptual weighting, in: Proc. Internat. Conf. Image Processing, IEEE, 1997, pp. 37–40.
[35] I. Hontsch, L.J. Karam, Locally adaptive perceptual quantization without side information for compression of visual data, Globecom '97, Global Telecommunications Conference, IEEE, 1997, pp. 1042–1046.
[36] D.C. Hood, M.A. Finkelstein, Sensitivity to light, in: Handbook of Perception and Human Performance, Boff, Kaufman, Thomas (Eds.),Vol. 1, Chapter 5, Wiley, New York, 1986.

[37] N. Jayant, J. Johnston, R. Safranek, Signal compression based on models of human perception, Proc. IEEE 81 (10) (1993) 1385–1421.
[38] J.D. Johnston, R.J. Safranek, Perceptually adaptive image coding system, U.S. Patent 5,517,581, 1996.
[39] S.A. Karunasekera, N.G. Kingsbury, A distortion measure for blocking artifacts in images based on human vision sensitivity, IEEE Trans. on Image Processing 4 (6) (1995) 713–724.
[40] V. Kayargadde, J.B. Martens, Perceptual characterization of images degraded by blur and noise: experiments, J. Opt. Soc. Amer. A 13 (1996) 1166–1177.
[41] V. Kayargadde, J.B. Martens, Perceptual characterization of images degraded by blur and noise: Model, J. Opt. Soc. Amer. A 13 (1996) 1178–1188.
[42] Y. Kim, I. Choi, I. Lee, T. Yun, K.T. Park, Wavelet transform image compression using human visual characteristics and a tree structure with a height attribute, Optical Engineering 35 (1) (1996) 204–212.
[43] F. Kingdom, P. Whittle, Contrast discrimination at high contrasts reveals the influence of local light adaptation on contrast processing, Vision Research 36 (6) (1996) 817–829.
[44] G.E. Legge, J.M. Foley, Contrast masking in human vision, J. Opt. Soc. Amer. A 70 (12) (1980) 1458–1471.
[45] G.E. Legge, D. Kersten, A.E. Burgess, Contrast discrimination in noise, J. Opt. Soc. Amer. A 4 (2) (1987) 391–404.
[46] J.O. Limb, Distortion criteria of the human viewer, IEEE Trans. Systems, Man, and Cybernetics 9 (12) (1979) 778–793.
[47] J. Lubin, The use of psychophysical data and models in the analysis of display system performance, in: A.B. Watson (Ed.), Digital Images and Human Vision, MIT Press, Cambridge, MA, 1993, pp. 163–178.
[48] J. Mannos, D. Sakrison, The effects of visual fidelity criterion on the encoding of images, IEEE Trans. Inform. Theory 20 (4) (1974) 525–536.
[49] J.B. Martens, V. Kayargadde, Image quality prediction in a multidimensional perceptual space, Internat. Conf. Image Processing, Vol. 1, Los Alamos, CA, 1996, pp. 877–880.
[50] J.-B. Martens, L. Meesters, Image dissimilarity, Signal Processing 70 (1998) 155–176.
[51] M. Miyahara, K. Kotani, V.R. Algazi, Objective picture quality scale (PQS) for image coding, submitted to IEEE Transaction on Communications, 1996.
[52] H. Mostafavi, D.J. Sakrison, Structure and properties of a single channel in the human visual system, Vision Research 16 (1977) 957–968.
[53] N.B. Nill, A visual model weighted cosine transform for image compression and quality assessment, IEEE Trans. Commun. 33 (6) (1985) 551–557.
[54] N.B. Nill, B.H. Bouzas, Objective image quality measure derived from digital image power spectra, Optical Engineering 31 (4) (1992) 813–825.
[55] T.P. O'Rourke, R.L. Stevenson, Human visual system based wavelet decomposition for image compression, Journal of Visual Communication and Image Representation 6 (2) (1995) 109–121.
[56] L.A. Olzak, J.P. Thomas, Seeing spatial patterns in: Boff, Kaufman, Thomas (Eds.), Handbook of Perception and Human Performance, Chapter 7, Wiley, New York, 1986.
[57] T.N. Pappas, T.A. Michel, R.O. Hinds, Supra-threshold perceptual image coding, in: Proc. Internat. Conf. Image Processing, 1996, pp. 237–240.
[58] E. Peli, Contrast in complex images, J. Opt. Soc. Amer. A 7 (10) (1990) 2032–2040.
[59] E. Peli, L.E. Arend, G.M. Young, R.B. Goldstein, Contrast Sensivity to patch stimuli: Effects of spatial bandwdith and temporal presentation, Spatial Vision 7 (1) (1993) 1–14.
[60] D.G. Pelli, The quantum efficiency of vision, in: C. Blakemore (Ed.), Vision: Coding and Efficiency, Cambridge University Press, Cambridge, 1990, pp. 3–24.
[61] D.G. Pelli, Pixel independence: Measuring spatial interactions on a CRT display, Spatial Vision 10 (4) (1997) 443–446.
[62] H.A. Peterson, A.J. Ahumada, A.B. Watson, An improved detection model for DCT coefficient quantization, Proc. SPIE 1913 (1993) 191–201.
[63] H.A. Peterson, A.J. Ahumada, A.B. Watson, The visibility of DCT quantization noise, Soc. Inf. Display Digest of Technical Papers 24 (1993) 942–945.
[64] H.A. Peterson, A.J. Ahumada, A.B. Watson, The visibility of DCT quantization noise: Spatial frequency summation, SID International Symposium Digest of Technical Papers 25 (1994) 704–707.
[65] R.F. Quick, A vector magnitude model of contrast detection, Kybernetik 16 (1974) 65–67.
[66] J.G. Robson, N. Graham, Probability summation and regional variation in contrast sensitivity across the visual field, Vision Research 21 (1981) 409–418.
[67] J.G. Rogers, W.L. Carel, Report HAC Ref. No. C 6619. Hughes Aircraft Company, Culver City, CA, (Office of Naval Research Contract Number: N00014-72-C-0451,NR213-107), December 1973.
[68] A.M. Rohaly, A.J. Ahumada Jr., A.B. Watson, Object detection in natural backgrounds predicted by discrimination performance and models, Vision Research 37 (1997) 3225–3235.
[69] R. Rosenholtz, A.B. Watson, Perceptual adaptive JPEG coding, in: Proc. IEEE Internat. Conf. Image Processing, Lausanne, Switzerland, Vol. 1, 1996, pp. 901–904.
[70] J.A.J. Roufs, Perceptual image quality: concept and measurement, Philips J. Res. 47 (1992) 35–62.
[71] A. Said, W.A. Pearlman, A new, fast, and efficient image codec based on set partitioning in heirarchical trees, IEEE Trans. Circ. Sys. Video Technology 6 (1996) 243–250.
[72] R. Safranek, A comparison of the coding efficiency of perceptual models, Proc. SPIE 2411 (1995) 83–91.
[73] R.J. Safranek, J.D. Johnston, A perceptually tuned subband image coder with image dependent quantization and post-quantization data compression, in: Proc. ICASSP 3 1989, pp. 1945–1948.
[74] J.A. Saghri, P.S. Cheatham, A. Habibi, Image quality measure based on a human visual system model, Optical Engineering 28 (7) (1989) 813–818.

[75] D.J. Sakrison, On the role of the observer and a distortion measure in image transmission, IEEE Trans. Commun. 25 (11) (1977) 1251–1267.
[76] E. Simoncelli, W.T. Freeman, E.H. Adelson, D.J. Heeger, Shiftable multiscale transforms, IEEE Trans. Inform. Theory 38 (2) (1992) 587–607.
[77] J.A. Solomon, A.B. Watson, A. Ahumada, Visibility of DCT basis functions: Effects of contrast masking, in: Proceedings of the Data Compression Conference, IEEE Computer Society Press, 1994, pp. 361–371.
[78] G. Sperling, B.A. Dosher, Strategy and optimization in human information processing, in: Boff, Kaufman, Thomas (Eds.), Handbook of Perception and Human Performance, Vol. 1, Chapter 2, Wiley, New York, 1986.
[79] C.S. Stein, A.B. Watson, L.E. Hitchner, Psychophysical rating of image compression techniques, Proc. SPIE 1077 (1989) 198–208.
[80] D.J. Swift, R.A. Smith, Spatial frequency masking and Weber's law, Vision Research 23 (1983) 495–506.
[81] P.C. Teo, D. Heeger, Perceptual image distortion, Proc. SPIE 2179 (1994) 127–139.
[82] W.S. Torgerson, Theory and Methods of Scaling, Wiley, New York, 1958.
[83] T. Tran, R. Safranek, A locally adaptive perceptual masking threshold model for image coding, in: Proc. ICASSP, 1996.
[84] A.M. van Digk, J.B. Martens, A.B. Watson, Quality assessment of coded images using numerical category scaling, advanced image and video communications and storage technologies, Amsterdam, Proc. SPIE 2451 (1995) 90–101.
[85] A.B. Watson, Summation of grating patches indicates many types of detectors at one retinal location, Vision Research 22 (1982) 17–25.
[86] A.B. Watson, D.G. Pelli, QUEST: A Bayesian adaptive psychometric method, Percept. Psychophys. 33 (1983) 113–120.
[87] A.B. Watson, The cortex transform: Rapid computation of simulated neural images, Computer Vision, Graphics, and Image Processing 39 (1987) 311–327.
[88] A.B. Watson, Estimation of local spatial scale, J. Opt. Soc. Amer. A4, 1987, pp. 1579–1582.
[89] A.B. Watson, DCTune: A technique for visual optimization of DCT quantization matrices for individual images, SID Digest of Technical Papers XXIV (1993) 946–949.
[90] A.B. Watson, Image Data Compression Having Minimum Perceptual Error, U.S. Patent: 5,426,512, 1995.
[91] A.B. Watson, J.A. Solomon, A model of visual contrast gain control and pattern masking, J. Opt. Soc. Amer. A 14 (1997) 2379–2391.
[92] A.B. Watson, R. Borthwick, M. Taylor, Image quality and entropy masking, Proc. SPIE 3016 (1997) 358–371.
[93] A.B. Watson, M. Taylor, R. Borthwick, DCTune perceptual optimization of compressed dental X-Rays, Proc. SPIE 3031 (1997) 358–371.
[94] A.B. Watson, G.Y. Yang, J.A. Solomon, J. Villasenor, Visibility of wavelet quantisation noise, IEEE Trans. Image Processing 6 (8) (1997) 1164–1175.
[95] S.J.P. Westen, R.L. Lagendijk, J. Biemond, Perceptual image quality based on a multiple channel HVS model, in: Proc. ICASSP, 1995, pp. 2351–2354.
[96] C. Zetzsche, G. Hauske, Multiple channel model prediction of subjective image quality, Proc. SPIE 1077 (1989) 209–215.