

# Visual Fidelity and Perceived Quality: Towards Comprehensive Metrics

Stefan Winkler

Signal Processing Laboratory  
Swiss Federal Institute of Technology  
1015 Lausanne, Switzerland  
<http://ltswww.epfl.ch/~winkler/>  
Stefan.Winkler@epfl.ch

## ABSTRACT

Traditional visual quality metrics measure *fidelity* instead of *quality*, even though fidelity, i.e. the accuracy of the reproduction of the original on the display, is just one of the many factors determining the overall perceived quality. In this paper, the addition of image appeal attributes is investigated in order to bridge this gap. Sharpness and colorfulness are identified among these attributes and are quantified by means of an isotropic local contrast measure and the distribution of chroma, respectively. The benefits of using these attributes are demonstrated with the help of data from subjective experiments.

**Keywords:** Image appeal, vision modeling, video quality assessment, perceptual distortion metric

## 1. INTRODUCTION

The widespread use of digital imaging systems on the consumer market has led to a rising demand for pertinent quality assessment tools on a perceptual level. Consequently, considerable effort has been put into the development of visual quality metrics in recent years, many of which rely on models of the human visual system. An overview of current modeling approaches was presented elsewhere by the author.<sup>23</sup> However, the design of reliable visual quality metrics is complicated by our limited knowledge of the human visual system.

An important shortcoming of most existing metrics is that they measure image *fidelity* instead of perceived *quality*. This approach has its limits with respect to prediction accuracy, even if sophisticated models of the human visual system are used.<sup>21</sup> Further fine-tuning of such metrics or their components for specific applications can improve the prediction performance only slightly.<sup>26</sup> Human observers, on the other hand, seem to require no such “tuning”, yet are able to give more reliable quality ratings.

One of the reasons for this is that the accuracy of the reproduction of the original on the display is only one of the many factors determining visual quality,<sup>18</sup> even if the characteristics of the human visual system are considered. However, high fidelity does not necessarily imply high quality. For example, sharp images with high contrast are usually more appealing to the average viewer. Likewise, subjects prefer slightly more colorful and saturated images despite realizing that they look somewhat unnatural.<sup>4,29</sup> These phenomena are well understood and exploited by professional photographers.<sup>1,15</sup>

In an attempt to overcome the limitations that seem to have been reached by fidelity metrics, we therefore focus on more subjective attributes of image quality, which we refer to as *image appeal* for better distinction. Two of these attributes, namely sharpness and colorfulness, are investigated in this paper.

The paper is structured as follows: Section 2 introduces and quantifies the image appeal attributes of sharpness and colorfulness and describes their integration with a perceptual distortion metric. Section 3 discusses the test sequences and subjective experiments that are used to analyze the benefits of these attributes. The results of this analysis are presented in Section 4.

## 2. MEASURING IMAGE APPEAL

### 2.1. Background

In a study of image appeal in consumer photography,<sup>18</sup> a list of positive and negative influences in the ranking of pictures was compiled based on experiments with human observers. The most important attributes for image selection are related to scene composition and location as well as the people in the picture and their expressions. However, due to the high semantic level of these attributes, it is an extremely difficult and delicate task to take them into account with a general metric for generic scenes.

Fortunately, a number of attributes that greatly influence the subjects' ranking decisions can be measured physically. In particular, colorful, well-lit, sharp pictures with high contrasts are considered attractive, whereas low-quality, dark and blurry pictures with low contrasts are often rejected.<sup>18</sup> The depth of field, i.e. the separation between subject and background, and the range of colors and shades have also been mentioned as contributing factors.<sup>3</sup> The importance of high contrast and sharpness as well as colorfulness and saturation for good pictures has been confirmed by studies on naturalness<sup>4,5,29</sup> and has also been emphasized by professional photographers.<sup>1,13,15</sup>

Based on the above-mentioned studies, *sharpness* and *colorfulness* are among the subjective attributes with the most significant influence on perceived quality. In order to work with these attributes, it is necessary to define them as measurable quantities.

### 2.2. Sharpness

For the computation of sharpness, we propose the use of a local contrast measure. The reasoning is that sharp images exhibit high contrasts, whereas blurring leads to a decrease in contrast. We employ the isotropic local contrast measure presented elsewhere by the author,<sup>27</sup> which is based on the combination of analytic oriented filter responses. Because of its design properties, it is a natural measure of isotropic contrast in complex images. It is briefly reviewed here.

The contrast measure is based on a class of non-separable filters that generalize the properties of analytic functions in 2-D. These filters are actually directional wavelets as defined by Antoine et al.,<sup>2</sup> which are square-integrable functions whose Fourier transform is strictly supported in a convex cone with the apex at the origin. It can be shown that these functions admit a holomorphic continuation in the domain  $\mathbb{R}^2 + jV$ , where  $V$  is the cone defining the support of the function. This is a genuine generalization of the Paley-Wiener theorem for analytic functions in one dimension. Furthermore, if we require that these filters have a flat response to sinusoidal stimuli, it suffices to impose that the opening of the cone  $V$  be strictly smaller than  $\pi$ . This means that at least three such filters are required to cover all possible orientations uniformly, but otherwise any number of filters is possible.

Working in polar coordinates  $(r, \varphi)$  in the Fourier domain, assume  $K$  directional wavelets  $\hat{\Psi}(r, \varphi)$  satisfying the above requirements and

$$\sum_{k=0}^{K-1} |\hat{\Psi}(r, \varphi - 2\pi k/K)|^2 = |\hat{\psi}(r)|^2, \quad (1)$$

where  $\hat{\psi}(r)$  is the Fourier transform of an isotropic dyadic wavelet, i.e.

$$\sum_{j=-\infty}^{+\infty} |\hat{\psi}(2^j r)|^2 = 1 \quad \text{and} \quad \sum_{j=-J}^{+\infty} |\hat{\psi}(2^j r)|^2 = |\hat{\phi}(2^J r)|^2,$$

where  $\phi$  is the associated 2-D scaling function.<sup>14</sup>

Now it is possible to construct an isotropic contrast measure  $C_j^I$  as the square root of the energy sum of these oriented filter responses, normalized by a low-pass band:<sup>27</sup>

$$C_j^I(x, y) = \frac{\sqrt{2 \sum_k |\Psi_{jk} * I(x, y)|^2}}{\phi_j * I(x, y)}, \quad (2)$$

where  $I$  is the input image, and  $\Psi_{jk}$  denotes the wavelet dilated by  $2^{-j}$  and rotated by  $2\pi k/K$ . If the directional wavelet  $\Psi$  is in  $L^1(\mathbb{R}^2) \cap L^2(\mathbb{R}^2)$ , the convolution in the numerator of Eq. (2) is again a square-integrable function,

and Eq. (1) shows that its  $L^2$ -norm is exactly what would have been obtained using the isotropic wavelet  $\psi$ . As can be seen in Figure 1(b),  $C_j^I$  is an orientation- and phase-independent quantity. Furthermore, being defined by means of analytic filters, it is equivalent to Michelson contrast  $C^M$  for sinusoidal gratings (i.e.  $C_j^I(x, y) \equiv C^M$  in this case).

The computation of a robust isotropic contrast measure can be accomplished with a translation-invariant multi-resolution representation based on 2-D analytic filters. This can be achieved by designing a special Dyadic Wavelet Transform (DWT) using 2-D non-separable frames. The very weak design constraints of these frames permit the use of analytic wavelets, for which condition (1) can easily be fulfilled.<sup>20</sup>

Since this construction mainly works in the Fourier domain, it is very easy to add directional sensitivity by multiplying all Fourier transforms with a suitable angular window:

$$\hat{\Psi}(r, \varphi) = \hat{\psi}(r) \cdot \hat{\eta}(\varphi). \quad (3)$$

For this purpose, we introduce an infinitely differentiable, compactly supported function  $\hat{\eta}(\varphi)$  such that

$$\sum_{k=0}^{K-1} |\hat{\eta}(\varphi - 2\pi k/K)|^2 = 1 \quad \forall \varphi \in [0, 2\pi] \quad (4)$$

in order to satisfy condition (1).

This allows us to build oriented pyramids using a very wide class of dyadic wavelet decompositions. The properties of the pertinent filters can then be tailored for specific applications. The wavelet used here is the *Log* wavelet or Mexican hat wavelet, i.e. the Laplacian of a Gaussian. Its frequency response is given by  $\hat{\psi}(r) = r^2 e^{-r^2/2}$ . For the directional separation of this isotropic wavelet, it is shaped in angular direction in the frequency domain according to Eq. (3). The shaping function  $\hat{\eta}(\varphi)$  used here is based on a combination of normalized Schwarz functions<sup>8</sup> that satisfies Eq. (4).

The number of filter orientations  $K$  is one parameter. The minimum number required by the analytic filter constraints, i.e. an angular support smaller than  $\pi$ , is three orientations. The human visual system emphasizes horizontal and vertical directions, so four orientations should be used as a practical minimum. To give additional weight to diagonal structures, eight orientations may be preferred. Although using even more filters might result in a better analysis of the local neighborhood, our experiments indicate that there is no apparent improvement when using more than eight orientations, and the additional computational load outweighs potential benefits.

The other parameter is the center frequency of the filters or the level of the pyramidal decomposition. The lowest level is chosen here, because it contains the high-frequency information, which intuitively appears most suitable for the representation of sharpness. An example of the resulting isotropic local contrast is shown in Figure 1(b). The close connection between contrast and the sharp rendition of image features is evident.

To reduce the contrast values at every pixel of a sequence to a single number, pooling is carried out by means of an  $L^p$ -norm. Several different exponents were tried, but best results were achieved with  $p = 1$ , i.e. plain averaging. Therefore, the sharpness rating of a sequence is defined as the mean isotropic local contrast over the entire sequence:

$$R_{\text{sharp}} = \mu_{C_0^I}. \quad (5)$$

### 2.3. Colorfulness

Colorfulness depends on two factors:<sup>5</sup> The first factor is the average distance of image colors from a neutral gray, which may be modeled as the average chroma. The second factor is the distance between individual colors in the image, which may be modeled as the spread of the distribution of chroma values. If lightness differences between images are neglected, chroma may be replaced by saturation. Conceptually, both saturation and chroma describe the purity of colors. *Saturation* is the colorfulness of an area judged in relation to its own brightness, and *chroma* is the colorfulness of an area judged in relation to the brightness of a similarly illuminated white area.<sup>10</sup>

CIE  $L^*u^*v^*$  color space<sup>28</sup> permits the computation of both measures. Using CIE 1931  $XYZ$  tristimulus values, the lightness component  $L^*$  is defined as:

$$L^* = \begin{cases} 116(Y/Y_0)^{1/3} - 16 & \text{if } Y/Y_0 > 0.008856, \\ 903.292 Y/Y_0 & \text{otherwise.} \end{cases}$$

The 0-subscript refers to the corresponding unit for the reference white being used. The chromaticity coordinates are computed as follows:

$$\begin{aligned} u^* &= 13L^*(u' - u'_0), & u' &= \frac{4X}{X+15Y+3Z}, \\ v^* &= 13L^*(v' - v'_0), & v' &= \frac{9Y}{X+15Y+3Z}. \end{aligned}$$

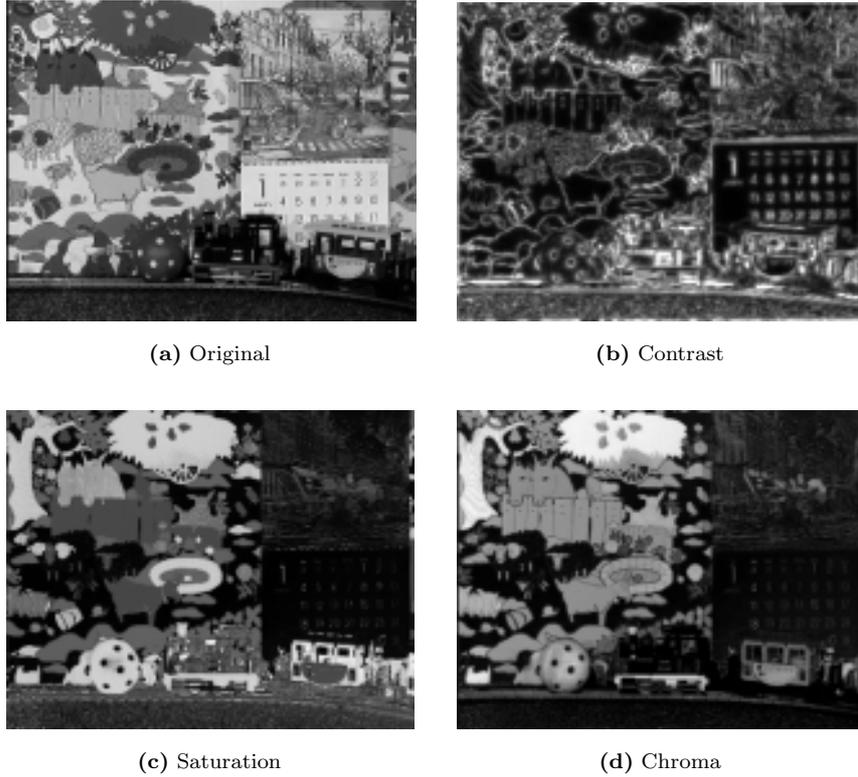
Saturation is then defined as:

$$S_{uv} = 13\sqrt{(u' - u'_0)^2 + (v' - v'_0)^2}, \quad (6)$$

and chroma is defined as:

$$C_{uv}^* = \sqrt{u^{*2} + v^{*2}} = S_{uv}L^*. \quad (7)$$

These quantities are shown for a sample frame in Figure 1(c,d).



**Figure 1:** Luminance contrast  $C_0^I$ , saturation  $S_{uv}$  and chroma  $C_{uv}^*$  for a frame of the mobile scene.

Several other color spaces with a saturation component exist. Examples are *HSI* (hue, saturation, intensity),<sup>9</sup> *HSV* (hue, saturation, value) and *HLS* (hue, lightness, saturation).<sup>6</sup> Their saturation components are very similar and easy to compute. Chroma could also be defined as the product of saturation and lightness as in Eq. (7). However, these color spaces suffer from the fact that they are not perceptually uniform, and that they exhibit a singularity for black. Their saturation components were also used as a measure of colorfulness in our experiments, but the results obtained were generally better with saturation and chroma based on CIE  $L^*u^*v^*$  color space from Eqs. (6) and (7).

We have found that the best overall colorfulness ratings are obtained using the distribution of chroma values. This significantly reduces the number of outliers. According to the dependence of colorfulness on the chroma distribution parameters discussed above, the colorfulness rating of a sequence is thus defined as the sum of mean and standard deviation of chroma values over the entire sequence:<sup>29</sup>

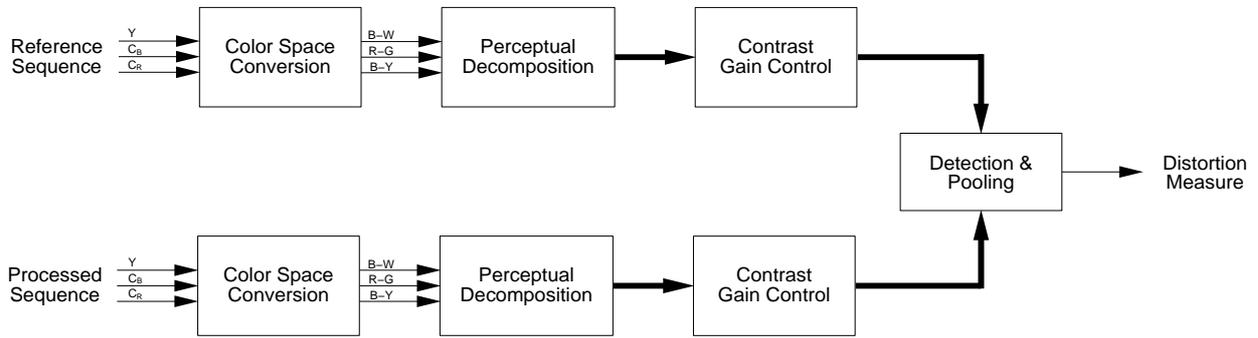
$$R_{\text{color}} = \mu_{C^*} + \sigma_{C^*}. \quad (8)$$

## 2.4. Integration

The underlying premise for using sharpness and colorfulness ratings as additional quality indicators is that a reduction of sharpness or colorfulness from the reference to the distorted sequence corresponds to a decrease in perceived quality. A great advantage of these image appeal attributes is that they can be computed on the reference and the distorted sequences independently. This means that it is not necessary to have the entire reference sequence available at the testing site, but only its sharpness and colorfulness ratings, which can easily be transmitted together with the video data. This paves the way for reduced-reference quality metrics.

However, the image appeal attributes defined above were not designed as stand-alone measures of quality; both have to be used in combination with a visual fidelity metric. In other words, the differences  $\Delta_{\text{sharp}} = R_{\text{sharp}} - \tilde{R}_{\text{sharp}}$  and  $\Delta_{\text{color}} = R_{\text{color}} - \tilde{R}_{\text{color}}$  may be combined with a vision model-based distortion for potentially more accurate predictions of overall visual quality. The benefits of such a combination will be investigated below.

The perceptual distortion metric (PDM) developed by the author<sup>24,25</sup> is used for measuring fidelity or perceptual distortion  $\Delta_{\text{PDM}}$  in this paper. Its structure is briefly reviewed here. The block diagram of the PDM and the underlying vision model is shown in Figure 2.



**Figure 2:** Block diagram of the perceptual distortion metric.

The input is first converted to an opponent color space with black-white (B-W), red-green (R-G), and blue-yellow (B-Y) difference signals, which was designed to separate the effects of color perception and pattern sensitivity.<sup>16,17</sup> Each of the resulting three components is subjected to a spatio-temporal perceptual decomposition, yielding a number of perceptual channels. This decomposition is performed first in the temporal and then in the spatial domain.

The temporal mechanisms in the human visual system are modeled with one low-pass and one band-pass filter.<sup>7</sup> The low-pass filter is applied to all three color channels, whereas the band-pass filter is applied only to the luminance channel. The decomposition in the spatial domain is carried out by means of the steerable pyramid transform,<sup>19</sup> which decomposes an image into a number of spatial frequency and orientation bands.\* In the present implementation, the basis filters have octave bandwidth and octave spacing; five subband levels with four orientation bands each plus one low-pass band are computed. After the temporal and spatial decomposition, each channel is weighted in such a way that the channels approximate the spatio-temporal contrast sensitivity functions of the human visual system.

The contrast gain control stage in the PDM implements pattern sensitivity and contrast masking. Contrast gain control was inspired by analyses of the responses of single neurons in the visual cortex, where this mechanism keeps neural responses within the permissible dynamic range while at the same time retaining global pattern information. It can be modeled by an excitatory nonlinearity that is inhibited divisively by a pool of responses from other neurons.<sup>22</sup> Masking occurs through the inhibitory effect of the normalizing pool. In the PDM, we rely on a generalized contrast gain control model that facilitates the integration of many kinds of channel interactions.<sup>24</sup>

Finally, all sensor differences are combined into a distortion measure according to rules of probability or vector summation, also known as pooling. In principle, any subset of dimensions can be used for this summation, depending on what kind of result is desired.

\* Source code and filter kernels for the steerable pyramid are available at <http://www.cis.upenn.edu/~eero/steerpyr.html>.

### 3. EXPERIMENTS

#### 3.1. VQEG Data

The Video Quality Experts Group (VQEG)\* collected subjective ratings for a large set of test sequences and evaluated the performance of different video quality metrics with respect to these sequences.<sup>21</sup> The sequences and subjective ratings from these experiments are used in the first part of the analysis of the proposed image appeal attributes in Section 4.1. The VQEG test conditions comprise mainly production- and distribution-class MPEG-2 encoded sequences with different profiles, levels and other parameter variations, including encoder concatenation, conversions between analog and digital video, and transmission errors. 20 scenes were encoded for 16 test conditions each. Before the sequences were shown to subjective viewers or assessed by the metrics, normalization with respect to temporal and spatial misalignments as well as chroma and luma gains was carried out.<sup>21</sup>

The Double Stimulus Continuous Quality Scale (DSCQS) method from ITU-R Rec. 500<sup>11</sup> was used in the subjective experiments. In this test, viewers are shown multiple sequence pairs consisting of a “reference” and a “test” sequence, which are rather short (8 seconds in this case). They are presented twice in alternating fashion, with the order of the two chosen randomly for each trial. Subjects are not informed which is the reference and which is the test sequence. They rate each of the two separately on a continuous quality scale ranging from “bad” to “excellent”. The difference in rating for each pair is calculated from an equivalent numerical scale from 0 to 100. The mean subjective rating differences between reference and distorted sequences, also known as differential mean opinion scores (DMOS), are used in the analyses.

The subjective experiments were carried out in eight different laboratories with a total of 297 non-expert viewers. For a detailed discussion of the experiments and their results, the reader is referred to the VQEG report.<sup>21</sup>

#### 3.2. Test Sequences

For further evaluating the usefulness of sharpness and colorfulness ratings in Section 4.2, additional subjective experiments were conducted with the test scenes shown in Figure 3 and the test conditions listed in Table 1.

The nine test scenes were selected from the set of VQEG scenes to include spatial detail, saturated colors, motion, and synthetic sequences. They are 8 seconds long with a frame rate of 25 Hz. They were de-interlaced and subsampled from the interlaced ITU-R Rec. 601<sup>12</sup> format to a resolution of  $360 \times 288$  pixels per frame for progressive display. It should be noted that this led to slight aliasing artifacts in some of the scenes. Because of the DSCQS testing methodology used, this does not affect the results of the experiment, however.

**Table 1:** Test conditions.

Number	Codec	Version	Bitrate	Comments
1	Intel Indeo Video	3.2	2 Mb/s	Vector quantization
2	Intel Indeo Video	4.5	2 Mb/s	Hybrid wavelet
3	Intel Indeo Video	5.11	1 Mb/s	Wavelet transform
4	Intel Indeo Video	5.11	2 Mb/s	Wavelet transform
5	MSSG MPEG-2	1.2	2 Mb/s	MC-DCT
6	Microsoft MPEG-4	2	1 Mb/s	MC-DCT
7	Microsoft MPEG-4	2	2 Mb/s	MC-DCT
8	Sorenson Video	2.11	2 Mb/s	Vector quantization

The test conditions in Table 1 were created by means of software codecs. Except for the MPEG-2 codec of the MPEG Software Simulation Group (MSSG),<sup>†</sup> they are Windows AVI- and QuickTime-codecs. In contrast to the VQEG test conditions with a heavy focus on MPEG, these codecs use several different compression methods, including block-based DCT with motion compensation, vector quantization, the wavelet transform and hybrid methods. Adobe Premiere<sup>‡</sup> was used for interfacing with the Windows codecs. A keyframe (I-frame) interval of 25 frames (1 second) was chosen. Two of the six codecs were operated at two different bitrates for comparison, yielding a total of eight test conditions and 72 test sequences. No normalization or alignment was carried out.

\* See <http://www.crc.ca/vqeg/>.

† The source code is available at <http://www.mpeg.org/~tristan/MPEG/MSSG/>.

‡ See <http://www.adobe.com/products/premiere/main.html> for more information.



(a) Scene 1: mobile



(b) Scene 2: barcelona



(c) Scene 3: harp



(d) Scene 4: graphics



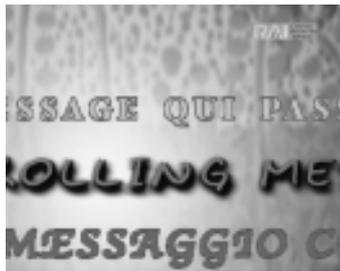
(e) Scene 5: canoe



(f) Scene 6: formula 1



(g) Scene 7: fries



(h) Scene 8: message



(i) Scene 9: rugby

**Figure 3:** Test scenes.

### 3.3. Subjective Experiments

The basis for the subjective experiments was ITU-R Rec. 500.<sup>11</sup> A total of 30 observers (23 males and 7 females) participated in the experiments. Their age ranged from 20 to 55 years; most of them were university students. The observers were tested for normal or corrected-to-normal vision with the help of a Snellen chart,<sup>\*</sup> and for normal color vision using three Ishihara charts.<sup>†</sup>

A 19 inch ADI PD-959 MicroScan monitor was used for displaying the sequences. Its refresh rate was set to 85 Hz, and its screen resolution was set to  $800 \times 600$  pixels, so that the sequences covered nearly one quarter of the display area. A black level adjustment was carried out for a peak screen luminance of  $70 \text{ cd/m}^2$ . The monitor gamma was determined through luminance measurements for different gray values  $y$ , which were approximated with the following function:

$$L(y) = \alpha + \beta \left( \frac{y}{255} \right)^\gamma, \quad (9)$$

with  $\alpha = -0.14 \text{ cd/m}^2$ ,  $\beta = 73.31 \text{ cd/m}^2$ , and  $\gamma = 2.14$ .

<sup>\*</sup> Available at <http://www.gimbel.com/check-yv.htm>.

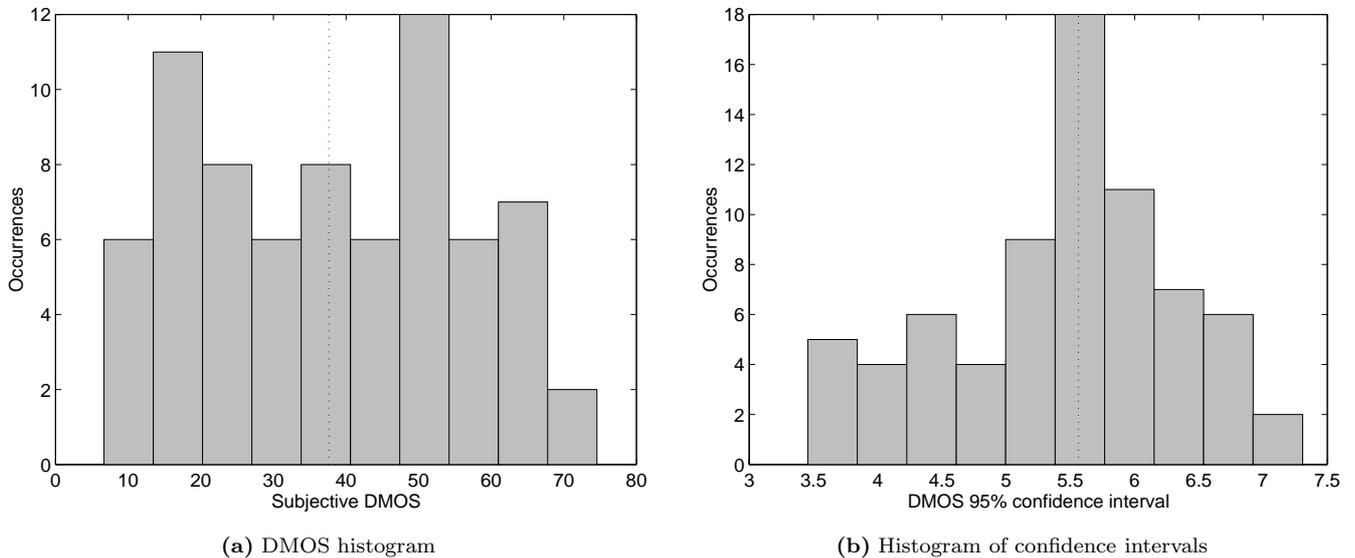
<sup>†</sup> Available at <http://www.toledo-bend.com/colorblind/Ishihara.html>.

As in the VQEG experiments, the Double Stimulus Continuous Quality Scale (DSCQS) method from ITU-R Rec. 500<sup>11</sup> (cf. Section 3.1) was selected for the subjective experiments. The subjects were introduced to the method and their task, and training sequences were shown to demonstrate the range and type of impairments to be assessed. The actual test sequences were presented to each observer in two sessions of 36 trials each. Their order was individually randomized so as to minimize effects of fatigue and adaptation. Windows Media Player 7\* with a handwritten “skin” (a uniform black background around the sequence) was used to display the sequences on the monitor. The viewing distance was 4-5 times the height of the active screen area.

After the experiments, post-screening of the subjective data was performed as specified in Annex 2 of ITU-R Rec. 500<sup>11</sup> to determine unstable viewers, but none of the subjects had to be removed.

### 3.4. Discussion

The resulting distributions of the differential mean opinion scores (DMOS) and their 95%-confidence intervals for all 72 test sequences are shown in Figure 4. As can be seen, the entire quality range is covered quite uniformly (the median of the rating differences is 38), as was the intention of the test, and in contrast to the VQEG experiments (where the median DMOS was 15). The size of the confidence intervals is also satisfactory (median of 5.6). As a matter of fact, they are not much wider than in the VQEG experiments.

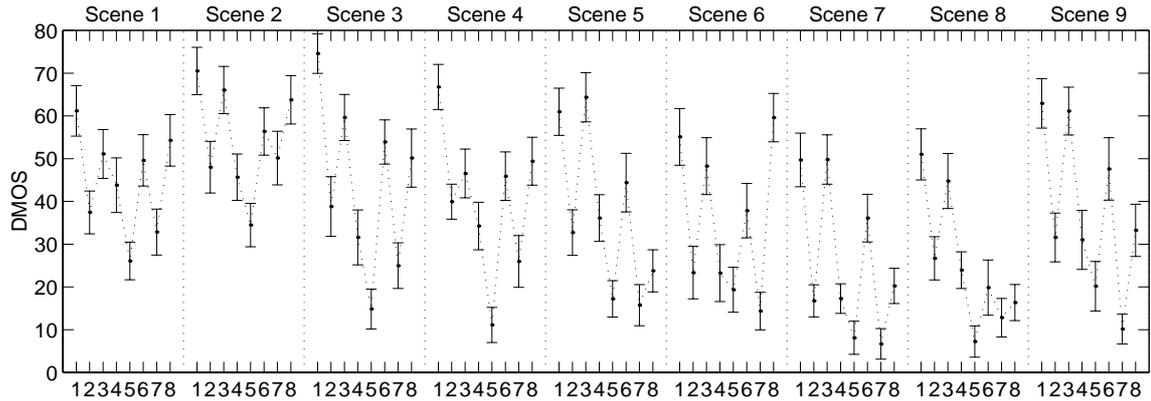


**Figure 4:** Distribution of differential mean opinion scores (a) and their 95% confidence intervals (b) over all test sequences. The dotted vertical lines denote the respective medians.

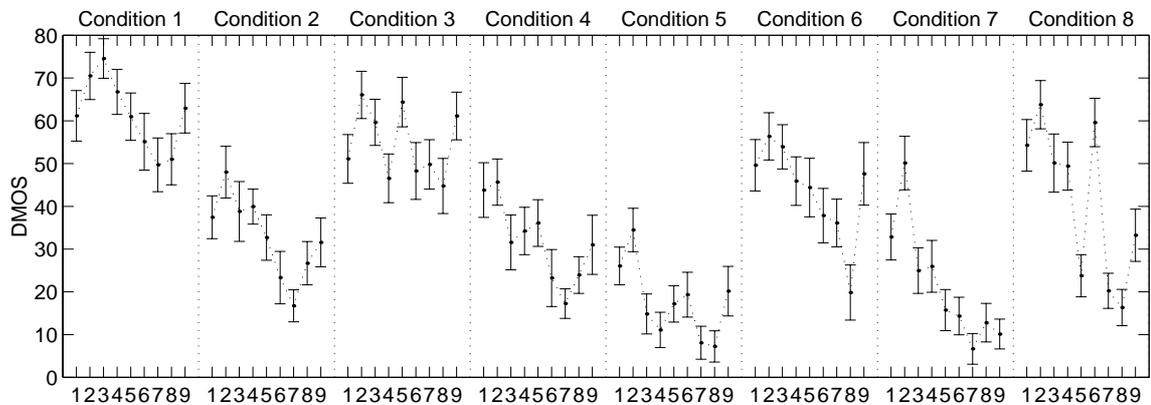
Figures 5(a) and 5(b) show the subjective DMOS and confidence intervals, separated by scene and by condition, respectively. The separation by test scene reveals that scene 2 (*barcelona*) is the most critical one with the largest distortions averaged over conditions, followed by scenes 1 (*mobile*) and 3 (*harp*). Scenes 7 (*fries*) and 8 (*message*) on the other hand exhibit the smallest distortions. Several subjects mentioned that scene 8 (a horizontally scrolling message) actually was the most difficult test sequence to rate, and this is also where most confusions between reference and compressed sequence (i.e. negative rating differences) occurred.

It is instructive to compare the compression performance of the different codecs and their compression methods. The separation by test condition in Figure 5(b) shows that condition 5 (MPEG-2 at 2 Mb/s) exhibits the highest quality over all scenes, closely followed by condition 7 (MPEG-4 at 2 Mb/s). Even at 1 Mb/s, the MPEG-4 codec (condition 6) outperforms conditions 1, 3, and 8. It should be noted that the Intel Indeo Video codecs and the Sorenson Video codec were designed for lower bitrates than the ones used in this test and obviously do not scale well at all, as opposed to MPEG-2 and MPEG-4. Comparing Figures 5(a) and 5(b) reveals that the perceived quality depends much more on the codec and bitrate than on the particular scene content in these experiments.

\* Available at <http://www.microsoft.com/windows/windowsmedia/en/software/Player7.asp>.



(a) DMOS for conditions 1 through 8 separated by scene



(b) DMOS for scenes 1 through 9 separated by condition

**Figure 5:** Subjective DMOS and confidence intervals for all test sequences separated by scene and by condition.

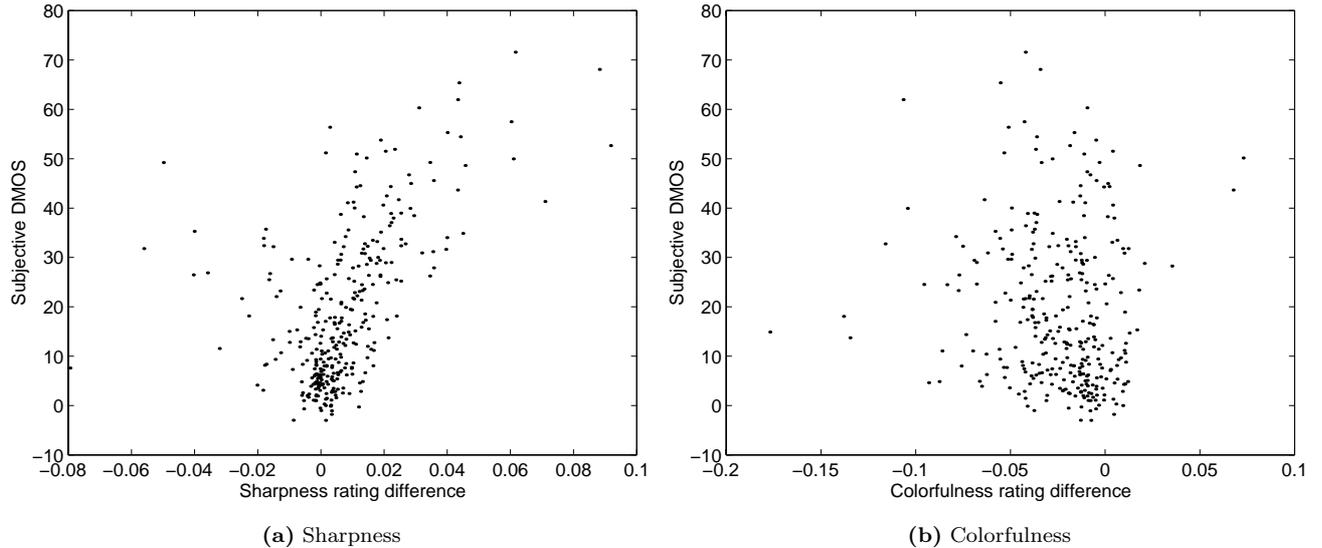
## 4. RESULTS

### 4.1. VQEG Data

The sharpness and colorfulness ratings were first computed for the 320 VQEG test sequences described in Section 3.1.<sup>21</sup> They are compared with the overall subjective quality ratings from the VQEG experiments in Figures 6(a) and 6(b). As can be seen, there exists a correlation of approximately 0.6 between the sharpness rating differences and the subjective quality ratings. The negative outliers are due almost exclusively to condition 1 (Betacam), which introduces noise and strong color artifacts, leading to an unusual increase of the sharpness rating.

Keep in mind that the sharpness rating was not conceived as an independent quality measure, but has to be combined with a fidelity metric such as the perceptual distortion metric (PDM). This combination is implemented as  $\Delta_{\text{PDM}} + w \max(0, \Delta_{\text{sharp}})$ , so that negative differences are excluded, and the sharpness ratings are scaled to a range comparable to the PDM predictions. Using the optimum  $w = 486$ , the correlation with subjective quality ratings increases by 5% compared to PDM-only predictions (see final results in Figure 8). This shows that the additional consideration of sharpness by means of a contrast measure improves the prediction performance of the PDM.

The colorfulness rating differences, on the other hand, are negative for most sequences, which is counter-intuitive and seems to contradict the above-mentioned premise. Furthermore, they exhibit no correlation at all with subjective quality ratings (see Figure 6(b)), not even in combination with the PDM predictions. This can be explained by the rigorous normalization with respect to global chroma and luma gains and offsets that was carried out on the VQEG



**Figure 6:** Perceived quality vs. sharpness and colorfulness rating differences for the VQEG sequences.

test sequences prior to the experiments.<sup>21</sup> When this normalization is reversed in the sequences, the colorfulness rating differences become positive for most sequences, as expected. However, the normalization cannot be undone for the VQEG subjective ratings, which were collected using the normalized sequences. Therefore, no conclusion about the effectiveness of the colorfulness rating can be drawn from the VQEG data. This is the reason additional subjective experiments with unnormalized test sequences were carried out as described above.

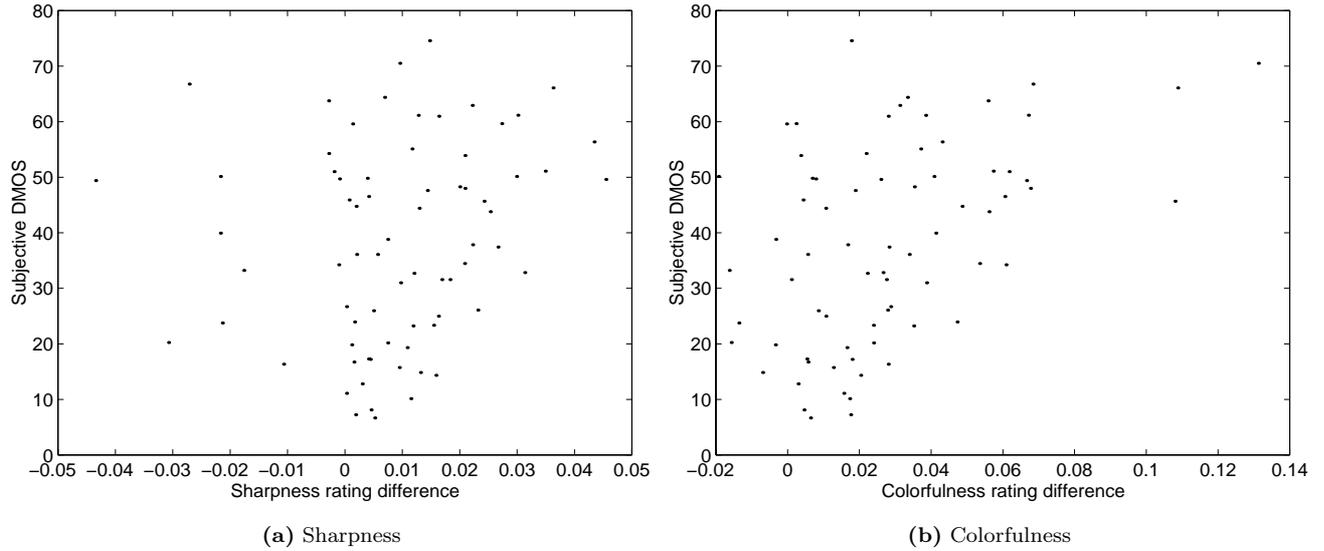
## 4.2. Test Sequences

The sharpness and colorfulness ratings were computed for the test sequences described in Section 3.2. The results are compared with the subjective quality ratings from Section 3.3 in Figures 7(a) and 7(b). The correlation between the subjective quality ratings and the sharpness rating differences is lower than for the VQEG sequences (cf. Section 4.1). This is mainly due to the extreme outliers pertaining to conditions 1 and 8. These conditions introduce considerable distortions leading to additional strong edges in the compressed sequences, which increase the overall contrast.

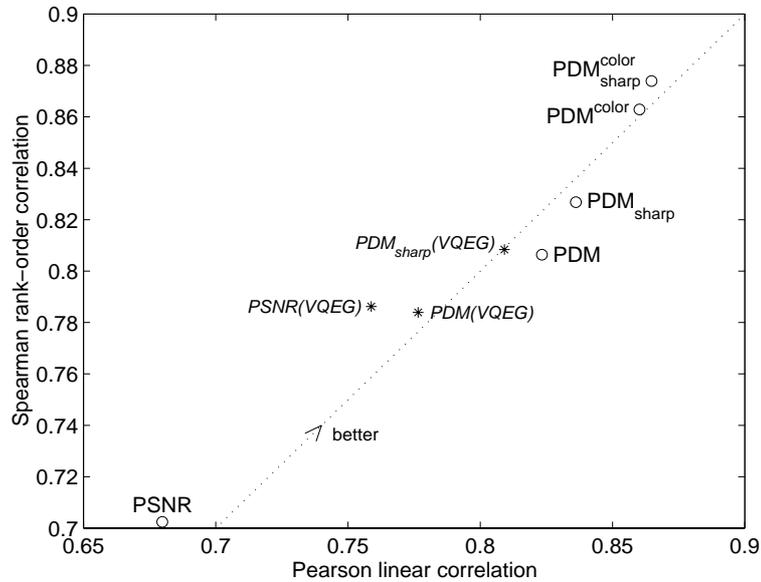
On the other hand, a correlation between colorfulness rating differences and subjective quality ratings can now be observed. This confirms our assumption that the counter-intuitive behavior of the colorfulness ratings for the VQEG sequences was due to their rigorous normalization. Without such a normalization, the behavior is as expected for the test sequences described above in Section 3.2, i.e. the colorfulness of the compressed sequences is reduced with respect to the reference for nearly all test sequences (see Figure 7(b)).

We stress again that neither the sharpness rating nor the colorfulness rating was designed as an independent measure of quality; both have to be used in combination with a visual fidelity metric. Therefore, the sharpness and colorfulness rating differences are combined with the output of the PDM as  $\Delta_{\text{PDM}} + w_{\text{sharp}} \max(0, \Delta_{\text{sharp}}) + w_{\text{color}} \max(0, \Delta_{\text{color}})$ . The rating differences are thus scaled to a range comparable to the PDM predictions, and negative differences are excluded. The results achieved with the optimum weights are shown in Figure 8.

It is evident that the additional consideration of sharpness and colorfulness improves the prediction performance of the PDM. The improvement with the sharpness rating alone is somewhat smaller than for the VQEG data. Together with the results discussed in Section 4.1, this indicates that the sharpness rating is more useful for sequences with relatively low distortions. The colorfulness rating, which is of low computational complexity, also gives a performance boost to the PDM predictions.



**Figure 7:** Perceived quality vs. sharpness and colorfulness rating differences for the test sequences described in Section 3.



**Figure 8:** Prediction performance of the PDM alone and in combination with image appeal attributes for the VQEG test sequences (stars) as well as the new test sequences (circles). PSNR correlations are shown for comparison.

## 5. CONCLUSIONS

Sharpness and colorfulness were identified as important attributes of image appeal. These attributes were quantified by defining a sharpness rating based on a measure of isotropic local contrast and a colorfulness rating derived from the distribution of chroma and saturation in the sequence. Extensive subjective experiments were carried out to establish a relationship between these ratings and perceived visual quality. The results show that a combination of PDM predictions with the sharpness and colorfulness ratings leads to improvements in prediction performance. The consideration of image appeal attributes such as sharpness and colorfulness as proposed here may prove useful in the development of reduced-reference metrics, which are much more versatile than traditional full-reference metrics.

## REFERENCES

1. P. Andrei: 1998, private communication.
2. J.-P. Antoine, R. Murenzi, P. Vandergheynst: "Directional wavelets revisited: Cauchy wavelets and symmetry detection in patterns." *Appl. Comp. Harm. Anal.* **6**(3):314–345, 1999.
3. C. Chiosso: 1998, private communication.
4. H. de Ridder, F. J. J. Blommaert, E. A. Fedorovskaya: "Naturalness and image quality: Chroma and hue variation in color images of natural scenes." in *Proc. SPIE*, vol. 2411, pp. 51–61, San Jose, CA, 1995.
5. E. A. Fedorovskaya, H. de Ridder, F. J. J. Blommaert: "Chroma variations and perceived quality of color images of natural scenes." *Color Res. Appl.* **22**(2):96–110, 1997.
6. J. D. Foley et al.: *Computer Graphics. Principles and Practice*. Addison-Wesley, 2<sup>nd</sup> edn., 1992.
7. R. E. Fredericksen, R. F. Hess: "Estimating multiple temporal mechanisms in human vision." *Vision Res.* **38**(7):1023–1040, 1998.
8. J.-F. Gobbers, P. Vandergheynst: "Directional wavelet frames: Design and algorithms." *IEEE Trans. Image Processing* 1999, submitted paper, preprint UCL-IPT-98-17.
9. R. C. Gonzalez, R. E. Woods: *Digital Image Processing*. Addison-Wesley, 1992.
10. R. W. G. Hunt: *The Reproduction of Colour*. Fountain Press, 5<sup>th</sup> edn., 1995.
11. ITU-R Recommendation BT.500-10: "Methodology for the subjective assessment of the quality of television pictures." ITU, Geneva, Switzerland, 2000.
12. ITU-R Recommendation BT.601-5: "Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios." ITU, Geneva, Switzerland, 1995.
13. J. Lynch: 1998, private communication.
14. S. Mallat, S. Zhong: "Characterization of signals from multiscale edges." *IEEE Trans. PAMI* **14**(7):710–732, 1992.
15. G. Marchand: 1999, private communication.
16. A. B. Poirson, B. A. Wandell: "Appearance of colored patterns: Pattern-color separability." *J. Opt. Soc. Am. A* **10**(12):2458–2470, 1993.
17. A. B. Poirson, B. A. Wandell: "Pattern-color separable pathways predict sensitivity to simple colored patterns." *Vision Res.* **36**(4):515–526, 1996.
18. A. E. Savakis, S. P. Etz, A. C. Loui: "Evaluation of image appeal in consumer photography." in *Proc. SPIE*, vol. 3959, pp. 111–120, San Jose, CA, 2000.
19. E. P. Simoncelli et al.: "Shiftable multi-scale transforms." *IEEE Trans. Inform. Theory* **38**(2):587–607, 1992.
20. P. Vandergheynst, M. Kutter, S. Winkler: "Wavelet-based contrast computation and its application to watermarking." in *Proc. SPIE*, vol. 4119, San Diego, CA, 2000, invited paper.
21. VQEG: "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment." 2000, available at <http://www.crc.ca/vqeg/>.
22. A. B. Watson, J. A. Solomon: "Model of visual contrast gain control and pattern masking." *J. Opt. Soc. Am. A* **14**(9):2379–2391, 1997.
23. S. Winkler: "Issues in vision modeling for perceptual video quality assessment." *Signal Processing* **78**(2):231–252, 1999.
24. S. Winkler: "A perceptual distortion metric for digital color video." in *Proc. SPIE*, vol. 3644, pp. 175–184, San Jose, CA, 1999.
25. S. Winkler: "Quality metric design: A closer look." in *Proc. SPIE*, vol. 3959, pp. 37–44, San Jose, CA, 2000.
26. S. Winkler: *Vision Models and Quality Metrics for Image Processing Applications*. Ph.D. thesis, École Polytechnique Fédérale de Lausanne, Switzerland, 2000.
27. S. Winkler, P. Vandergheynst: "Computing isotropic local contrast from oriented pyramid decompositions." in *Proc. ICIP*, vol. 4, pp. 420–424, Kyoto, Japan, 1999.
28. G. Wyszecki, W. S. Stiles: *Color Science: Concepts and Methods, Quantitative Data and Formulae*. John Wiley & Sons, 2<sup>nd</sup> edn., 1982.
29. S. N. Yendrikhovskij, F. J. J. Blommaert, H. de Ridder: "Perceptually optimal color reproduction." in *Proc. SPIE*, vol. 3299, pp. 274–281, San Jose, CA, 1998.