Color Image Quality on the Internet

Sabine Süsstrunk, Stefan Winkler

Audiovisual Communications Laboratory (LCAV)
Ecole Polytechnique Fédérale de Lausanne (EPFL)
1015 Lausanne, Switzerland

ABSTRACT

Color image quality depends on many factors, such as the initial capture system and its color image processing, compression, transmission, the output device, media and associated viewing conditions. In this paper, we are primarily concerned with color image quality in relation to compression and transmission. We review the typical visual artifacts that occur due to high compression ratios and/or transmission errors. We discuss color image quality metrics and present no-reference artifact metrics for blockiness, blurriness, and colorfulness. We show that these metrics are highly correlated with experimental data collected through subjective experiments. We use them for no-reference video quality assessment in different compression and transmission scenarios and again obtain very good results. We conclude by discussing the important effects viewing conditions can have on image quality.

Keywords: Compression artifacts, transmission errors, image and video quality metrics, blockiness, blurriness, colorfulness, jerkiness, JPEG2000, MPEG-4, viewing conditions

1. INTRODUCTION

The literature relating to color image quality is vast, and encompasses such different research areas as camera design and visual psychology. Any "good" image requires an adequate capturing system where optics and sensors are well matched. The image processing chain from raw sensor data to image data that represent the desired color appearance on a defined output medium under defined viewing conditions has been studied extensively. There are several international standards that define image-state architecture,[1] standard image encodings,[2,3] and means to transform between them or to output devices.[4] Content-specific color image processing algorithms, such as adaptive tone reproduction,[5] color constancy,[6] and retinex,[7] inspired by the way the human visual system processes visual information, have added to the overall quality of digital imaging results. In short, most of today's digital image capture systems are capable of producing acceptable, if not excellent color images and/or videos.

However, images are often compressed for storage or for transmission over a network. High compression rates can only be achieved with "lossy" compression schemes, where part of the original image data is thrown away in the compression process. Additionally, bit errors and packet losses can occur when transmitting the bitstreams over a network. As a result, the decompressed image or video at the destination output device may contain visible artifacts that are detrimental to the overall quality. In Section 2, we discuss some common compression and transmission artifacts. In Section 3, we outline approaches to measure the resulting artifacts. In Section 4, we describe three artifact metrics for blockiness, blurriness, and colorfulness. Their applicability as quality predictors is discussed in Section 5. In Section 6, we describe how we tested their performance for quality prediction in video applications.

No discussion on artifacts, either due to wrong capture and processing parameters, or due to compression and transmission, is complete without considering their visual impact. One of the parameters often overlooked in the coding community are the viewing conditions of the output. For example, a perfect image on one computer screen might not be acceptable on a different monitor. In Section 7, we discuss different output and viewing condition parameters and show how they influence the appearance of an image or video.

E-mail of corresponding author: sabine.susstrunk@epfl.ch

2. ARTIFACTS

Due to their large data size, images and video have to be compressed for storage and transmission. Lossy compression is typically required for a substantial reduction of the data rate. This is usually achieved by transforming the image data to a domain suitable for compression and applying quantization to the resulting transform coefficients. Examples of such transforms include the Discrete Cosine Transform (DCT), as used in JPEG or MPEG compression standards, and the Discrete Wavelet Transform (DWT), as used in JPEG2000.[8] Depending on the specific compression algorithm, this can introduce a variety of artifacts.[9] We list some of the more common ones here:

- The *blocking effect* or blockiness refers to a block pattern in the compressed image or video (cf. Figure 1). It is due to the independent quantization of individual blocks (usually of 8×8 or 16×16 pixels in size) in block-based DCT coding schemes, leading to discontinuities at the boundaries of adjacent blocks. Due to the regularity and extent of the resulting pattern, the blocking effect is easily noticeable.
- *Blurriness* manifests itself as a loss of spatial detail and a reduction of edge sharpness (cf. Figure 1). It is due to the suppression of high-frequency coefficients by coarse quantization, which is applied in almost any lossy compression scheme. It can be further aggravated by deblocking filters, which are sometimes used in the decoder to reduce the above-mentioned blocking effect.
- *Ringing* is fundamentally associated with Gibbs' phenomenon and is thus most evident along high-contrast edges in otherwise smooth areas. It is a direct result of quantization leading to high-frequency irregularities in the reconstruction. Ringing and blurriness constitute the main artifacts of wavelet compression.
- *Color bleeding* is the smearing of the color between areas of strongly differing chrominance. It results from the suppression of high-frequency coefficients of the chroma components. Due to chroma subsampling, color bleeding extends over an entire macroblock. Many observers perceive this also as a loss of colorfulness (cf. Figure 4).
- *Jerkiness* represents artifacts of motion rendition in video, for example due to a varying or reduced frame rate chosen by the encoder. It can also occur in live streaming over a network due to missing or erroneous parts of the bitstream.
- *Flickering* appears when a video has high texture content. Texture regions are compressed with varying quantization factors over time, which results in a visible flickering effect.

3. QUALITY METRICS

Depending on the compression ratio, these artifacts are more or less severe. In order to measure and control their visual impact, reliable quality metrics are needed. However, the accurate measurement of quality as perceived by a human observer is a great challenge, because the amount and visibility of the artifacts strongly depend on the underlying image content.

Subjective experiments, which to date are the only widely recognized method of determining actual perceived quality, are complex and time-consuming, both in their preparation and execution. Basic fidelity measures like mean-squared error (MSE), peak signal-to-noise ratio (PSNR), or Delta E (ΔE) on the other hand may be simple and very popular, but they do not necessarily correlate well with perceived quality. Additionally, these measures assume that there exists a reference in the form of an "original" to compare to, which restricts their usability. Thus, reliable automatic methods for visual quality assessment are needed. Ideally, such a quality assessment system would perceive and measure image or video impairments just like a human being. Two approaches can be taken:

- The "psychophysical approach", which is based on models of the human visual system.[10] Their general structure is usually determined by the modeling of visual effects, such as color appearance, contrast sensitivity, and visual masking, to name a few. Due to their generality, these metrics can be used in a wide range of video applications; the downside to this is the high complexity of the underlying vision models. Besides, the visual effects modeled are best understood at the threshold of visibility, whereas image distortions are often supra-threshold.

- The “engineering approach”, where metrics make certain assumptions about the types of artifacts that are introduced by a specific compression technology or transmission link. Such metrics look for the strength of these distortions in the video and use their measurements to estimate the overall quality.

Quality metrics can be further classified into the following categories:

- Full-reference (FR) metrics perform a direct comparison between the image or video under test and a reference or “original”. They thus require the entire reference content to be available, usually in uncompressed form, which is quite an important restriction on the usability of such metrics. Another practical problem is the alignment of the two, especially for video sequences, to ensure that the frames and image regions being compared actually correspond. As mentioned above, fidelity metrics such as MSE/PSNR and ΔE belong to this class as well.
- No-reference (NR) metrics look only at the image or video under test and have no need of reference information. This makes it possible to measure the quality of any visual content, anywhere in an existing compression and transmission system. The difficulty here lies in telling apart distortions from regular content, a distinction humans are able to make from experience.
- Reduced-reference (RR) metrics lie between these two extremes. They extract a number of features from the reference image or video (e.g. spatial detail, amount of motion). The comparison with the image/video under test is then based only on those features. Additionally, image metadata as available with some file formats (e.g. EXIF, JPEG2000) can also be used. This makes it possible to avoid some of the pitfalls of pure no-reference metrics.

4. NO-REFERENCE ARTIFACT METRICS

Our focus in this paper are no-reference (NR) metrics because of their versatility and flexibility. Since no information about the reference is required, quality can be measured even in cases when the reference is not accessible, for example at the receiver side of an Internet streaming transmission, or completely unavailable, such as an image taken with a digital camera, where the camera-internal processing parameters are unknown and the metadata lost. Furthermore, there are no alignment issues whatsoever. To be able to measure quality in the absence of reference information, NR metrics must make certain assumptions about the types of artifacts that are introduced by a specific compression technology or transmission link. They consequently look at the strength of the different artifacts in the image or video.* The separate measurements can then be combined into an estimate of overall quality (see Sections 5 and 6).

4.1. Blockiness Metric

Our no-reference blockiness metric assumes that the blocks introduced in the image form a regular grid.[11] This regularity becomes apparent by analyzing the image in the Fourier domain. We first compute horizontal and vertical difference signals of each row and column, respectively. By applying 1-D discrete Fourier transforms to these signals, we compute the power spectra and average them over all rows and columns, respectively. An example of such an average 1-D spectrum obtained from a JPEG-compressed image is shown in Figure 2. The peaks in this averaged spectrum are due to periodic block structures. They appear at specific locations in the spectrum, depending on the block size (e.g. at multiples of $N/8$ for blocks of size 8×8 pixels, where N is the DFT/image size). The power spectrum of the image without the blocks at the locations of the peaks can be approximated by median-filtering these curves. The overall blockiness measure is then computed as the difference between these two power spectra at the locations of the peaks. To further enhance the measurement, the integration of visual masking effects has also been proposed.[12]

* While the algorithms described here were designed primarily as no-reference metrics, they can also be used in a full-reference scenario; in that case, the assumptions or estimations of the reference image would be replaced by the actual values.

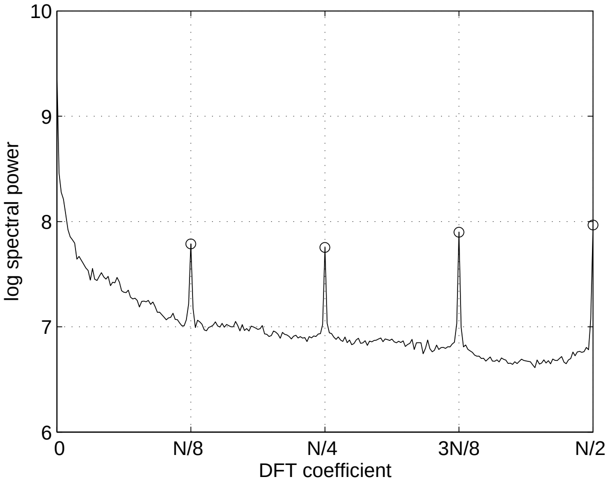

Figure 2. Blockiness metric: 1-D power spectrum of horizontal differences, averaged over all rows, for a JPEG-encoded image. The circles mark the spectrum peaks due to the 8×8-block structure in the image. The height of these peaks relative to the median-filtered version of the spectrum is used as a measure of blockiness.

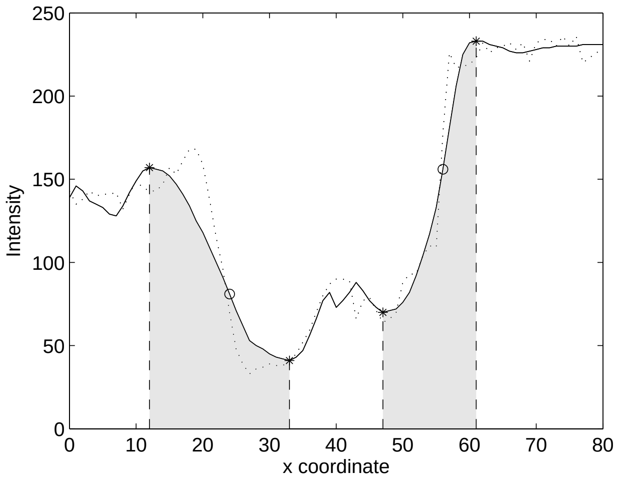

Figure 3. Blurriness metric: One row of a JPEG2000-encoded image (solid line). Note the smearing of edges compared to the sharper original (dotted line). The circles indicate the significant edges detected, the stars and gray patches mark the local minima and maxima around the edge. The edge width between these local extrema is used as a measure of blurriness.

4.2. Blurriness Metric

Our technique for measuring blurriness is based on the assumption that most significant edges in an image, which often represent borders of objects, are sharp. Compression has a smearing effect on these edges, the extent of which our blurriness metric attempts to measure. The algorithm is summarized as follows. First, an edge detector (e.g. a Sobel filter) is applied to the luminance component of the image. Thresholding the edge gradients removes noise and insignificant edges. The start and end positions of each significant edge in the image are defined as the locations of the local intensity minimum and maximum closest to the edge. The distance between these two points is identified as the local blur measure for this edge location. An example for an image row is illustrated in Figure 3. The global blurriness for the whole image is estimated by averaging the local blur values over all significant edges found. This metric was confirmed to be a reliable predictor of perceived blur both for images filtered with Gaussian low-pass filters and JPEG2000-encoded images.[13]

4.3. Colorfulness Metric

Measuring the perceived color quality of an image is extremely difficult. The quality of the reproduction depends on content, media and viewing conditions (see Section 7). On the other hand, compressing images can also reduce their colorfulness at high compression rates. Suppressing the high-frequency coefficients of the chroma components not only introduces color bleeding, but also makes images appear less colorful, especially if an image contains a lot of chromatic details (two examples are shown in Figure 4). We have developed a metric for such artifacts. Our measure for colorfulness is based on the mean and standard deviations of two axes in a simple opponent color representation with $\alpha = R - G$ and $\beta = \frac{1}{2}(R + G) - B$. The metric is defined as follows:[14]

$$M = \sqrt{\sigma_\alpha^2 + \sigma_\beta^2} + 0.3\sqrt{\mu_\alpha^2 + \mu_\beta^2}$$

where $\sigma_\alpha, \sigma_\beta$ and μ_α, μ_β are the standard deviation and mean of the pixel cloud along directions α and β, respectively. For 84 test images, the correlation of this metric with experimental data on perceived colorfulness was 95%. Note, however, that while our metric is capable of predicting the perceived colorfulness, it does not provide a quality prediction and as such cannot be used directly for no-reference quality assessment. It could be applicable in a reduced-reference framework where the colorfulness of the original image is known.

5. IMAGE QUALITY PREDICTION

Once a metric has been designed, its performance as an image quality predictor has to be evaluated. This can only be done by comparing the results of the metric to observer ratings. How to design an appropriate subjective experiment is not trivial, and it is important not to underestimate this task.[15] Indeed, determining the correct type of experiment, such as rank-ordering, categorical, forced choice, etc. that is most appropriate is very important. Additionally, the number of visual samples, the viewing conditions (see Section 7), the number and experience of the observers all influence the results. Another critical parameter of the subjective experiment are the instructions given to the observers. The ISO is working on an international standard for still image quality evaluation (ISO/DIS 20462-2[16–18]), and the ITU has established standard test procedures for subjective video quality evaluations (ITU-T Rec. P.910[19] and ITU-R Rec. BT.500[20]). Consequently, if one metric is compared to another, the results are only meaningful if they are applied to the same dataset under the same test conditions with the same instructions, as was done for example by the Video Quality Experts Group (VQEG).[21, 22]

Here we summarize the results of an evaluation of the prediction performance of the NR blurriness metric as a predictor of overall perceived image quality.[23] The LIVE Image Quality Assessment Database[24] was used for these tests. The images in this database were created by compressing 29 color images (typically of size 768×512 pixels) using Kakadu's JPEG2000 encoder. Compression ratios range from 7.5 to 800, for a total of 169 compressed images. The subjective experiments were conducted in two separate sessions with 29 and 25 observers, respectively; the original uncompressed images were included in both. Observers provided their quality ratings on a continuous scale from 1 (lowest quality) to 100 (highest quality).

As shown in Table 1, PSNR is already an excellent predictor of perceived quality for this database: the correlation with the mean opinion score (MOS) is about 91%. These good results can be attributed largely to the fact that the database contains exclusively images created with a single type of encoder (JPEG2000) and thus contain mainly varying degrees of the same distortions.

The hypothesis for using the NR blurriness metric from Section 4.2 is that the quality prediction is a simple non-linear transform of the measured blur for this dataset. To test this, we separated the LIVE test images into a training set and a test set, using 100 different random divisions of the dataset. As shown in Table 1, our metric achieves correlations of around 85% with MOS on the test sets, which is quite a good prediction performance for an NR metric.

A closer look at the data reveals that the most significant outliers are due to two specific pictures, namely one close-up and one macro shot, both with very small depths of field. Since our blurriness metric does not distinguish between blur as a compression artifact and any other blur in the image, its MOS predictions for these images are too low in comparison to the observers' ratings, who do not consider this type of blur a quality degradation. In one form or another, this problem is intrinsic to any no-reference metric when its assumptions about the source of the artifacts are violated. An added detector for distinguishing central objects from the potentially out-of-focus background could help alleviate this problem when using our metric for the assessment of compression artifacts. In fact, when these two images are removed from the test set, the prediction performance of our NR metric approaches that of PSNR (see Table 1).

Table 1. Prediction performance of the proposed quality metric. The bottom row refers to the exclusion of the images with very small depth of field (see text), an effect that is difficult to distinguish from compression-induced blur for a no-reference metric.

	Linear correlation	Rank-order correlation	Prediction error
PSNR	91%	92%	9.7
NR quality metric	86%	84%	12.1
NR metric w/o outliers	90%	88%	10.1

6. VIDEO QUALITY PREDICTION

The metrics described above were also used for no-reference video quality assessment. Here we briefly present experiments and results for two sets of applications, namely video streaming over a packet network[25] and mobile/wireless video transmission.[26] In addition to video compression artifacts, we also consider the effects of transmission of the compressed bitstream over a network.

The test conditions are described in more detail in the respective sections below. They were chosen so as to produce typical video quality for each of the two applications and at the same time achieve a good distribution of qualities for the different scenes. The source sequences were chosen to cover a wide range of typical content for streaming applications, such as news, sports and music video clips. Furthermore, they were selected to span a wide range of coding complexity. Specifically, two one-minute test sequences were created by concatenating scenes taken from clips used in previous tests by MPEG[27] and VQEG[21] as well as other sources.

The subjective ratings obtained in the experiments (see below) were again used to tune and evaluate the MOS predictions based on the no-reference metrics for blockiness and blurriness described above in Section 4. In addition to these two spatial image metrics, which are computed on a frame-by-frame basis, a *jerkiness metric* is used to specifically take into account temporal distortions, such as frame drops or video freeze. It is based on an estimate of the instantaneous frame rate and the motion content in the video. All three artifact metrics are computationally very light, a very useful property for video quality prediction. This makes it possible to compute them in real-time on a standard PC, in parallel to decoding and displaying the video.

Due to the different types of artifacts that are produced by the codecs used in the tests, individual mappings were determined for each codec separately. For example, the MOS prediction for the MPEG-4 videos relies mainly on the blockiness metric. Tuning was again performed on a randomly selected half of the data, and the other half was used for evaluation in each case.

6.1. Subjective Experiments

Subjective assessment was based on ITU-R Rec. BT.500[20] and ITU-T Rec. P.910.[19] We used Single Stimulus Continuous Quality Evaluation (SSCQE), which is specified in ITU-R Rec. BT.500, as the assessment method for our subjective experiments. In an SSCQE session, a series of video sequences is presented to the viewer. The video sequences may or may not contain impairments. Subjects evaluate the *instantaneous* quality in real time using a slider with a continuous scale. The SSCQE method yields quality ratings at regular time intervals and can thus capture the perceived time variations in quality. The ratings are absolute in the sense that viewers are not explicitly shown the reference sequences. This corresponds well to an actual home viewing situation, where the reference is also not available to the viewer. In our experiments, the slider position was recorded every 50 ms, on a scale from 0 ("bad") to 100 ("good").

For our test setup, we found subjects to be comfortable at a viewing distance of 3-4 times the height of the video picture, which corresponds to about 30-40 cm in our setup. Since mobile devices and the majority of PC's sold today have an LCD screen, the monitors used in the subjective assessments are also LCD screens.[25, 26]

20 non-expert viewers – mostly university students – participated in each test. They were screened for normal visual acuity or corrective glasses and normal color vision.

6.2. Video Over Packet Network

The purpose of this test was to simulate video streaming over a packet network such as the Internet. At the compression stage, the following encoders and video formats were used:

- Windows Media Video 8;*
- Real Video 8;*
- ISO MPEG-4[28] (Microsoft implementation).†

* Version 8 of the Real and Windows Media codecs was the latest version at the time of the tests.

† To facilitate streaming with the tools at hand, the ISO MPEG-4 codec provided with the Windows Media Encoder was used. It encapsulates the MPEG-4 stream inside the WMV file format.

These codecs are probably the most popular ones for Internet streaming applications. At the transmission stage, an IP network simulator was used to simulate different network conditions. A large number of trials was required to obtain representative test sequences. Specifically, different packet loss rates (PLR) were selected that would result in test videos with noticeable artifacts, but without completely destroying the video. The test conditions are summarized in Table 2. Videos had a frame size of 360 × 288 pixels and a frame rate of 25 fps.

Table 2. Test conditions for video over packet network.

#	**Codec**	**Bitrate**	**PLR**
1	WMV8	256 kb/s	–
2	RV8	256 kb/s	–
3	RV8	256 kb/s	2%
4	WMV8	512 kb/s	–
5	RV8	512 kb/s	–
6	RV8	512 kb/s	3%
7	MPEG-4	512 kb/s	–

At the receiving end, the video was decoded while keeping track of the exact timing of frame display as encountered during playback (including picture freeze and playback irregularities). This allowed us to reproduce these temporal distortions during the subjective experiments.

The prediction performance is summarized in Table 3. The overall quality of the MOS predictions is characterized by a correlation of 78% and an average prediction error of 9.1 (on the 0-100 SSCQE scale), which is roughly the same as the confidence interval size in the subjective experiments. Quality prediction works well for all three codecs, considering that it is based only on no-reference metrics. A comparison with PSNR or other full-reference metrics is not possible here because the encoders and the packet loss simulations introduced frame rate variations and delays in the video which made an alignment with the source sequences impossible.

Table 3. MOS prediction performance

	Linear correlation	Rank-order correlation	Prediction error
Real Media	76%	76%	10.2
Real Media (no PL)	84%	83%	9.1
Windows Media	84%	85%	7.7
MPEG-4	83%	84%	6.8
Overall	**78%**	**79%**	**9.1**

One problem is the noticeable deterioration of the prediction accuracy with the inclusion of conditions with packet loss. The reason for this appears to be that people respond rather slowly to the sudden effects packet losses have on the video – it takes them some time to realize that the video has frozen. This gradual viewer response obviously cannot be taken into account with memoryless metrics. Modifying the metric predictions to produce the same gradual response could be done quite easily; on the other hand, an immediate response of the metrics in such a case may be preferred in monitoring applications.

6.3. Video Over Wireless

In this experimental setup we simulated the transmission of video sequences over a WCDMA wireless channel. The video source was first compressed using MPEG-4* or Motion JPEG2000.†,[8] Both coding standards include a number of tools to improve their resilience to transmission errors, which makes them well suited for mobile/wireless video applications. By exploiting inter-frame redundancy, MPEG-4 has a higher coding efficiency at the cost of a higher complexity. The dependencies between coded frames and the resulting propagation of

* MoMuSys reference software implementation,[29] available for download from http://megaera.ee.nctu.edu.tw/mpeg/

† Kakadu software, available for download from http://www.kakadusoftware.com/

errors across consecutive frames also imply a lower error resilience. Conversely, Motion JPEG2000, which is based on intra-frame coding, has a lower coding efficiency at the benefit of a reduced complexity. Additionally, it is more resilient to transmission errors, because each frame is coded independently.

After compression, H.223[30] was used for packetization and cyclic redundancy check (CRC) of the bitstream. Transmission errors were introduced using bit error patterns representative of WCDMA.[31] We selected two distinct error patterns with a Bit Error Rate (BER) of 10^{-4}. The random nature of transmission errors was simulated by applying different circular shifts to the bit error patterns. Again, a large number of trials was required to obtain representative test sequences.*

The test conditions are summarized in Table 4. The sequences were downsampled spatially and temporally as specified in order to accommodate the low bitrates. Contrary to what was observed in the tests with packet losses (Section 6.2), the bit errors did not lead to dropped frames or delays in the decoded video.

Table 4. Test conditions (MJ2K = Motion JPEG2000).

#	Frame size	Frame rate	Codec	Bitrate	Bit Error Rate
1	180×144	4 fps	MPEG-4	64 kb/s	—
2	180×144	4 fps	MJ2K	64 kb/s	—
3	180×144	4 fps	MJ2K	64 kb/s	10^{-4} (I)
4	180×144	4 fps	MJ2K	64 kb/s	10^{-4} (II)
5	180×144	6 fps	MPEG-4	128 kb/s	—
6	180×144	6 fps	MJ2K	128 kb/s	—
7	180×144	6 fps	MJ2K	128 kb/s	10^{-4} (I)
8	180×144	6 fps	MJ2K	128 kb/s	10^{-4} (II)
9	360×288	8 fps	MPEG-4	384 kb/s	—
10	360×288	8 fps	MJ2K	384 kb/s	—
11	360×288	8 fps	MJ2K	384 kb/s	10^{-4} (I)
12	360×288	8 fps	MJ2K	384 kb/s	10^{-4} (II)

The prediction performances are summarized in Table 5. The MOS prediction works very well – it is characterized by correlations of around 90%, even higher than what was obtained in the tests with video over a packet network (Section 6.2), despite the fact that transmission error effects are not always measured correctly by the three artifact metrics. For comparison, PSNR correlation with the same MOS data is only around 40%. This shows that no-reference metrics, even rather simple ones, can be very effective in estimating perceived quality if they are designed to measure application- and codec-specific artifacts.

Table 5. MOS prediction performance.

	Linear correlation	Rank-order correlation	Prediction error
MPEG-4	91%	89%	8.2
M-JPEG2000	93%	89%	7.1
Overall	**93%**	**89%**	**7.4**
PSNR	39%	43%	—

* While this setup is a simplification over implementing the complete WCDMA protocol stack and air-interface, this methodology is similar to the one used by 3GPP.[32]

7. VIEWING CONDITIONS

As mentioned previously, the viewing conditions have a significant influence on the appearance of an image or video, because they can amplify or diminish the visibility of artifacts. Thus, standard viewing conditions have been established for critical evaluations of images viewed on screen or in print.[33] Subjective evaluation standards[16–20] define the viewing conditions under which the tests should take place. Here, we review the parameters that influence appearance on a color monitor. For more information about print viewing conditions, refer to ISO 3664.[33] Note that in general, quality requirements for prints are a lot higher than quality requirements for images or videos viewed on a screen.

7.1. Visual Phenomena

There are many visual phenomena that influence the appearance of images on a monitor, but here we restrict the discussion to two: contrast sensitivity and adaptation. Contrast sensitivity is the ability of the human visual system to distinguish changes in luminance or chromaticity. Any given contrast sensitivity depends on the luminance level of the contrasting stimuli, their spatial frequency, their chromaticity, and on the state of adaptation of the human observer.

Contrast is usually modeled with the Weber law $C = \frac{\Delta L}{L}$, where ΔL is the difference in luminance between a stimulus and its surround, and L is the luminance of the surround. The threshold contrast, i.e. the minimum change in luminance necessary to detect a change, remains nearly constant over the luminance range important for imaging applications, i.e. from 10 - 1000 cd/m^2. However, the sensitivity to contrast also depends on the spatial and temporal frequency, and the chromaticity of the stimuli (see Figure 5).

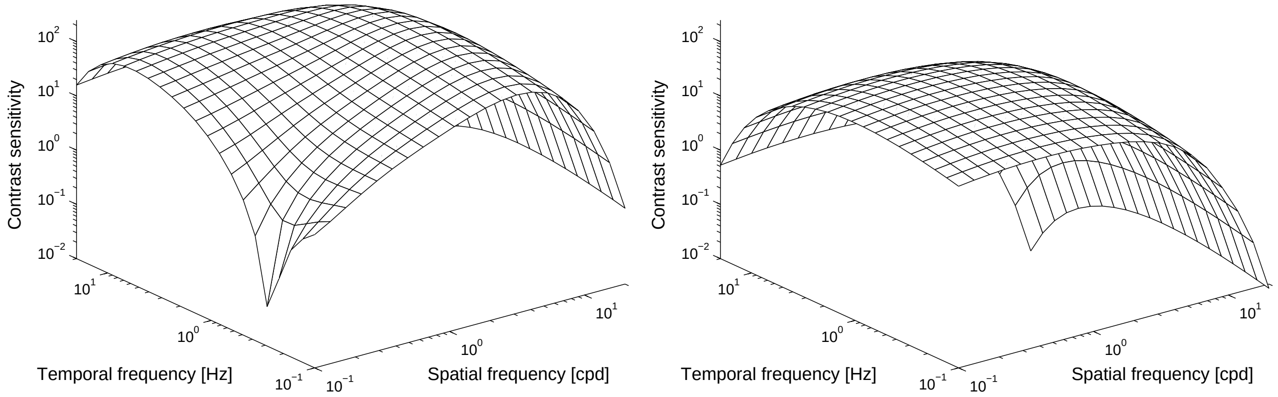

Figure 5. Approximations of achromatic (left) and chromatic (right) spatio-temporal contrast sensitivity functions.[34–36]

The contrast sensitivity behavior of the human visual system has been incorporated into JPEG and MPEG compression standards: High-frequency components are suppressed first; the lower chromatic contrast sensitivity is taken into account by subsampling the chroma. The assumption is that at a given compression rate, the artifacts introduced in such a way are not visible, i.e. the decoded image contains errors that are small and of low enough contrast to be classified as invisible. This is known as visually lossless compression. Of course, the definition of “small” depends on the output device and viewing conditions, such as display resolution and viewing distance. “Low contrast” depends on the monitor and ambient illuminant conditions discussed below.

Adaptation can be considered as a dynamic mechanism of the human visual system to optimize the visual response to a particular viewing condition. Dark and light adaptation are the changes in visual sensitivity when the level of illumination is decreased or increased, respectively. Chromatic adaptation is the ability of the human visual system to discount the color of the illumination to approximately preserve the appearance of an object. It can be observed by examining a white object under different types of illumination, such as daylight and incandescent light. Daylight contains far more short-wavelength energy than incandescent, and is "bluer." However, the white object retains its white appearance under both illuminants, as long as the viewer is adapted to the light source.

7.2. Image Appearance

Standard output-referred color image encodings, such as sRGB[2] and ROMM,[3] and to a lesser extent video encodings in ITU-R Rec. BT.709,[37] are based on a reference output device and reference viewing conditions (a CRT monitor environment in the case of sRGB and a print environment in the case of ROMM). Output-referred image data represent the color-space coordinates of the elements of an image that has undergone color rendering appropriate for a specified real or virtual output device and viewing conditions. A single scene can be color rendered to a variety of output-referred representations depending on the anticipated output viewing conditions, media limitations, and/or artistic intent.[1]

If the real output device characteristics and viewing conditions do not correspond to the intended output, the image appearance will change and image quality can decrease when the difference becomes too large. In case of monitor output, the following parameters should be considered:

- The *monitor luminance* level of the white-point should be greater than 100 cd/m^2. While LCD monitors for desktop PCs can easily achieve this today, it is limiting for CRTs and especially for battery-powered LCD displays. The maximum luminance level has an influence on the dynamic range of the display.
- The *level of ambient illumination* should be significantly lower than the luminance level of the monitor white point. This is partly to ensure that the observer is reasonably adapted to the monitor, but primarily to ensure that the full contrast range of the monitor is not significantly reduced by the effects of veiling glare (see below). For monitors with a white-point luminance of 100 cd/m^2, the illuminance falling on the monitor should be around 64 lux or less. For comparison, the illuminance falling on a monitor in typical office environments is around 300 lux. The *chromaticity* of the ambient illumination should be close to the chromaticity of the monitor.
- *Veiling glare* is reflected light that does not originate from the monitor. Although primarily influenced by the amount of ambient illumination, any other reflection can contribute, such as a window in the field of view, light clothing, etc. Veiling glare lightens and reduces the contrast of the darker parts of an image.
- The *chromaticity* of the white-point of the monitor should match the white-point defined in the color image encoding, i.e. D65 for sRGB and ITU-R Rec. BT.709. When just viewing an image on a monitor, this point is less critical. We adapt to the white-point of the monitor when the ambient illuminant is low.
- The dynamic range, determined by the monitor white-point luminance, the black-point luminance, the viewing glare, and the chromaticities of the phosphors determine the *color gamut* of the monitor. The smaller the gamut, the less colorful an image will appear.
- The *surround conditions* of an image or video on the screen should be neutral to avoid any simultaneous contrast effects.

Reducing the dynamic range, and by consequence the color gamut, can be of advantage to "hide" certain artifacts. If the dynamic range is reduced, contrast is reduced. Thus, certain artifacts might fall below the visibility threshold of the observer. The same applies when comparing moving and still pictures. Motion can mask a lot of artifacts that would be visible in still pictures. Therefore, it is important that a quality metric be evaluated in similar conditions as the intended output. Ideally, any given metric is flexible enough to "scale" according to output conditions.

However, if an image or video is displayed in conditions significantly different from the intended output, other artifacts may occur. For example, veiling glare will not only reduce the contrast in the shadows, but also the overall contrast. As a result, fewer image details are visible. A simulation of this effect is illustrated in Figure 6.

Additionally, color rendering systems in printers usually try to render an image based on the assumed image appearance communicated by the color image encoding. Today, a printer usually assumes the image is encoded in sRGB,* and tries to match that appearance with device specific paper and colorants. If the actual viewing conditions differ substantially from the intended conditions, a cross-comparison between monitor and print becomes very difficult. The print will then look substantially different from the monitor image. This is especially important if any artistic rendering decisions are taken based on the monitor image.

In general, the quality requirements for images/videos viewed on a monitor are lower than the quality requirements for prints. This is primarily due to the limits on dynamic range and gamut that reduce apparent contrast and mask quantization artifacts. Furthermore, the spatial resolution requirements for screen display are usually much lower than for a print. On the other hand, the adaptation of observers to the monitor white-point ensures that no color cast is perceived independent of which white-point setting is chosen. As far as video is concerned, the decrease in contrast sensitivity with increasing temporal frequencies together with the masking effect of high motion activity hides artifacts that would otherwise be visible on a single frame.

8. DISCUSSION AND CONCLUSIONS

Color image quality depends on many factors. In this paper, we were primarily concerned with artifacts introduced by current compression algorithms and transmission errors. We presented three artifact metrics and discussed their prediction performance on images. We also tested them as no-reference video quality predictors and obtained high correlations with subjective mean opinion scores.

Nonetheless, it may be worth recalling what the British politician Benjamin Disraeli (1804 - 1881) once said: "There are three kinds of lies: lies, damned lies, and statistics." This statement, unfortunately, also applies to many image quality and subjective evaluation studies. Care should be taken to make sure that the image samples, the testing conditions, the evaluation parameters and the statistical analysis are really meaningful to assess as subjective an attribute as "quality." For example, while the statistics used in this paper to characterize the prediction performance of video quality metrics, namely correlation coefficients and error residuals, give a certain indication of metric performance, they are probably not the best way to analyze this type of data. The time series data obtained in the SSCQE experiments and from the metrics have a relatively high auto-correlation, i.e. each sample is dependent on the previous and following samples. This problem of analyzing such data is not addressed in existing recommendations and standards. It also makes it difficult to separate the data into tuning and test sets in a meaningful fashion. As a possible solution, it has been proposed to subsample the time series data until they become independent, but this is not completely satisfactory. We are examining approaches that are better suited for the comparison of such time series data.

The other important point we would like to emphasize is the choice of metrics. No-reference quality metrics are certainly the most versatile, as they can be used in many situations where the reference is unavailable, or where an exact alignment with the original is not possible. As such, they can be implemented on the client side of streaming applications. If the output conditions are known, the quality metric can even be "tuned" to the specific conditions, and give feedback to the streaming server.

Still, no-reference quality assessment does have its problems, as we have pointed out in this paper as well. Any additional information about the "original" content can only improve the prediction performance of a quality metric (if the practical issues of making available well-aligned reference information is solved). In still image print applications, for example, some color re-rendering algorithms rely on camera specific metadata to improve the quality of their output, such as white-point, exposure, and color cast corrections. If the metadata is stripped away, mostly through opening and saving an image in an application that does not support metadata, the efficiency of these algorithms is decreased. In video applications, reduced-reference quality metrics have become quite popular and successful. Several metrics in the latest VQEG evaluation are based on this approach,[22] even though the tests were performed in a full-reference setting and the metrics had access to the uncompressed reference videos. Some of them are of relatively low complexity, achieve very good results and clearly outperform PSNR. In the end, the application at hand is probably the determining factor for choosing the "right" type of quality metric.

[*] This applies to printers that are not ICC color management compatible, and/or to image files that are not tagged with an ICC profile specifying a different encoding.

REFERENCES

1. ISO 22028-1:2003, "Photography and graphic technology – extended colour encodings for digital image storage, manipulation and interchange – Part 1: architecture and requirements." International Organization for Standardization, 2003.
2. IEC 61966 2-1:1999, "Multimedia systems and equipment – colour measurement and management – Part 2-1: Colour management – default RGB colour space – sRGB." International Electrotechnical Commission, 1999.
3. ANSI/I3A IT10.7666, "Electronic still picture imaging – reference output medium metric RGB color encoding ROMM-RGB." American National Standards Institute, 2002.
4. ICC.1:2001-12, "File format for color profiles, version 4.0.0." International Color Consortium, 2001.
5. J. Holm, "Photographic tone and colour reproduction goals," in *Proc. CIE Expert Symposium on Colour Standards for Image Technology*, pp. 51–56, (Vienna, Austria), March 25–27 1996.
6. G. Finlayson, S. Hordley, and P. Hubel, "Color by correlation: A simple unifying framework for color constancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23**(11), pp. 1209–1221, 2001.
7. E. Land and J. McCann, "Lightness and retinex theory," *Journal of the Optical Society of America* **61**(1), pp. 1–11, 1971.
8. D. S. Taubman and M. W. Marcellin, *JPEG2000: Image Compression Fundamentals, Standards and Practice*, Kluwer Academic Publishers, 2002.
9. M. Yuen and H. R. Wu, "A survey of hybrid MC/DPCM/DCT video coding distortions," *Signal Processing* **70**, pp. 247–278, Nov. 1998.
10. S. Winkler, *Vision Models and Quality Metrics for Image Processing Applications.* PhD thesis, École Polytechnique Fédérale de Lausanne, Switzerland, 2000.
11. S. Winkler, A. Sharma, and D. McNally, "Perceptual video quality and blockiness metrics for multimedia streaming applications," in *Proc. International Symposium on Wireless Personal Multimedia Communications*, pp. 547–552, (Aalborg, Denmark), Sept. 9–12, 2001.
12. Z. Wang, A. C. Bovik, and B. L. Evans, "Blind measurement of blocking artifacts in images," in *Proc. International Conference on Image Processing*, **3**, pp. 981–984, (Vancouver, Canada), Sept. 10–13, 2000.
13. P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "A no-reference perceptual blur metric," in *Proc. International Conference on Image Processing*, **3**, pp. 57–60, (Rochester, NY), Sept. 22–25, 2002.
14. D. Hasler and S. Süsstrunk, "Measuring colourfulness in natural images," in *Proc. SPIE/IS&T Human Vision and Electronic Imaging*, **5007**, pp. 87–95, (Santa Clara, CA), Jan. 21–24, 2003.
15. P. Engeldrum, *Psychometric Scaling: A Toolkit for Imaging Systems Development*, Imcotek Press, 2000.
16. ISO/DIS 20462-1, "Psychophysical experimental method to estimate image quality – Part 1: Overview of psychophysical elements." International Organization for Standardization, 2003.
17. ISO/DIS 20462-2, "Psychophysical experimental method to estimate image quality – Part 2: Triplet comparison method." International Organization for Standardization, 2003.
18. ISO/DIS 20462-3, "Psychophysical experimental method to estimate image quality – Part 3: Quality ruler method." International Organization for Standardization, 2003.
19. ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications." International Telecommunication Union, Geneva, Switzerland, 1996.
20. ITU-R Recommendation BT.500-11, "Methodology for the subjective assessment of the quality of television pictures." International Telecommunication Union, Geneva, Switzerland, 2002.
21. VQEG, "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment," April 2000. Available at http://www.vqeg.org/.
22. VQEG, "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment – Phase II," August 2003. Available at http://www.vqeg.org/.
23. P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "Perceptual blur and ringing metrics: Application to JPEG2000," *Signal Processing: Image Communication* **19**, Jan. 2004.
24. H. R. Sheikh, A. C. Bovik, L. Cormack, and Z. Wang, "LIVE image quality assessment database." http://live.ece.utexas.edu/research/quality, 2003.

25. S. Winkler and R. Campos, “Video quality evaluation for Internet streaming applications,” in *Proc. SPIE/IS&T Human Vision and Electronic Imaging*, **5007**, pp. 104–115, (Santa Clara, CA), Jan. 21–24, 2003.
26. S. Winkler and F. Dufaux, “Video quality evaluation for mobile applications,” in *Proc. SPIE/IS&T Visual Communications and Image Processing*, **5150**, pp. 503–693, (Lugano, Switzerland), July 8–11, 2003.
27. T. Alpert, V. Baroncini, D. Choi, L. Contin, R. Koenen, F. Pereira, and H. Peterson, “Subjective evaluation of MPEG-4 video codec proposals: Methodological approach and test procedures,” *Signal Processing: Image Communication* **9**, pp. 305–325, May 1997.
28. T. Ebrahimi and F. Pereira, *The MPEG-4 Book*, Prentice Hall, 2002.
29. MPEG, “AHG report on editorial convergence of MPEG-4 reference software,” Tech. Rep. M8041, ISO/IEC JTC1/SC29/WG11, 2002.
30. ITU-T Recommendation H.223, “Multiplexing protocol for low bit rate multimedia communication.” International Telecommunication Union, Geneva, Switzerland, 2001.
31. ITU-T Study Group 16, “WCDMA error patterns at 128 and 384 kbps,” Tech. Rep. Q15-G28, ITU, 1999.
32. 3GPP, “QoS for speech and multimedia codec: Quantitative performance evaluation of H.324 Annex C over 3G,” Tech. Rep. 26.912 (Release 4), 3GPP, 2001.
33. ISO 3664:2000, “Viewing conditions – Graphic technology and photography.” International Organization for Standardization, 2000.
34. D. H. Kelly, “Motion and vision. II. Stabilized spatio-temporal threshold surface,” *Journal of the Optical Society of America* **69**, pp. 1340–1349, Oct. 1979.
35. C. A. Burbeck and D. H. Kelly, “Spatiotemporal characteristics of visual mechanisms: Excitatory-inhibitory model,” *Journal of the Optical Society of America* **70**, pp. 1121–1126, Sept. 1980.
36. D. H. Kelly, “Spatiotemporal variation of chromatic and achromatic contrast thresholds,” *Journal of the Optical Society of America* **73**, pp. 742–750, June 1983.
37. ITU-R Recommendation BT.709-5, “Parameter values for the HDTV standards for production and international programme exchange.” International Telecommunication Union, Geneva, Switzerland, 2002.

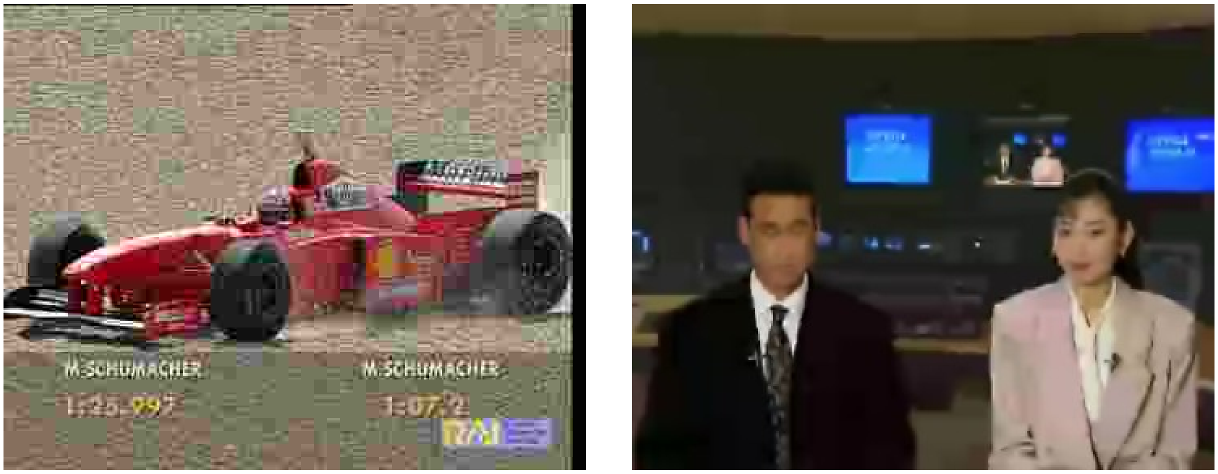

Figure 1. Examples of artifacts commonly introduced by image or video compression: blockiness (left) and blurriness (right).

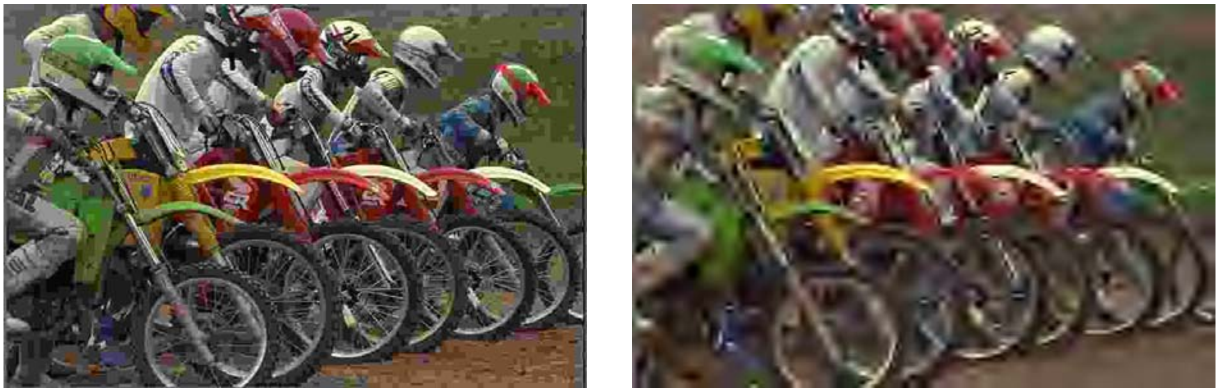

Figure 4. Examples for color bleeding and loss of colorfulness due to JPEG (left) and JPEG2000 (right) encoding.

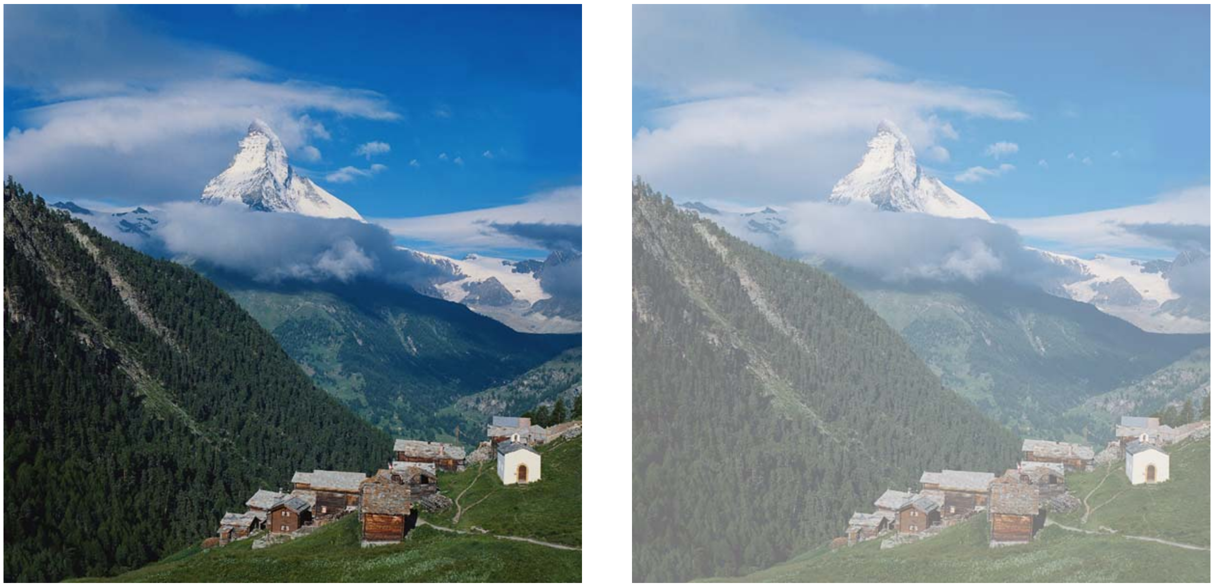

Figure 6. Simulation of an image without veiling glare (left) and with 4% glare of the adapted white-point luminance (right).