Quality Metric Design: A Closer Look

Stefan Winkler

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

ABSTRACT

The design of reliable visual quality metrics is complicated by our limited knowledge of the human visual system and the resulting variety of pertinent vision models. We have begun to analyze and compare a number of implementation choices for some components found in most of today's visual quality metrics that are based on a model of human vision and present the first results here.

Keywords: Vision model evaluation, video quality assessment, perceptual distortion metric

1. INTRODUCTION

The introduction of digital TV systems on the consumer market has lead to a rising demand for pertinent measurement tools, not only on the bitstream and network level, but also for video quality assessment on a perceptual level. Consequently, considerable effort has been put in 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[14]). As a matter of fact, there are so many different implementations described in the literature today, each one claiming performance advantages over the others, that it is difficult to determine which one actually is the most promising design.

Some investigations have already attempted to conduct an impartial comparison of selected visual quality metrics. Most notably, the Video Quality Experts Group (VQEG)* has collected subjective ratings for a large set of sequences and has evaluated the performance of different video quality assessment systems with respect to these sequences. Its final report is being prepared at the time of writing.[12] First performance results for the author's submission to the VQEG test, the perceptual distortion metric (PDM),[15] are presented in this paper.

Nevertheless, this and other comparative studies have focused on evaluating the performance of entire systems. Hardly any analyses of single components of visual quality metrics have been published. Such an evaluation, which we think is important in order to achieve further improvements in this domain, is the purpose of this paper. We have begun to analyze a number of implementation choices for some components found in most of today's quality assessment systems that are based on a vision model and present the first results here. These different implementations are equivalent from the point of view of simple threshold experiments, but can produce differing results for complex test sequences.

The paper is structured as follows: Section 2 gives a brief overview of the experiments carried out in the framework of VQEG, as we use the data obtained there in the subsequent analyses. Then the perceptual distortion metric (PDM),[15] which has been developed by the author and is one of the metrics in the VQEG test, is described and its performance is discussed in section 3. Based on this metric, a number of different color spaces and some common pooling algorithms are compared in sections 4 and 5, respectively. Finally, areas for future work are outlined.

* See http://www.crc.ca/vqeg/ for an overview of its activities.

2. EXPERIMENTS

Subjective experiments are necessary in order to evaluate models of human vision. Similarly, subjective ratings form the benchmark for visual quality metrics. In this paper, experimental data collected by the Video Quality Experts Group (VQEG) is used. VQEG was formed in 1997 with the objective to collect reliable subjective ratings for a well-defined set of test sequences and to evaluate the performance of different video quality assessment systems with respect to these sequences.[12] In this section, the sequence characteristics and the testing methodology are described.

The emphasis of the first phase of VQEG has been out-of-service testing (the full reference sequence is available to the metrics) of production- and distribution-class video. Therefore, the test conditions comprise mainly MPEG-2 encoded sequences with different profiles, levels and other parameter variations, including encoder concatenation, conversions between analog and digital video, and transmission errors. A set of 8-second scenes with different characteristics (e.g. spatial detail, color, motion) was selected by independent labs; the scenes were disclosed to the proponents only after the submission of their metrics. In total, 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.[12]

Formal subjective testing is described in ITU-R Recommendation 500,[3] which suggests standard viewing conditions, criteria for observer selection, assessment procedures, and analysis methods. In the VQEG tests, the Double Stimulus Continuous Quality Scale (DSCQS) from this recommendation was used for the subjective experiments.

The presentation sequence for a DSCQS trial is illustrated in Figure 1. Viewers are shown multiple sequence pairs consisting of a "reference" and a "test" sequence, which are rather short (8 seconds in the VQEG experiments). The reference and test sequence are presented twice in alternating fashion, with the order of the two chosen randomly for each trial. Subjects are not informed which is the reference and which is the test sequence. They rate each of the two separately on a continuous quality scale ranging from "bad" to "excellent" as shown in Figure 1. Analysis is based on the difference in rating for each pair, which is calculated from an equivalent numerical scale from 0 to 100.

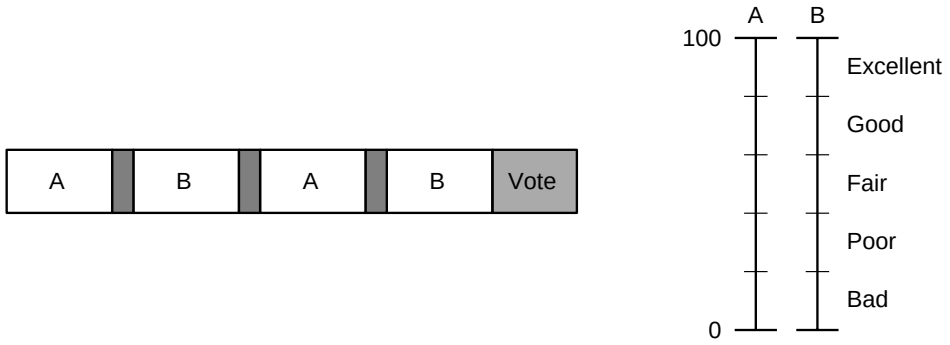

Figure 1. Presentation sequence and rating scale for the DSCQS method.

The subjective experiments were carried out in eight different laboratories. Half of the labs ran the tests with 50 Hz sequences, and the other half with 60 Hz sequences. Furthermore, each lab ran two separate tests for low-quality and high-quality sequences. A total of 297 non-expert viewers participated in the experiments, and over 26000 individual ratings were recorded. Post-screening of the subjective data was performed in accordance with ITU-R Rec. 500 in order to discard incomplete and unstable viewers. The mean subjective ratings over the screened viewers are used in this paper. For a detailed discussion of the subjective experiments and their results the reader is referred to the VQEG report.[12]

3. PERCEPTUAL DISTORTION METRIC

The perceptual distortion metric (PDM) has been developed by the author.[15] It is one of the ten metrics submitted to the VQEG test for evaluation. In this section, the vision model used in the PDM is reviewed, and its performance for the VQEG sequences is discussed.

3.1. Vision Model

The vision model of the perceptual distortion metric works as follows: After conversion to opponent-colors space, each of the resulting three components is subjected to a perceptual decomposition, yielding several perceptual channels. They undergo weighting and a contrast gain control stage. Finally, all the sensor differences are combined into a distortion measure. A block diagram of the PDM and the underlying vision model is shown in Figure 2.

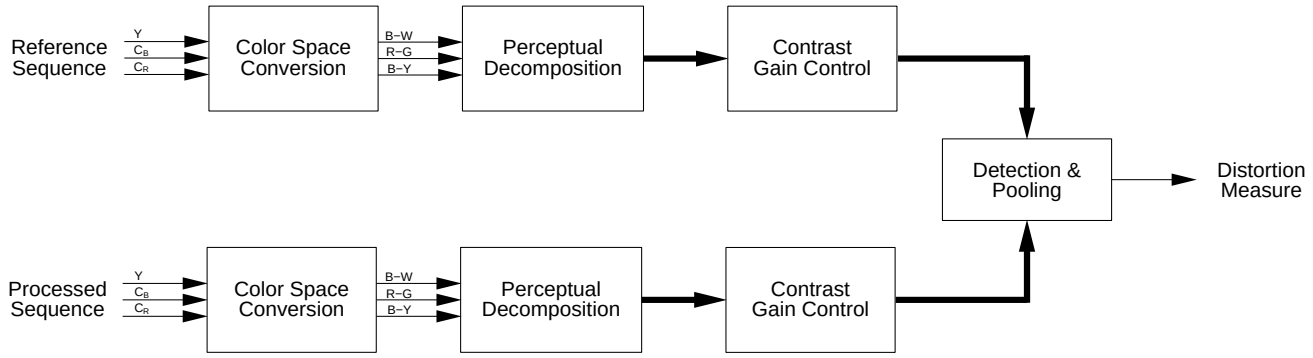

Figure 2. Block diagram of the perceptual distortion metric.

The input is first converted to an opponent-colors space, comprising black-white (B-W), red-green (R-G), and blue-yellow (B-Y) difference signals. The specific opponent-colors space used in the PDM was derived by Poirson and Wandell[4,5] and separates the effects of color perception and pattern sensitivity. This module will be analyzed in more detail in section 4.

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 filters used in the PDM are based on work by Fredericksen and Hess.[2] The temporal mechanisms in the human visual system are modeled with one low-pass and one band-pass mechanism, whose frequency responses are shown in Figure 3. For use in the metric, corresponding filters were designed with the constraint to keep the delay to a minimum, which is crucial in certain applications of visual quality assessment. Therefore the mechanisms are approximated with two recursive infinite impulse response (IIR) filters, which have been found to yield the shortest delay while still maintaining a good approximation of the frequency responses as shown in Figure 3. In the present implementation, the low-pass filters are applied to all three color channels, but the band-pass filter is applied only to the luminance channel. This simplification is based on the fact that color contrast sensitivity is rather low for higher frequencies.

The decomposition in the spatial domain is carried out by means of the steerable pyramid transform† proposed by Simoncelli et al.[8] This transform decomposes an image into a number of spatial frequency and orientation bands; its basis functions are directional derivative operators (the frequency response of one such filter is shown in Figure 4). For use within a vision model, it has the advantage of being rotation-invariant and self-inverting, and it minimizes the amount of aliasing in the subbands. 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, and the same decomposition is used for all three color channels.

After the temporal and spatial decomposition, each channel is weighted in such a way that the sum of all channels approximates the spatio-temporal contrast sensitivity function of the human visual system.

The contrast gain control stage in the PDM implements pattern sensitivity and contrast masking. Contrast gain control models were 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. Contrast gain control can be realized by an excitatory nonlinearity that is inhibited divisively by a pool of responses from other neurons. Thus masking occurs through the inhibitory effect of the normalizing pool. In the PDM, we rely on a generalized contrast gain control model by Watson and Solomon,[13] which facilitates the integration of many kinds of channel interactions.

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

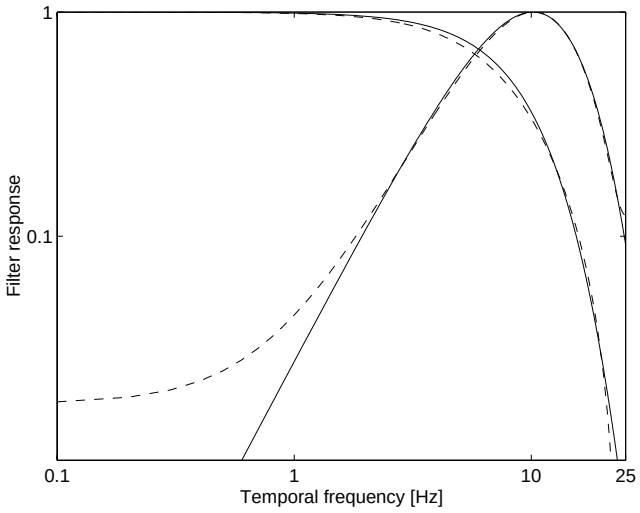

Figure 3. Frequency responses of sustained (low-pass) and transient (band-pass) mechanisms of vision[2] (solid) and their IIR filter approximations for a sampling frequency of 50 Hz (dashed).

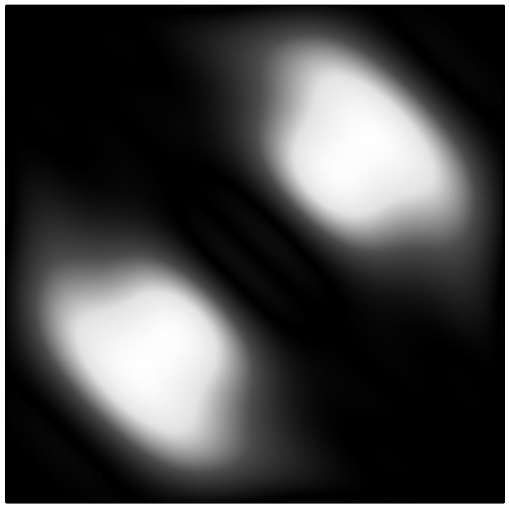

Figure 4. Frequency response of a filter used in the steerable pyramid transform.[8] In the PDM, these filters are applied at four different levels and four different orientations.

Finally, all the sensor differences are combined into a distortion measure according to rules of probability or vector summation, also known as pooling.[7] In principle, any subset of dimensions can be used for this summation, depending on what kind of result is desired. We will take a closer look at possible pooling methods in section 5.

3.2. Performance

There are a number of attributes that characterize a visual quality metric in terms of its estimation performance with respect to the subjective ratings. These attributes are accuracy, monotonicity, and consistency. Accuracy is the ability of a metric to predict subjective ratings with minimum average error and can be determined by means of the Pearson linear correlation coefficient; for a set of N data pairs (x_i, y_i), it is defined as follows:

$$r_P = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2}\sqrt{\sum(y_i - \overline{y})^2}},$$

where $\overline{x}$ and $\overline{y}$ are the means of the respective data sets.

Monotonicity is another important attribute as it measures if increases in one variable are associated with increases in the other variable. Ideally, changes of a metric's rating between different sequences should always have the same sign as the changes of the corresponding subjective ratings. The degree of monotonicity can be quantified by the Spearman rank-order correlation coefficient, which is defined as follows:

$$r_S = \frac{\sum(\chi_i - \overline{\chi})(\gamma_i - \overline{\gamma})}{\sqrt{\sum(\chi_i - \overline{\chi})^2}\sqrt{\sum(\gamma_i - \overline{\gamma})^2}} = 1 - \frac{6(\chi_i - \gamma_i)^2}{N\,(N^2 - 1)},$$

where χ_i is the rank of x_i and γ_i is the rank of y_i, and $\overline{\chi}$ and $\overline{\gamma}$ are the respective midranks. The Spearman rank-order correlation is non-parametric, i.e. it makes no assumptions about the form of the relationship between the x_i and y_i.

Considering these attributes, the PDM performs well over all test cases (see Figure 5). Several of its outliers belong to the lowest-bitrate (H.263) sequences of the test. As the metric is based on a threshold model of human vision, performance degradations for such clearly visible distortions can be expected. A number of other outliers are due to a single 50 Hz-scene with a lot of movement. They may be due to inaccuracies in the temporal filtering of

the submitted version, which is being investigated at present. These sequences will be disregarded in the analyses below in order to minimize the influence of the outliers.

Figure 6 shows the correlations between PDM and subjective ratings over all sequences and for a number of subsets of test conditions, namely the 50 and 60 Hz sets, the low- and high-quality sets as defined for the subjective experiments, the H.263 and non-H.263 sequences, the sequences with and without transmission errors, as well as the MPEG and non-MPEG sequences. As can be seen, the PDM can handle MPEG as well as non-MPEG kinds of distortions and also behaves well with respect to sequences with transmission errors. The Pearson linear[‡] and the Spearman rank-order correlation coefficients for most of the subsets are around 0.8.

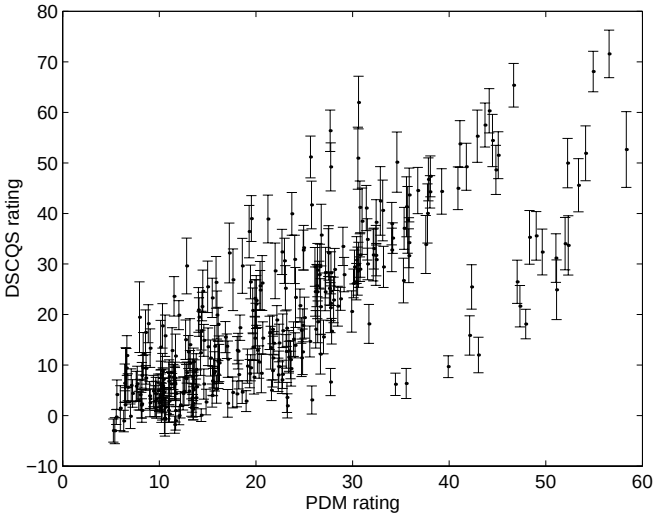

Figure 5. Scatter plot of PDM vs. mean DSCQS ratings. The error bars indicate the 95% confidence intervals of the subjective ratings.

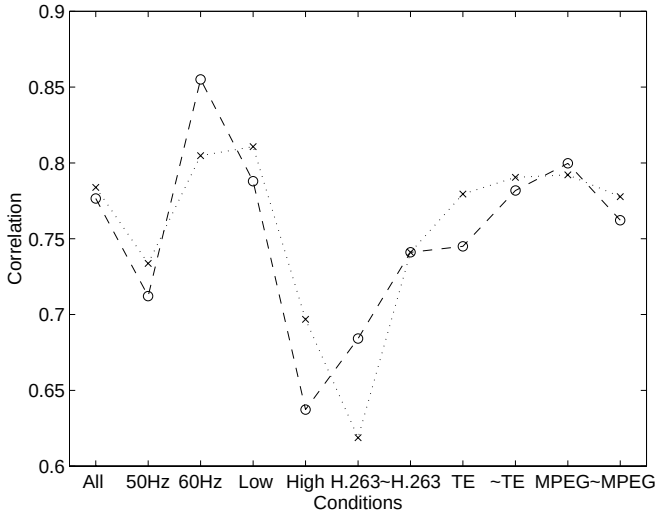

Figure 6. Pearson linear (circles) and Spearman rank-order (crosses) correlations for several subsets of test conditions in the VQEG test (see text).

4. COLOR SPACE

The PDM uses an opponent-colors space by Poirson and Wandell that was designed to separate color perception from pattern sensitivity,[4,5] which has been considered an advantage for the modular design of the metric. However, this color space is based on color-matching experiments and not on human perception of color differences, which is what a visual quality metric is really supposed to evaluate. Color spaces such as CIE $L^*a^*b^*$ and CIE $L^*u^*v^*$ on the other hand, which have been used successfully in other metrics, were designed for color difference measurements, but lack pattern-color separability. Even simple YC_BC_R implements the opponent-color idea (Y encodes luminance, C_B the difference between blue primary and luminance, and C_R the difference between red primary and luminance) and provides the advantage of requiring no conversions from the digital component video input material (see e.g. Poynton[6] for details about this color space).

In addition to evaluating the different color spaces, we also compare the full-color version of each implementation with its luminance-only version. Before analyzing these color spaces with respect to the achievable rating accuracy, however, let us briefly review the definitions of CIE $L^*a^*b^*$ and CIE $L^*u^*v^*$.

Conversion from CIE 1931 XYZ tristimulus values to CIE $L^*a^*b^*$ and CIE $L^*u^*v^*$ color spaces is defined as follows.[17] The conversions make use of the function

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

[‡] Note that only the direct linear correlations have been calculated here and elsewhere in this paper; when using a mapping function to account for nonlinearities in the scatter plot as in the official VQEG analysis, the correlations are higher. The relations between them generally remain the same, though.

Both CIE $L^*a^*b^*$ and CIE $L^*u^*v^*$ space share a common lightness component $L^* = 116f(Y/Y_0) - 16$. The chromaticity coordinates are defined as $a^* = 500\,[f(X/X_0) - f(Y/Y_0)]$, $b^* = 200\,[f(Y/Y_0) - f(Z/Z_0)]$ and $u^* = 13L^*(u' - u'_0)$, $v^* = 13L^*(v' - v'_0)$, where $u' = 4X/(X + 15Y + 3Z)$ and $v' = 9Y/(X + 15Y + 3Z)$ (the 0-subscripts refer to the corresponding units for the reference white $X_0Y_0Z_0$ being used). The total perceptual color differences are given by the squared sum of the respective component differences: $\Delta E^*_{Lab} = \sqrt{(\Delta L^*)^2 + (\Delta a^*)^2 + (\Delta b^*)^2}$ and $\Delta E^*_{Luv} = \sqrt{(\Delta L^*)^2 + (\Delta u^*)^2 + (\Delta v^*)^2}$.

The above-mentioned color spaces are similar in that they are all based on color differences. Therefore, they can be used interchangeably in the PDM by doing the respective color space conversion in the first module and ensuring that the threshold behavior of the metric does not change.

The results are shown in Figure 7. As can be seen, the differences in correlation are quite significant. Common to all color spaces is the fact that the additional consideration of the color components leads to a performance increase over the luminance-only version, although this improvement is not very large. In fact, the slight increases may not justify the double computational load imposed by the full-color PDM. However, one has to bear in mind that under most circumstances video encoders are good-natured and distribute distortions more or less equally between the three color channels, therefore a result like this can be expected. Certain conditions with high color saturation or unusually large distortions in the color channels may well be overlooked by a simple luminance metric, though.

Component video YC_BC_R exhibits the worst performance of the group. This is unfortunate, because it is the color space of the digital video input, so no further conversion is required. However, the conversions from YC_BC_R to the other color spaces incur only a relatively small penalty on the total computation time (on the order of a few percent) despite the nonlinearities involved. Furthermore, it is interesting to note that both CIE $L^*a^*b^*$ and CIE $L^*u^*v^*$ slightly outperform the Poirson-Wandell opponent-colors space in the PDM. This may be due to the better incorporation of perceived lightness in these color spaces.

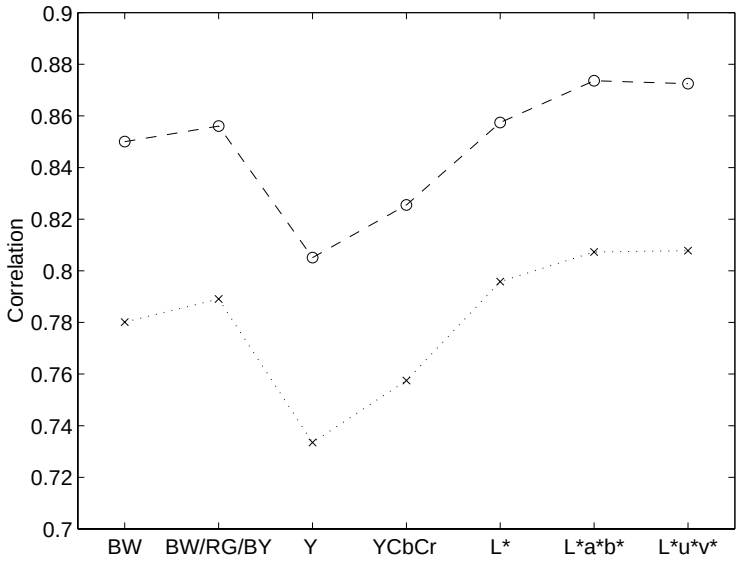

Figure 7. Pearson linear (circles) and Spearman rank-order (crosses) correlations for different color spaces. The differences in correlation are quite significant.

5. POOLING ALGORITHM

It is believed that the information represented in various channels of the primary visual cortex is integrated in higher-level areas of the brain. This process can be simulated by gathering the data from these channels according to rules of probability or vector summation, also known as pooling.[7] However, little is known about the nature of the actual integration in the brain, and pooling mechanisms remain one of the most debated and uncertain aspects of vision modeling.

The combination of N mechanism responses x_i can be computed by means of vector summation (also known as Minkowski summation):

$$x = \sqrt[\beta]{\frac{1}{N}\sum x_i^\beta}.$$

Different exponents β have been found to yield good results for different experiments and implementations. $\beta = 2$ corresponds to the ideal observer formalism under independent Gaussian noise, which assumes that the observer has complete knowledge of the stimuli and uses a matched filter for detection.[9] In a study of subjective experiments with coding artifacts, $\beta \approx 2$ was found to give good results.[1] Intuitively, a few high distortions may draw the viewer's attention more than many lower ones. This behavior can be emphasized with higher exponents, which have been used in several other vision models, for example $\beta = 4$.[10,11] The best fit of a contrast gain control model to masking data was achieved with $\beta \approx 5$.[13]

In the PDM, pooling over channels and pixel locations is carried out with $\beta = 2$, whereas $\beta = 4$ is used for pooling over frames. We take a closer look at the latter part here. First, the temporal pooling exponent is varied between 0.1 and 6, and the correlations of PDM and subjective ratings are computed. As can be seen from Figure 8, the maximum Pearson correlation is obtained at $\beta = 2.9$, and the maximum Spearman correlation at $\beta = 2.2$. However, neither of the two peaks is very distinct. This result may be explained by the fact that for a large majority of the test sequences, the distortions are distributed quite uniformly over time, so that the overall ratings computed with $\beta = 0.1$ and $\beta = 6$ differ by less than 15%.

As an alternative, the distribution of ratings over frames can be used statistically to derive an overall rating. A simple method is to take the distortion rating that separates the lowest 80% of frame ratings from the highest 20%, for example. It can be argued that such a procedure emphasizes high distortions which are annoying to the viewer no matter how good the quality of the rest of the sequence is. Again, however, the specific histogram threshold chosen is rather arbitrary. Figure 9 shows the correlations computed for different values of this threshold. Here the influence is much more pronounced; the maximum Pearson correlation is obtained for thresholds between 55% and 75%, and the maximum Spearman correlation for thresholds between 45% and 65%, leading to the conclusion that a threshold of around 60% is the best choice overall for this method.

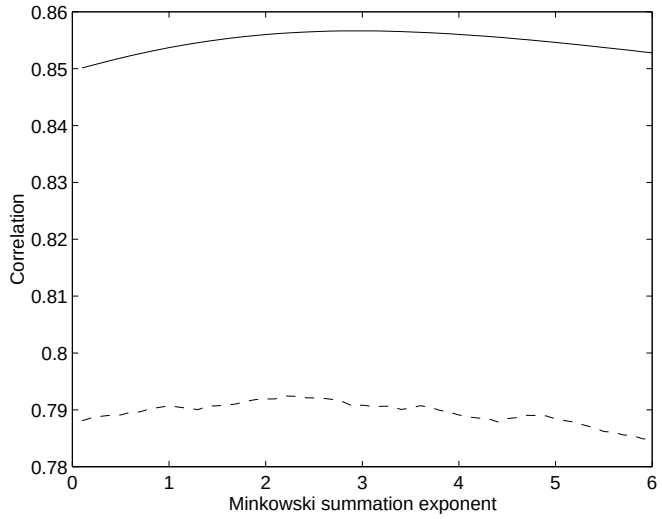

Figure 8. Pearson linear (solid) and Spearman rank-order (dashed) correlation vs. Minkowski summation exponent β. Both curves have their maxima around $\beta \approx 2.5$, but they are not very distinct.

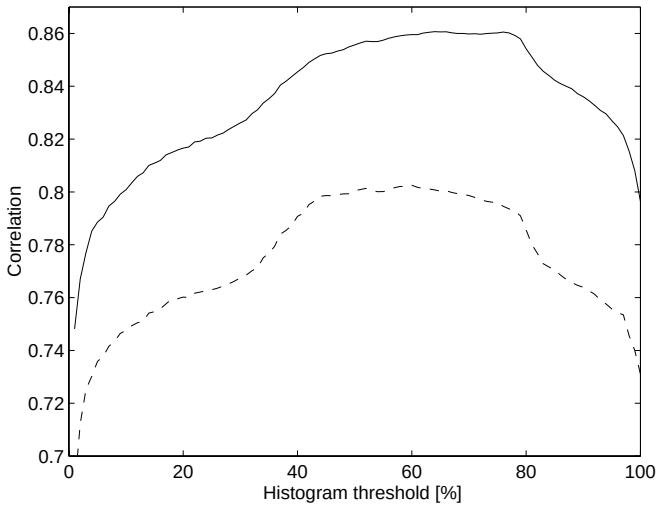

Figure 9. Pearson linear (solid) and Spearman rank-order (dashed) correlation vs. histogram threshold. Best overall results are obtained at thresholds around 60%.

In any case, the pooling operation need not be carried out over all pixels in the entire sequence or frame. In order to take into account the focus of attention of observers, for example, pooling can be carried out separately for

spatio-temporal blocks of the sequence that cover roughly 100 milliseconds and two degrees of visual angle each.[11] Alternatively, the distortion can be computed locally for every pixel, yielding perceptual distortion maps for better visualization of the temporal and spatial distribution of distortions. Such a distortion map can help the expert to locate and identify problems in the processing chain or shortcomings of an encoder, for example. This can be more useful and more reliable than a global measure in many quality assessment applications.

6. CONCLUSIONS AND FUTURE WORK

The PDM is a reliable measurement tool for perceptual video quality and performs well over a wide range of test sequences. The results presented above also show that visual quality metrics which are essentially equivalent at the threshold level can exhibit significant performance differences for complex sequences depending on the implementation choices made for the color space and the pooling algorithm used in the underlying vision model.

Within the scope of this paper, only a small number of components could be investigated. Future research will focus on other parts of the vision model. One of them is the perceptual decomposition; many different filters have been proposed as approximations to the decomposition of visual information taking place in the human visual system. The PDM uses the steerable pyramid; other options include Gabor filters, the Cortex transform, and wavelets. We have found that the exact shape of the filters is not of paramount importance, but the goal here is also to obtain a good tradeoff between implementation complexity, flexibility, and correlation with subjective data.

Another component of interest is masking. Contrast gain control models such as the one used in the PDM have recently increased in popularity. However, these models can be rather awkward to use in the general case, because they require a computation-intensive parameter fit for every change in the setup. Simpler models such as the so-called nonlinear transducer model are often more "user-friendly", but are also less powerful.

Finally, the computation of contrast should be investigated. Contrast is relatively easy to define verbally, but for complex stimuli a multitude of mathematical contrast definitions have been proposed. Results show that a local measure of contrast is important for complex stimuli, but which filter combination should be used to compute it? We have recently looked into some of these issues[16] and intend to carry out an evaluation with the PDM.

REFERENCES

1. H. de Ridder: "Minkowski-metrics as a combination rule for digital-image-coding impairments." in *Proc. SPIE*, vol. 1666, pp. 16–26, San Jose, CA, 1992.
2. R. E. Fredericksen, R. F. Hess: "Estimating multiple temporal mechanisms in human vision." *Vision Res.* **38**(7), 1023–1040, 1998.
3. ITU-R Recommendation BT.500-8: "Methodology for the subjective assessment of the quality of television pictures." ITU, Geneva, Switzerland, 1998.
4. A. B. Poirson, B. A. Wandell: "Appearance of colored patterns: Pattern-color separability." *J. Opt. Soc. Am. A* **10**(12), 2458–2470, 1993.
5. A. B. Poirson, B. A. Wandell: "Pattern-color separable pathways predict sensitivity to simple colored patterns." *Vision Res.* **36**(4), 515–526, 1996.
6. C. A. Poynton: *A Technical Introduction to Digital Video*. John Wiley & Sons, 1996.
7. R. F. Quick, Jr.: "A vector-magnitude model of contrast detection." *Kybernetik* **16**, 65–67, 1974.
8. E. P. Simoncelli et al.: "Shiftable multi-scale transforms." *IEEE Trans. Information Theory* **38**(2), 587–607, 1992.
9. P. C. Teo, D. J. Heeger: "Perceptual image distortion." in *Proc. SPIE*, vol. 2179, pp. 127–141, San Jose, CA, 1994.
10. C. J. van den Branden Lambrecht: "Color moving pictures quality metric." in *Proc. ICIP*, vol. 1, pp. 885–888, Lausanne, Switzerland, 1996.
11. C. J. van den Branden Lambrecht, O. Verscheure: "Perceptual quality measure using a spatio-temporal model of the human visual system." in *Proc. SPIE*, vol. 2668, pp. 450–461, San Jose, CA, 1996.
12. Video Quality Experts Group: "Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment." VQEG, 2000, available at ftp://ftp.its.bldrdoc.gov/dist/ituvidq/.
13. 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.
14. S. Winkler: "Issues in vision modeling for perceptual video quality assessment." *Signal Processing* **78**(2), 231–252, 1999.
15. S. Winkler: "A perceptual distortion metric for digital color video." in *Proc. SPIE*, vol. 3644, pp. 175–184, San Jose, CA, 1999.
16. S. Winkler, P. Vandergheynst: "Computing isotropic local contrast from oriented pyramid decompositions." in *Proc. ICIP*, vol. 4, pp. 420–424, Kyoto, Japan, 1999.
17. G. Wyszecki, W. S. Stiles: *Color Science: Concepts and Methods, Quantitative Data and Formulae*. John Wiley & Sons, 2nd edn., 1982.