Design and performance of a digital video quality metric

Andrew B. Watson, James Hu, John F McGowan III, Jeffrey B. Mulligan

NASA Ames Research Center, Moffett Field, CA 94035

ABSTRACT

The growth of digital video has given rise to a need for computational methods for evaluating the visual quality of digital video. We have developed a new digital video quality metric, which we call DVQ (Digital Video Quality)[1].Here we provide a brief description of the metric, and give a preliminary report on its performance. DVQ accepts a pair of digital video sequences, and computes a measure of the magnitude of the visible difference between them. The metric is based on the Discrete Cosine Transform. It incorporates aspects of early visual processing, including light adaptation, luminance and chromatic channels, spatial and temporal filtering, spatial frequency channels, contrast masking, and probability summation. It also includes primitive dynamics of light adaptation and contrast masking. We have applied the metric to digital video sequences corrupted by various typical compression artifacts, and compared the results to quality ratings made by human observers.

1. INTRODUCTION

The emerging infrastructure for digital video lacks a critical component: a reliable means for automatically measuring visual quality. Such a means is essential for evaluation of codecs, for monitoring broadcast transmissions, and for ensuring the most efficient compression of sources and utilization of communication bandwidths. In a previous paper[1], we gave a preliminary description of a new video quality metric, which we called DVQ. In this report, we provide a brief review of the design of the metric, and then describe application of this metric to a set of video materials for which human subjective ratings are available. This provides a test of whether the metric can accurately predict human subjective ratings.

2. DVQ

All video quality metrics are inherently models of human vision. For example, if root-mean-squared-error (RMSE) is used as a quality metric, this amounts to the assumption that the human observer is sensitive to the summed squared deviations between reference and test sequences, and is insensitive to aspects such as the spatial frequency of the deviations, their temporal frequency, or their color. The DVQ (Digital Video Quality) metric is an attempt to incorporate many aspects of human visual sensitivity in a simple image processing algorithm. Simplicity is an important goal, since one would like the metric to run in real-time and require only modest computational resources. One of the most complex and time consuming elements of other proposed metrics[2, 3, 4, 5, 6] are the spatial filtering operations employed to implement the multiple, bandpass spatial filters that are characteristic of human vision. We accelerate this step by using the Discrete Cosine Transform (DCT) for this decomposition into spatial channels. This provides a powerful advantage since efficient hardware and software are available for this transformation, and because in many applications the transform may have already been done as part of the compression process.

Figure 1 is an overview of the processing steps of the DVQ metric. These steps are described in greater detail elsewhere[1], here we provide only a brief review. The input to the metric is a pair of color image sequences: reference (R), and test (T). The first step consists of various sampling, cropping, and color transformations that serve to restrict processing to a region of interest and to express the sequences in a perceptual color space. This stage also deals with de-interlacing and de-gamma-correcting the input video. The sequences are then subjected to a blocking (BLK) and a Discrete Cosine Transform (DCT), and the results are then transformed to local contrast (LC). Local contrast is the ratio of DCT amplitude to DC amplitude for the corresponding block. The next step is a temporal filtering operation (TF) which implements the temporal part of the contrast sensitivity function. This is accomplished through a suitable recursive discrete second order filter. The results are then converted to just-noticeable differences by dividing each DCT coefficient by its respective visual threshold . This implements the spatial part of the contrast sensitivity function (CSF). At the next stage the two sequences are subtracted. The difference sequence is then subjected to a contrast masking operation (CM), which also depends upon the reference sequence. Finally the masked differences may be pooled in various ways to illustrate the perceptual error over various dimensions (POOL), and the pooled error may be converted to visual quality (VQ).

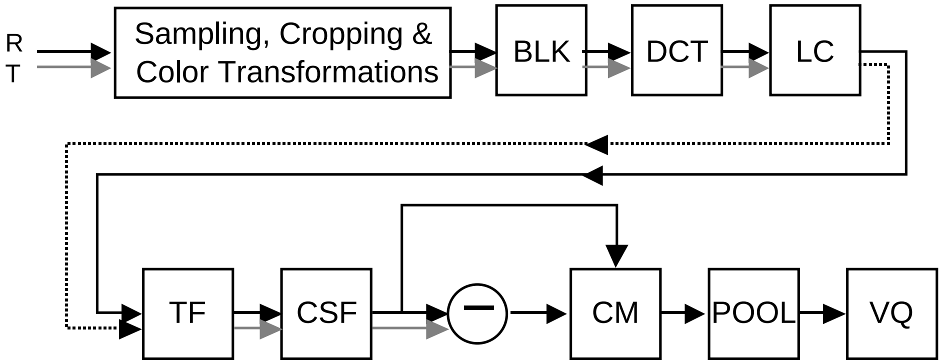


Figure 1. Overview of DVQ processing steps.

The parameters of the metric have been estimated from psychophysical data, both from the existing literature and from measurements of visibility of dynamic DCT quantization error made by the author[1].

3. TEST MATERIALS

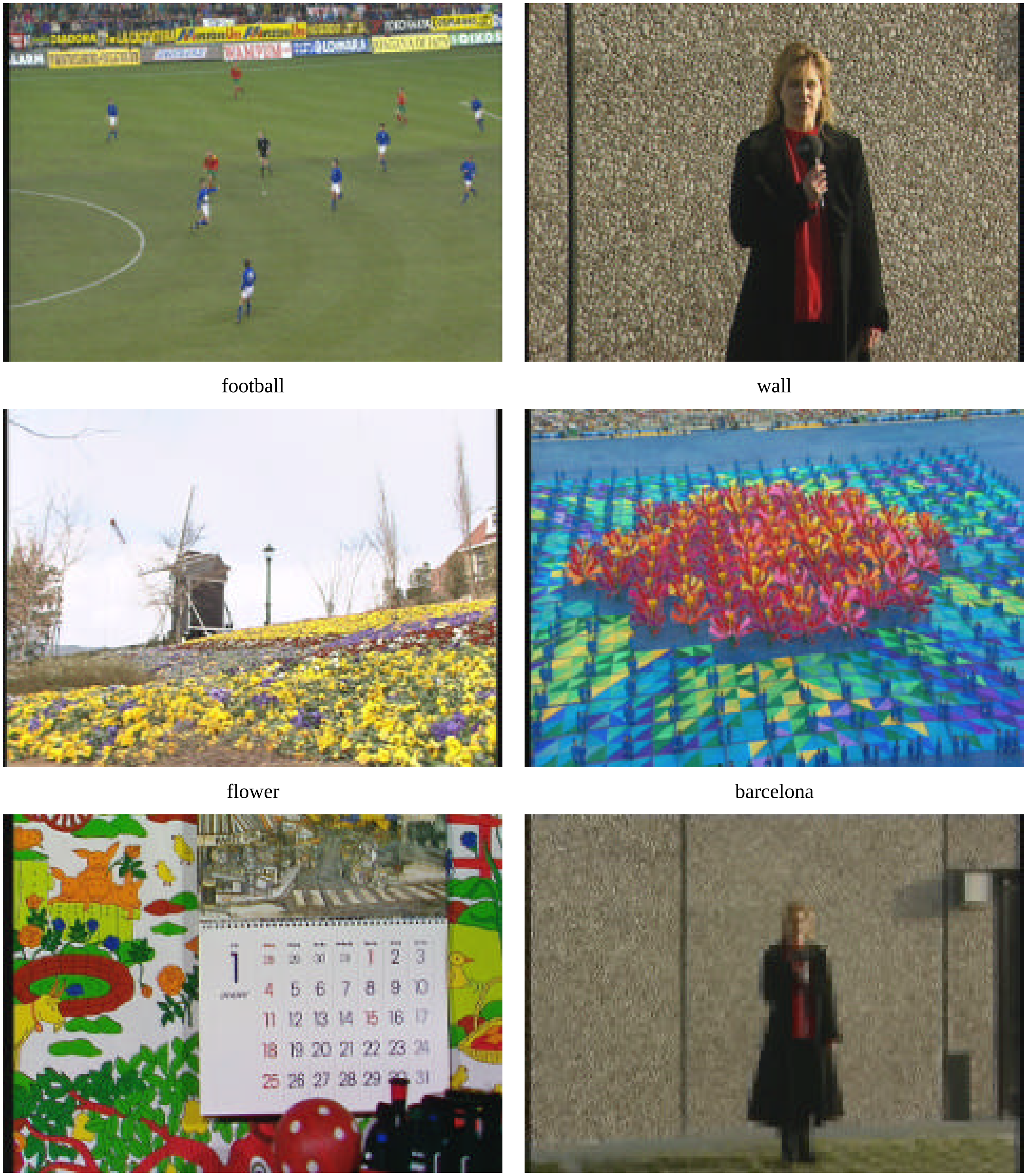
To test the DVQ metric we need to compare its results for specific video sequences to the quality estimates provided by human observers. There are few, if any, publicly available data sets of this kind. Most such materials are developed in commercial settings and subject to proprietary concerns. We were able to secure one data set, consisting of digital image sequences and associated subjective ratings. The data set is described more extensively elsewhere[7, 8]; here we provide only summary details regarding the video sequences and subjective data.

1. Video Sequences

The video materials consisted of a total of 65 sequences, of which 5 were reference sequences and 60 were processed sequences, obtained by passing the 5 reference sequences through 12 hypothetical reference circuits (HRCs) that will be described below. Each sequence was in ITU-601 PAL format[9] (576x720, interlaced, 4:2:2 sampling), and was 9 seconds (225 frames) in duration.

2. Reference Sequences

The five reference sequences were selected to span a wide spectrum of typical video, and to emphasize various challenges to video compression, such as saturated color, panning, rapid movement, fine detail, etc. The first frame of each of the five reference sequences is shown in Figure 2, along with one frame from a processed sequence.



football

wall

flower

barcelona

mobile

wall (HRC 1)

Figure 2. The first five images show the first frames of the reference sequences. The last image shows one frame from HRC 1 applied to "wall".

3. Hypothetical Reference Circuits

Processed sequences were produced by passing the five reference sequences through each of 13 hypothetical reference circuits (HRCs). Each HRC consisted of a particular combination of MPEG codec, bit-rate, and possibly conversion from digital to analog (PAL) and back again to digital prior to encoding. The identities of the particular codecs are not important for present purposes, so we assign them the arbitrary labels A and B. An HRC consisting of no processing is also included. The thirteen HRC's are defined in Table 1.

HRC	Codec	Bit-rate	PAL processing
1	A	2	
2	A	3	
3	A	4.5	
4	A	7	
5	A	10	
6	B	2	
7	B	3	
8	B	4.5	
9	B	7	
10	B	10	
11	A	3	yes
12	B	3	yes
13	none		

Table 1. HRC definitions.

4. Subjective data.

The subjective data were obtained from 25 observers using the Double Stimulus Continuous Quality Scale (DSCQS)[10]. In this method on each trial the observer views in succession a processed (test) sequence and the corresponding reference sequence, and assigns to each a quality rating between 0 and 100. Here we examine the difference in the scores for reference and test, which we will call the *impairment* score.

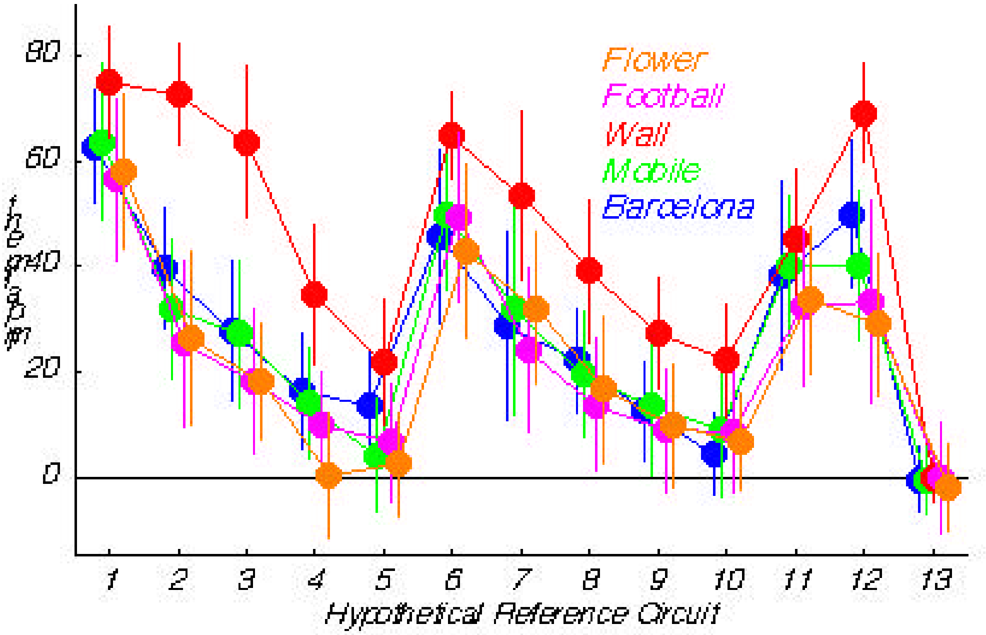


Figure 3. Impairment scores for five reference sequences processed by thirteen HRC's. Means and ± one standard deviation are shown

In Figure 3 we plot the mean impairment scores for each condition (combination of source and HRC). Each point is the mean of a single trial from each of 25 observers. The general pattern of results is similar for all five sources, though the “wall” sequence yields generally higher levels of impairment than the others. As expected, impairment declines with bit-rate, approaching zero at a bit-rate of 10 Mbits/sec for both codecs. Note the considerable variability among observers, which places a limit on the predictability of the results.

4. DVQ PREDICTIONS

The DVQ metric was prototyped in the Mathematica programming language, and subsequently implemented in c. Predictions shown here were computed on a Silicon Graphics Octane computer. To accelerate computations, we have also sometimes made use of software to distribute calculations for different sequences to an array of computers.

The DVQ metric computes elementary perceptual errors indexed by DCT frequency, block, color, and field of the video sequence, and allows selective pooling over subsets of these dimensions. In Figure 4, we show the perceptual error pooled over all dimensions except time (video field). The results are shown for the “flower” source and HRC’s 1-5. Particularly for the lowest bit rate, the perceptual error is clearly bursty and periodic, the periodicity presumably due to the GOP structure.

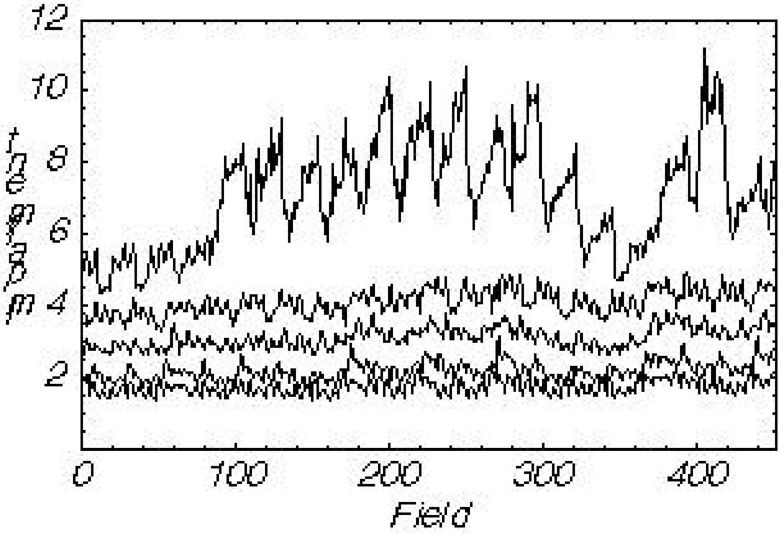


Figure 4. Impairment in each field for the source “flower” and HRC’s 1-5 (codec A at bit-rates of 2-10 Mbit/sec).

Time-varying predictions of this sort would be useful for predicting time-varying subjective assessments of video quality[11], but regrettably the requisite sequences and ratings are not publicly available. The total impairment score for a sequence is obtained by pooling over fields, using a Minkowski metric with an exponent of 4[1]. The five traces above, pooled in this way, yield scores of {34.7, 19.14, 14.56, 10.3, 8.21}.

5. COMPARING DATA AND METRIC

A simple method of comparing the impairment predictions generated by DVQ and the subjective data is to compute the root-mean-squared (rms) error between predictions and data. However, it has frequently been observed that when plotted against simple objective measures such as rms error or bit-rate, subjective impairment scores show a sigmoidal form. This is attributed to two aspects of human judgements: the flattening at the low end is attributed to a threshold, while the flattening at the high end is attributed to a saturation. In simple terms, impairments below a certain level are invisible, and impairments above a certain level are all considered about equally bad. Although such thresholds and saturation phenomena should be a part of an accurate model, they may be absent from current models, and thus in comparing model and data the possibility of a further non-linear transformation between model and data is often entertained. The non-linear transformation we have considered is a cubic polynomial. Thus we have first found the best-fitting cubic polynomial, and then computed the residual rms error. This fit is shown in Figure 5A. Finally, we present the results by plotting the data against the transformed predictions, as shown in Figure 5B. Following transformation, a unit slope straight line (as shown) is the best fitting polynomial. For this fit, the parameters of the polynomial were {a_0 = -1.635, a_1 = 0.573, a_2 = 0.0603, a_3 = -0.00085}.

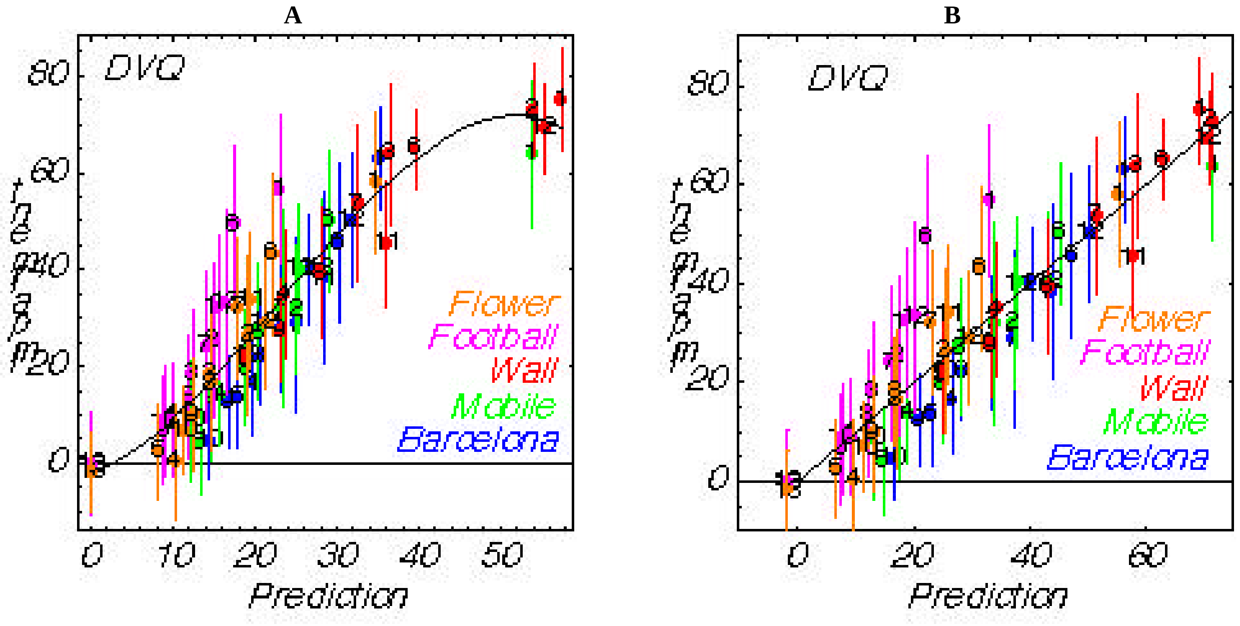


Figure 5. Comparison of DVQ prediction and subjective data. A: DVQ predictions vs data with fitted polynomial; B: transformed predictions. Numbers within points indicate the HRC.

The rms error of the fit of the DVQ metric is 14.61. For comparison, we have also computed the predictions of a model consisting of the rms error between the digital values of the reference and processed sequences (RMSE), and a model consisting simply of the inverse bit-rate of each sequence (MBPS). We have also assessed the predictions of the Sarnoff model[12] for these data, as reported in [8]. All of these may be compared to the rms error of the individual observers relative to their mean for a given condition (source and HRC), which is a measure of the underlying variability of the data. The rms errors for all of these models are noted in Table 2.

While more detailed statistical conclusions must await further analysis, some preliminary observations may be warranted. First, there is unlikely to be a statistical difference between DVQ and Sarnoff models for this data set. A choice between the two metrics must be based on other considerations, such as speed, simplicity, availability, or cost. Second, the variability in the data puts a bound on how well any metric can fit the data, and on how much effort we should expend on improvements to the models. Third, both Sarnoff and DVQ metrics appear to perform significantly better than the RMSE model. Furthermore, this data set does not exercise one feature of these models that should be expected to set them clearly above the RMSE metric, namely variations in display resolution and viewing distance.

Metric	**Rms Error**
DVQ	14.61
Sarnoff	14.40
RMSE	16.80
Mbits/sec	16.70
Condition Means	12.65

Table 2. Rms error of fit of various metrics to the subjective data.

Although the overall DVQ fit is reasonable, the “football” source appears to depart systematically from the model predictions. We have not yet discovered the explanation for this discrepancy. Interestingly, it is also evident in the fit of the RMSE and Sarnoff models. One possibility is that the scales used by observers are different for each source sequence. In this case, for example, the observers may be more demanding in their quality expectations for the “football” sequence.

6. CONCLUSIONS

We have evaluated the performance of the DVQ video quality metric by comparing its predictions to judgements of impairment made by 25 human observers viewing 5 reference sequences as processed by 12 HRCs. The DVQ metric performs about as well as the de facto standard (the Sarnoff model), and considerably better than models based on simple bit-rate or rms error. The quality of the predictions suggests the metric may be useful in practical applications. However the metric shows what appear to be systematic failures of prediction (the “football” sequence at low bit-rates) whose explanation awaits further research.

7. ACKNOWLEDGEMENTS

We thank Alexander Schertz of the Institut fuer Rundfunktecknik GmbH, and Tektronix, Inc. for providing the digital video sequences and subjective data used in this report. This work was supported by NASA Grant 199-06-12-39 from the Office of Life Science and Microgravity Research.

8. REFERENCES

1. A.B. Watson, "Toward a perceptual video quality metric," Human Vision, Visual Processing, and Digital Display VIII, 3299, 139-147 (1998).

2. K.T. Tan, M. Ghanbari and D.E. Pearson, "A video distortion meter," Picture Coding Symposium, 119-122 (1997).

3. T. Hamada, S. Miyaji and S. Matsumoto, "Picture quality assessment system by three-layered bottom-up noise weighting considering human visual perception," Society of Motion Picture and Television Engineers, 179-192 (1997).

4. C.J.v.d.B. Lambrecht, "Color moving pictures quality metric," International Conference on Image Processing, I, 885-888 (1996).

5. A.B. Watson, "Multidimensional pyramids in vision and video," Representations of vision: trends and tacit assumptions in vision research, A. Gorea, , 17-26, Cambridge University Press, Cambridge (1991).

6. A.B. Watson, "Perceptual-components architecture for digital video," Journal of the Optical Society of America A, 7(10), 1943-1954 (1990).

7. M. Ravel, J. Lubin and A. Schertz, "Pruefung eines Systems fuer die objektive Messung der Bildqualitaet durch den Vergleich von objektiven und subjektiven Testergebnissen," **FKT vol.52 nr. 10**, (1998).

8. A. Schertz, "IRT/Tektronix Investigation of Subjective and Objective Picture Quality for 2-10 Mbit/s MPEG-2 Video," Technischer Bericht des IRT **nr. B 159/97**, (1997).

9. ITU-R, "Recommendation ITU-R BT.601-5, Studio encoding parameters of digital television for standard 4:3 and wide screen 16:9 aspect ratios," , (1995).

10. ITU-R, "Methodology for the subjective assessment of the quality of television pictures," International Telecommunications Union **BT.500-7**, (1995).

11. R. Hamberg and H.d. Ridder, "Continuous assessment of perceptual image quality," Journal of the Optical Society A, 12(12), 2573-2577 (1995).

12. J. Lubin, "A Human Vision System Model for Objective Picture Quality Measurements," International Broadcasters' Convention, Conference Publication of the International Broadcasters' Convention, 498-503 (1997).