ELSEVIER

Signal Processing: *Image Communication* 17 (2002) 781–798

SIGNAL PROCESSING: IMAGE COMMUNICATION

www.elsevier.com/locate/image

PVQM – A perceptual video quality measure

A.P. Hekstra[a,*], J.G. Beerends[a], D. Ledermann[b], F.E. de Caluwe[c,1], S. Kohler[b,2,3], R.H. Koenen[d,1], S. Rihs[b,2,4], M. Ehrsam[e,2,5], D. Schlauss[b]

[a] *KPN Research, P.O. Box 421, 2260 AK, Leidschendam, The Netherlands*
[b] *Swisscom Innovations, CH-3050 Bern, Switzerland*
[c] *Nokia Nederland B.V., Loire 148, 2491 AK Den Haag, The Netherlands*
[d] *InterTrust Technologies, 4800 Patrick Henry Drive, Santa Clara, CA 95054, USA*
[e] *Sohard AG, Galgenfeldweg 18, 3000 Bern 32, Switzerland*

Received 1 March 2002; accepted 20 June 2002

Abstract

Modern video coding systems such as ISO-MPEG1,2,4 exploit properties of the human visual system, to reduce the bit rate at which a video sequence is coded, given a certain required video quality. As a result, to the degree in which such exploitation is successful, accurate prediction of the quality of the output video of such systems, should also take the human visual system into account. In this paper, we propose a perceptual video quality system, that uses a linear combination of three indicators. The indicators are, the "edginess" of the luminance, the normalized color error and the temporal decorrelation. In the benchmark by the Video Quality Expert Group (VQEG), a combined ITU-T and ITU-R expert group, the model showed the highest variance weighted regression overall correlation of all models.

Keywords: Objective video quality measurement

1. Introduction

With the introduction and standardization of new, perception based, digital video codecs [7,9], classical methods for measuring video quality, like peak signal-to-noise ratio and video bandwidth, became

*Corresponding author.
E-mail addresses: andries.hekstra@ieee.org (A.P. Hekstra), j.g.beerends@kpn.com (J.G. Beerends), daniel.ledermann@swisscom.com (D. Ledermann), f.e.decaluwe@hetnet.nl (F.E. de Caluwe), stefan.kohler@swisscom.com (S. Kohler), rob.koenen@ieee.org (R.H. Koenen), samuel.rihs@swisscom.com (S. Rihs), matthias.ehrsam@disetronic.ch (M. Ehrsam), denis.schlauss@swisscom.com (D. Schlauss).
[1] This work was done while the author was working at KPN Research.
[2] This work was done while the author was working at Swisscom Innovations.
[3] Currently working for Swisscom mobile, CH-3050 Bern, Switzerland.
[4] Currently working for Swisscom Broadcasting Services, CH-3050 Bern, Switzerland.
[5] Now at Disetronic Medical Systems AG, Postfach 3401, Burgdorf, Switzerland.

0923-5965/02/$ - see front matter
PII: S0923-5965(02)00056-5

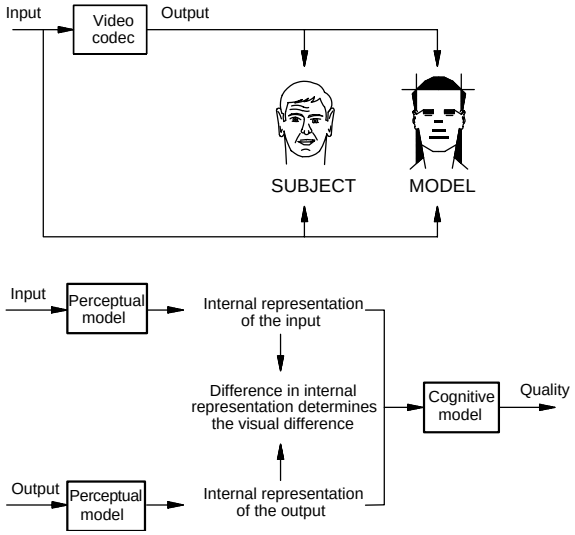

Fig. 1. Overview of the basic philosophy used in the development of the PVQM. A computer model of the subject, consisting of a perceptual and a cognitive model, is used to compare the output (degraded) of the video codec with the input (reference, ideal).

questionable ways of quantifying perceptual video quality. During the standardization process of these codecs the quality of the different proposals was therefore assessed only subjectively (see e.g. [15,1]). Subjective assessments are however time consuming, expensive and difficult to reproduce.

A fundamental question is whether objective methods can be formulated that can be used for prediction of the subjective quality of such perceptual coding techniques in a reliable way. A difference with classical approaches to video quality assessment is that system characterizations, based on test signal evaluations, are no longer useful because of the time varying, signal adaptive, techniques that are used in these codecs. Even simple techniques like motion estimation introduce time varying properties which are difficult to quantify from the perspective of the subjectively perceived video quality.

This paper will present a general method for measuring the quality of video codecs. The basic idea of the method is given in Fig. 1. It uses the same video material as would have been used in the subjective test and the quality of the video codec is based on a comparison of the internal representations of the input (reference, ideal, undistorted original) and output (degraded). The internal representations of the input and output are calculated with a perceptual model of the visual system.

2. Perception-based video quality measurement

The aspects that have to be modelled to obtain an accurate internal representation and map it onto the perceived video quality are [16]:

(1) spatio-temporal alignment, one has to know which parts of the reference and degraded picture are compared by subjects to base their opinion on;
(2) adaptation at the pupil and the retina, watching at video in a dark room is completely different from watching in a well-lighted room;
(3) brightness and contrast, what aspects of the luminance signal are important, some changes in brightness and contrast can even improve the picture quality of the output video sequence;

(4) color saturation, what aspects of color are important, some changes in color can even improve the picture quality of the output video sequence;
(5) spatio-temporal filtering, some changes are not visible;
(6) non-linear relation between luminance and brightness, a certain increase in luminance in the world causes a certain increase in brightness in the visual system;
(7) spatio-temporal masking, one spatio-temporal event may decrease or increase the visibility of another spatio-temporal event;
(8) eye movements, causing changes in the spatio-temporal filtering, e.g. when tracking a slow movement details remain visible while loosing tracking causes these details to get blurred or even invisible;
(9) distortion preferences, sometimes differences that are big in the internal domain cause little degradation while sometimes differences that are small are very disturbing.

Realistic modelling of all these aspects is not feasible and therefore practical systems use a simplified approach or do not take into account certain aspects [4,14].

One could doubt whether it is useful to model all aspects of human visual perception if one realizes that in the end picture quality is a concept that is dominated by cognitive aspects, such as knowledge of what aspects dominate the overall quality perception. If we take for example a picture of a table with many objects and we would be able to design a coding scheme that would degrade the picture by leaving out one object on this table, and replacing it by the background pattern, the resulting picture would be of high quality. If, however, we take the approach as given in Fig. 1, and directly compare the internal representations of the reference and degraded picture, we would detect a large difference between them, whatever perceptual model we would use. Furthermore, eye movements are nearly impossible to predict and thus, even at a perceptual level, the spatio-temporal filtering characteristics are largely unknown.

In this paper we will thus take the pragmatic approach of designing a video quality measurement system that uses a simple perceptual model combined with a model that takes into account the most dominant cognitive effects in quality measurement. The same strategy was taken in the development of a speech quality measurement system [3,2]. A simple perceptual model, in combination with the modelling of two cognitive effects that dominate speech quality perception, resulted in a quality measure (the Perceptual Speech Quality Measure, PSQM) that has a high correlation between objective and subjective speech quality measurement. In an ITU-T benchmark between five proposals the PSQM method was found to be superior in terms of correlation between objective and subjective measurements [10]. PSQM was standardized by the ITU-T as recommendation P.861 [11]. In a second phase, a few years later, a new improved version of PSQM (PSQM99) was integrated with a method from British Telecom [17] to form recommendation P.862, the current world standard for speech quality measurement [13].

For video, the Video Quality Experts Group (VQEG) [6], formed by experts from various backgrounds and affiliations, including ITU-T and ITU-R, performed a benchmark for 10 different quality metrics in the period 1998–1999. In this test over 26,000 subjective scores were generated on 20 different video sequences processed by 16 different video systems. The proposal from KPN Research, called the Perceptual Video Quality Measure (PVQM) showed highest correlations using a variance weighted regression correlation between subjective and objective quality scores [12]. Currently, the VQEG group is working towards a standard for objective video quality measurement [6].

The complete PVQM approach is split into two parts, the first part (Section 3) gives the spatio-temporal and luminance–chrominance alignment that is used before the comparison of input and output is being carried out. The second part (Section 4) gives a description of the core PVQM algorithm based on the fundamentals of human visual perception. Finally, Section 5 presents the most important conclusions and recommendations.

3. Spatio-temporal and luminance-chrominance alignment

3.1. Introduction

The PVQM as described in this paper is based on a comparison of the degraded video sequence, the output of the device under test, with a reference in the same way as carried out in a subjective experiment. When comparing two video sequences subjects will more or less compare identical parts of the sequences and judge the quality of the degraded sequence on the basis of the perceived difference. Furthermore, it is known that changes in brightness, contrast and chrominance lead to big differences between input and output values when making pixel by pixel comparisons, but only to minor changes in the perceived overall quality. This makes it necessary to align the reference (input) with the degraded (output) in both the spatial-temporal and luminance–chrominance domain.

The first step in the alignment procedure matches the histograms of the luminance and chrominance values. It transforms the luminance and chrominance values in such a way that the cumulative probability density functions become (nearly) identical. The video codec under test can introduce a linear or nonlinear gain or offset into the component signals of the degraded video signal. Such amplification effects need to be undone before a proper comparison can be made between with the reference and degraded video signal. The basics of the histogram matching are given in Section 3.2.

The histogram correction assumes that each pixel that is counted in the histograms of the reference sequence is also present in the degraded sequence. Initially, it is unknown whether the raw degraded video sequence contains shifts or horizontal stretching. Thus, it is also unknown what the effective margins are. Furthermore, in case of a positive delay, the first few frames of the degraded video sequence may not match the contents of the reference video sequences and must (later) be discarded for processing by PVQM in order not to inadvertedly make PVQM react as if there are serious visual errors. As a result, it is impossible to get everything right in one pass of the degraded and reference sequences. Therefore the histogram matching is carried out for a second time after the sequences have been spatio-temporally aligned.

The second step in the alignment procedure finds matching parts of the video sequences that are used by the subject to base his judgment on. This spatial-temporal alignment is carried out in two phases, a probing phase, in which for a subset of degraded frames the best matching reference frames are searched for (Section 3.3), and a matching phase where for all frames the best matches are searched for (Section 3.4), using a spatio-temporal search range smaller than in the probing phase. Both phases use the root mean square distance between reference and degraded frames. The searches are carried out simultaneously in time and space by using a block matching technique. For each block of 32×32 pixels the best matching block in a reference frame is searched for, using the mean square distance between reference and degraded block of pixels. In the probing phase a search range of about 10 frames (temporally) and 15×4 (horizontal $\times$ vertical, spatially) pixels is used. The results of this first alignment are used to calculate a rough alignment including a possible stretching. In the second step the search range is limited to small deviations (e.g. plus and minus two pixels) from the first alignment. A consequence of this dynamic spatio-temporal alignment is that sequences with a low temporal resolution (many frame repeats) will get a significantly better predicted MOS than would have been predicted without this alignment.

The alignment that is carried out does not use a test signal but uses the same input and output video sequences as were used in the subjective test.

3.2. Basics of histogram matching

Each of the components Y, Cb, Cr of the degraded video signal may have undergone a transformation with respect to its counterpart in the reference sequence by a linear or nonlinear curve. Linear amplifications with a gain larger than one are commonly found in video systems as the designers hope that

the increased brightness and color saturation increases the perceived quality of their systems. Because Y, Cb, Cr take on values from a bounded domain [0..255], amplification by a gain exceeding one must be nonlinear over some parts of the domain since otherwise the range of this mapping would exceed [0..255]. The luminance–chrominance alignment can correct for arbitrary increasing functions of the component signals. The information to carry out this correction is extracted from the histograms of the component signals of the reference sequence and the component signals of the degraded sequence. For each of the three components, a curve is computed that makes the component histograms of the corrected raw degraded video sequence approximately equal to that of the reference sequence. Some restrictions apply, such as a minimum or maximum slope constraint on the computed correction curves.

It is assumed that the pixel values of reference and degraded can only take on integer values in the range [0..255]. As a result, the histograms, for each component (Y, Cb,Cr) of all values in [0..255] can be determined by a simple counting procedure. The cumulative histograms indicate which fraction of the pixels lies below or exactly at a certain value. From these cumulative histograms one can calculate the correction curve that makes the histograms of the input (x) and output (y) component signals approximately equal. A complication is that the histogram values can be zero, causing the cumulative histogram functions to be locally noninvertible. In that case, a simple remedy is to start out in the middle of the domain near the value of 128 where the cumulative histograms are invertible and proceeding to the tails resort to a continuation of the correction curves with local slope of 1 in case the histogram values (fractions) are too small.

Another complication in matching the histograms of input and output occurs when the codec under test is clipping the range of luminance values. In this case information has been lost by the clipping operation and the question rises what the best inverse value for the clipping value is. The best possible answer to this system has been found to map the clipping level back to the center of its histogram mass in the reference histogram. This is most easily achieved by a slight smearing of the cumulative histograms.

In practice, clipping at extreme values of 0 or 255 occurs less often for the chrominance values. Thus for the chrominance corrections unsmeared cumulative histogram functions are used. A peculiarity of the chrominance signal is that the reference sequence may be colorless, whereas the degraded sequence contains cross color artifacts, e.g. due to PAL processing. A simple (gain, offset) correction instead of the histogram correction would set the gain for the color components equal to zero, the offset to 128 and thus remove the artifacts. The objective measure would not see the introduced colors in its input. This is a fundamental problem for any alignment algorithm that does not use test patterns in the video sequences. With histogram correction, the problem can be circumvented partially, be requesting a minimum slope of the correction curve. This is equivalent to the requirement of a minimum gain of e.g. 0.5 in a (gain, offset) correction. Note that with histogram correction in addition a slope of one outside the peak of the reference color components at 128 is enforced.

The procedure described in this section is first used on the raw degraded video sequences y_{raw} (preliminary histogram correction). After the probing and matching alignment phases the histograms of the reference sequence and the degraded sequences, displaced by the final displacement fields, are recomputed and used to compensate the degraded sequence in final histogram correction.

3.3. The probing phase

After the preliminary histogram correction the spatio-temporal shift between the corrected degraded sequence and the input reference sequence is determined using a minimum mean square objective function. This probing is carried out with temporal subsampling factor of 17. The initial delay and displacement vector search ranges may be large and the goal is to narrow them down. The choice of a prime number ensures that in the case of degraded sequences with frame repeat, the subsampling and the frame repeat cannot run in phase with each other. With respect to the start and stop frame index of the sequences some

margins must be kept to allow for a delay between degraded and reference sequence up to values specified by the initial delay search range and still have a matching reference frame for each degraded frame.

During the probing phase, the degraded frames are corrected with the preliminary histogram correction curves. For the subset of degraded frames the average of the matching error is computed by full search block matching. Some analogue systems, like VHS systems, can regroup fields into new frames. The decision whether the degraded frames have underwent regrouping of fields into frames is taken depending on which of the two matching errors, on a sequence level, is the smallest. This decision is then fixed.

In case the initial search ranges are large, significant savings in computation time without loss of performance can be achieved by narrowing the search range vector down starting from its initial value for each of the probing passes. Such modifications are ad hoc and lie beyond the scope of this paper.

The final matching of degraded frames to reference frames searches through a maximum of the delays that have been found on the probing set. Also the block matchings that are carried out on the subset are used to find the average horizontal and vertical shift. Furthermore, the block matchings allow us to determine a horizontal stretching curve as some codecs use horizontal spatial scaling.

3.4. The matching phase

The probing phase has yielded the information about the matching delay range, the horizontal and vertical shifts and the horizontal stretching curve. The matching delay is computed using the minimal matching error over all reference frames within the specified delay range and for each reference frame over all possible displacement fields that lie close to the stretching curve (e.g. plus or minus two pixels).

With frame repeat sequences, the matching delay can show a saw tooth behavior, 0, 1, 2, 3, 0, 1, 2, 3, 0, 1, 2, etc. In that case one would like to distinguish the frame repeat from the delay and let that be zero in the example. Thus, the perceptual delay is defined as the local minimum of the delay.

4. The perceptual video quality measure

4.1. Introduction

After the frames of the degraded video sequence are spatio-temporally and luminance–chrominance aligned with the reference sequence a number of quality indicators are calculated for each of the three video components Y, CR, CB, luminance and color-difference signals for the ITU-R 601 video representation. From a perceptual point of view a transformation from Y, CR, CB to a representation that is more close to the internal representation of the picture should give higher correlations between objective and subjective quality evaluation results. Although during the development of PVQM the $L*, u*, v*$ representation was used, the most widely used perceptual color representation for color television, it turned out that only minor improvements resulted from this transformation while the overall complexity of the model increased significantly.

The basic idea of the PVQM as given in Fig. 1 is applied to both the luminance Y and chrominance signals CR, CB. Both the input x (reference) and the output (in most cases degraded) video signal y are mapped onto differences in three different quality indicators, one describing the difference in edginess of the luminance signal (Section 4.3), one temporal indicator that describes the amount of movement or change in the reference video sequence which can be used to decrease sensitivity for details in fast moving pictures (Section 4.4). Furthermore, there is one indicator describing the perceived difference in chrominance (Section 4.5). The way these three indicators are derived from the video signal is strongly dominated by the way we perceive moving pictures. In a final stage the way in which variations over time are degraded and

the mapping from the indicators to the overall perceived video quality is given (Section 4.6). This stage of modelling is dominated by more central, cognitive processes (see Fig. 1).

4.2. Perceptual modelling: spatio-temporal filtering

Like the perceptual color representation $L*, u*, v*$ only showed a minor improvement in correlation over the standard Y, CR, CB representation it turned out that perceptual-based complex spatio-temporal filtering only showed a minor improvement in correlation over very simple spatio-temporal filtering like e.g. a simple gradient filtering.

It is well known that the HVS is much more sensitive to the sharpness of the luminance component than that of the chrominance components. Furthermore, the HVS has a contrast sensitivity function that decreases at high spatial frequencies. Since PVQM is essentially a field-based algorithm, high vertical frequencies are not picked up by the model. Since PVQM operates on the two fields of an interlaced frame separately, for ease of processing, the odd lines of a frame are reordered and placed in the top half of a frame TOP, and the even lines are reordered to the bottom half of a frame BOT.

In order to reflect the reduced sensitivity to high horizontal spatial frequencies a horizontal $(1, 2, 1)$ filter is applied to both the degraded luminance signal and the reference luminance signal ($i = 1, 2, \ldots, M - 2$). The details of the filtering are given in Appendix A.

4.3. Perceptual modelling: determination of the edginess of the luminance

One of the features most dominantly lacking from the SNR as a predictor of quality is that it does not reflect the sharpness of the images presented. The HVS has a pronounced sensitivity to edges and local changes in the luminance. In this paper the local edginess edge of the reference and y_{edge} of the degraded frames are computed and compared. They are computed as an approximation to the local gradient of the luminance signal. Furthermore, the edginess signals x_{edge} and y_{edge} are dilated to allow for the fact that single pixel errors are more visible then otherwise predicted.

Subjective experiments are normally carried out with a background illumination that corresponds to the mid-range gray level of the display. The HVS of the test persons adapts to this background illumination. As a result, the visual quality of dark and light areas in the picture is of lesser importance to the quality judgment. In order to reflect this effect, the absolute deviation of the luminance from 100 is computed for both the reference video signal and the degraded video signal. It is the maximum of these absolute deviations that is used in the edginess frame indicator.

If for a particular pair of reference and degraded sequences, wherever x'_{edge} is nonzero, y'_{edge} is in general smaller than x'_{edge}, this is perceived as a loss of sharpness ([5], see Fig. 2). Some reference sequences have an overall higher edginess than others. In turns out that the relative decrease of y'_{edge} with respect to x'_{edge} is a better indicator of loss of sharpness than the absolute difference. The indicator computed in this section not only indicates loss of sharpness ($y'_{\text{edge}}[i,j,t]$ minus $x'_{\text{edge}}[i,j,t]$ negative). Also, the introduction of sharpness is registered as a distortion ($y'_{\text{edge}}[i,j,t]$ minus $x'_{\text{edge}}[i,j,t]$ positive). The relativeness of the effect also manifests itself locally for introduced edginess. The introduction of edginess in areas with a lot of edginess is less disturbing than the introduction of sharpness where little edginess is originally present.

$$e[i,j,t] = \frac{y'_{\text{edge}}[i,j,t] - x'_{\text{edge}}[i,j,t]}{x'_{\text{edge}}[i,j,t] + 80 + \text{dev}[i,j,t]},$$

$$\text{dev}[i,j,t] = \max\{\text{abs}(x3[0,i,j,t] - 100),\ \text{abs}(y3[0,i,j,t-100)\}.$$

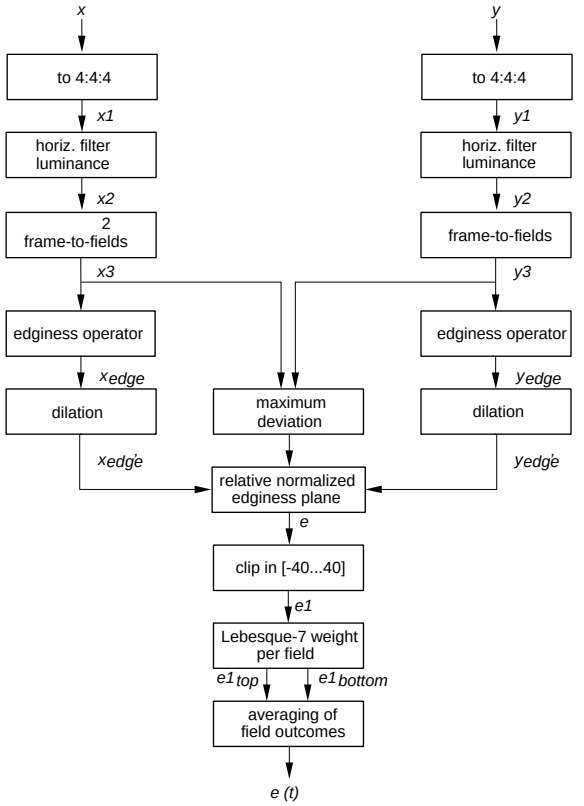

Fig. 2. Objective quality measurement procedure, luminance indicator. The variables are listed in Table 1.

The normalized change in edginess e is locally clipped to the range $[-40, 40]$.

$$e1[i,j,t] = \begin{cases} -40 & \text{when} \quad e[i,j,t] \leqslant -40, \\ 40 & \text{when} \quad e[i,j,t] \geqslant 40, \\ e[i,j,t] & \text{otherwise.} \end{cases}$$

4.4. Perceptual modelling: determination of the temporal decorrelation indicator

The edginess indicator $e1$ is a pure spatial indicator. However, the spatial content of a sequence is judged more critically in case of still images than for images with fast motion and rapid changes. To reflect this, the positive contributions to the DMOS_P from $e1$ (and n) indicator should be compensated by a contribution from an indicator that measures the temporal variability of the reference or degraded video sequence. The (peak) temporal variability of the degraded video signal is also influenced by transmission errors and the presence or absence of frame repeats. As a result, the temporal variability is best measured on the luminance of the reference sequence, see Fig. 3. The decorrelation $d[t]$ is defined as 1 minus the correlation between the current ($[t]$) and previous frame ($[t-1]$):

$$d[t] = 1 - \frac{\sum_{(i,j)\in \mathrm{TOP}\cup\mathrm{BOT}} x3[0,i,j,t]x3[0,i,j,t-1]}{\sum_{(i,j)\in \mathrm{TOP}\cup\mathrm{BOT}} x3[0,i,j,t]^2 \sum_{(i,j)\in \mathrm{TOP}\cup\mathrm{BOT}} x3[0,i,j,t-1]^2}.$$

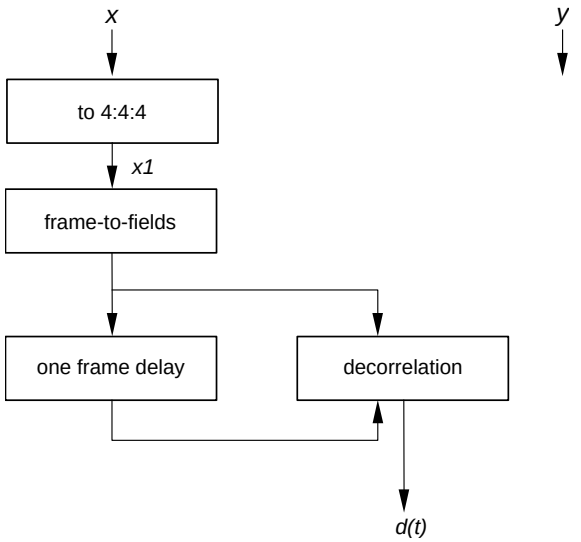

Fig. 3. Objective quality measurement procedure, decorrelation indicator. The variables are listed in Table 1.

4.5. Perceptual modelling: determination of the chrominance indicator

In the Cb, Cr domain, the HVS is more sensitive to areas than to edges. Furthermore, the HVS is much more sensitive to the quality of the Cr component than that of the Cb component. In regression experiments on our training databases, on the sequence level, the Cr indicator dominates the contribution to the DMOS_P unless the error in the Cb component is twice bigger than that in the Cr component.

The HVS is less sensitive to numerical errors in the color difference components that occur in areas that have saturated colors. This is even more true in bright areas. To reflect this, the color saturation is computed as follows. As for the deviation signal, the maximum of the color saturation of the reference signal and the degraded video signal is taken:

$$\mathrm{sat}_x[i,j,t] = \sqrt{(x3[1,i,j,t]-128)^2 + (x3[2,i,j,t]-128)^2}$$

and equivalently for sat_y. We define

$$\mathrm{sat}[i,j,t] = \max\{\mathrm{sat}_x[i,j,t], \mathrm{sat}_y[i,j,t]\}.$$

For each pixel, the normalized color error is determined ($c = 1, 2$).

$$n[c,i,j,t] = \frac{\mathrm{abs}(y3[c,i,j,t] - x3[c,i,j,t])}{25 + 0.3\mathrm{sat}[i,j,t]}.$$

See Fig. 4.

4.6. Cognitive modelling: aggregation of the indicators over space and time

When for each pixel position the three quality indicators are calculated one has more or less modelled the most prominent parts of visual perception. Two describes a spatio-temporal distribution of the error of the luminance and chrominance signal ($e1$ and n) and one describes the amount of movement over time. Clearly, this perceptual model is far too simplistic to account for all perceptual phenomenon that play a role in the perception of video quality. However, from a certain level of sophistication further

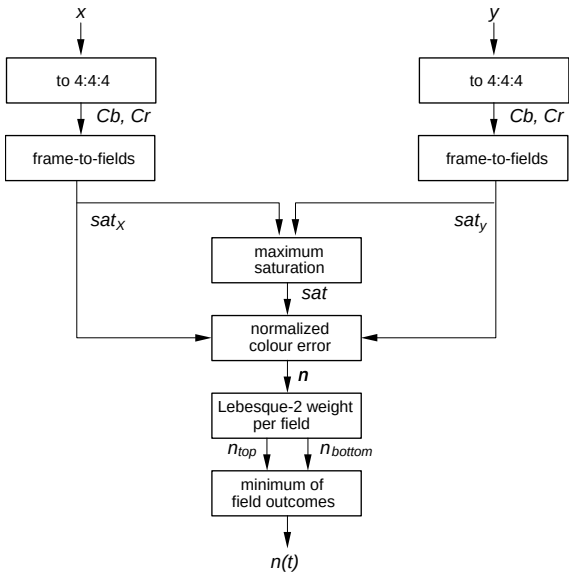

Fig. 4. Objective quality measurement procedure, chrominance indicator. The variables are listed in Table 1.

improvements in the perceptual domain only give minor improvements in the overall performance of the video quality measure in terms of correlations with the subjectively perceived video quality. During the development of PVQM it turned out that the next step in improving the correlation could be found in the modelling of a more central, cognitive phenomenon that dominates video quality perception. This modelling is focussed on the way people aggregate the error over space and time.

When we calculate the $e1$ indicator over the whole picture and over the whole sequence and compare the spatio-temporal distribution with quality judgments one will find three cognitive effects that are important and easy to model. The first one is that errors at the edges are perceived to be less disturbing then errors in the middle of the picture. Therefore the $e1$ and n indicators are aggregated over top and bottom field separately using heavier weighting of the changes that have occurred in the central part of the display using a sinusoidal weighting function. Because the two fields are placed on top of each other in the processing of the frames this weighting function has two peaks, one in the center of each field.

$$w[i,j] = \sin\left(\pi\frac{i}{M}\right)\mathrm{abs}\left(\sin\left(2\pi\frac{j}{N}\right)\right), \quad i = 1,2,\ldots,M,\ j = 1,2,\ldots,N.$$

The second cognitive effect that can be modelled very simple is the dominance on the perceived quality of small areas that are severely degraded. By aggregation of the $e1$ indicator with a Lebesgue-7 measure over space one will put a higher weight on large spatially localized distortions,

$$e1_{\mathrm{top}}[t] = \sqrt[7]{\frac{\sum_{(i,j)\in \mathrm{TOP}} \mathrm{abs}(e1[i,j,t])^7 w[i,j]}{\sum_{(i,j)\in \mathrm{TOP}} w[i,j]}}$$

and equivalently for e_{bot}. The outcome of the Lebesgue-7 weights can be dominated by relatively small areas of the image with gross errors. In order for a frame t to have a large $e[t]$ indicator, both fields must have large Lebesgue-7 weights.

$$e[t] = \frac{e1_{\mathrm{top}}[t] + e1_{\mathrm{bot}}[t]}{2}.$$

In experiments with time localized errors (e.g. transmission errors) one can show that local errors have a big impact on the DMOS. In order to reflect this in the PVQM, the edginess frame indicators had to be aggregated over time using a Lebesgue-7 weight (P = number of frames).

$$E = \sqrt[7]{\frac{\sum_{t=0,1,\ldots,P-1} e[t]^7}{P}}.$$

In the quality indicator optimization process it turned out that small values of E were invisible. The correlation between objective and subjective results could be improved by introducing a deadzone around zero:

$$E' = \begin{cases} 0 & \text{if } E < 7, \\ E - 7 & \text{otherwise.} \end{cases}$$

The decorrelation indicator that compensates the edginess indicator and normalized color error indicator, is likewise aggregated using a Lebesgue-7 weight.

$$D = \sqrt[7]{\frac{\sum_{t=1,2,\ldots,P-1} d[t]^7}{P-1}}.$$

The color indicator n is just averaged over time. Using Lebesgue-N, $N \leqslant 1$ weights which led to lower correlations on the available training databases.

The DMOS_P is predicted as a linear combination of indicators

$$\mathrm{DMOS}_P = 3.95E' + 0.74N - 0.78D - 0.4,$$

$$\mathrm{DMOS}'_P = \begin{cases} 0 & \text{if } \mathrm{DMOS}_P < 0, \\ 85 & \text{if } \mathrm{DMOS}_P > 85, \\ \mathrm{DMOS}_P & \text{otherwise.} \end{cases}$$

4.7. Results

Fig. 5 gives the results of the PVQM trained on a large medium to high quality database that included a wide range of digital codec distortions (such as H.263 with and without frame repeat, MPEG2, ETSI codecs) and analog PAL, VHS and Betacam distortions. It also contained a wide variety of alignment problems such as horizontal stretching over more than 50 pixels, strange increases in saturation, etc. The correlation between subjective quality, DMOS, and objective quality, DMOS_P, is 0.934.

5. Conclusions

This paper presents a new method for assessing video quality. It uses ITU-R 601 [8] format video sequences and compares the input, the reference, video sequence with a degraded output. Because global changes in the brightness and contrast only have a limited impact on the subjectively perceived quality the method first aligns the luminance and chrominance of the signals. After a spatio-temporal alignment, that is based on a kind of block matching procedure, it makes a pixel by pixel comparison for both the luminance and chrominance signals. For the luminance signal the pixel by pixel comparison is focused on the edges of the picture. Furthermore the objective quality measure takes into account the fact that the eye cannot follow rapid time variations, especially when different motions take place within a video sequence.

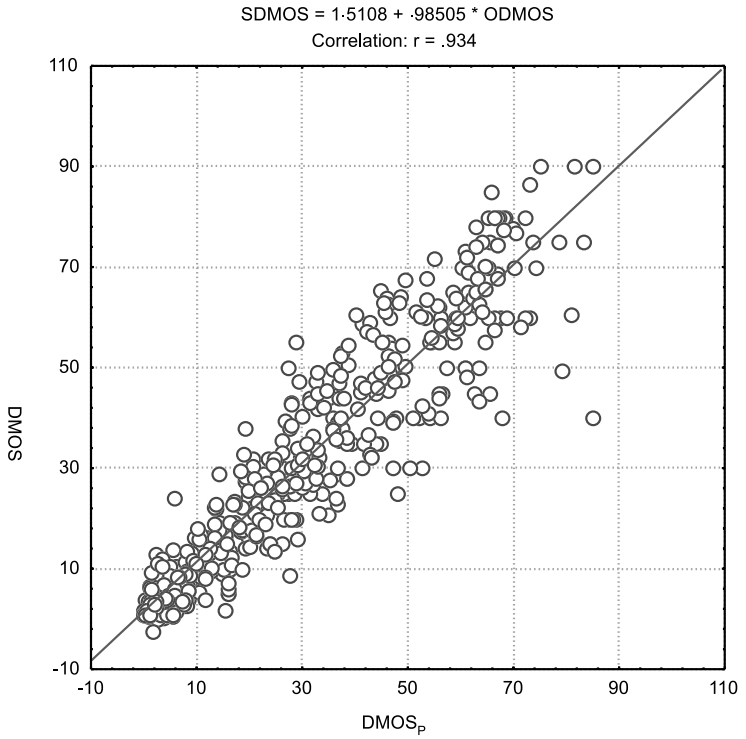

Fig. 5. Scatter plot of DMOS versus $DMOS_P$ produced by PVQM on a pool of databases processed with H.263 (with and without frame repeat) codecs, hardware and software MPEG2 codecs, ETSI codecs, PAL and Betacam systems, and spatial filters.

The final objective metric, called the Perceptual Video Quality Measure (PVQM), shows a good correlation between objective and subjective scores. On unknown material PVQM, showed a variance weighted regression correlation between subjective and objective quality scores of around 0.85, better[6] than peak signal-to-noise ratio ($r = 0.80$) and other video quality metrics ($r \approx 0.80$) [12].

Appendix A. Detailed description of PVQM

Table 1 gives an overview of the variables in the perceptive model of PVQM.

A.1. Spatio-temporal filtering

The HVS has a contrast sensitivity function that decreases at high spatial frequencies. Since PVQM is essentially a field-based algorithm, high vertical frequencies are not picked up by the model. In order to reflect the reduced sensitivity to high horizontal spatial frequencies a horizontal $(1, 2, 1)$ filter is applied to both the degraded luminance signal and the reference luminance signal ($i = 1, 2, \ldots, M - 2$). This leads to the following spatial filtering mapping:

$$x2[0, 0, j, t] = x1[0, 0, j, t],$$

[6] Within VQEG, these differences were judged to be statistically insignificant.

Table 1
Variables in PVQM

c	Index in component planes (0 =“Y”, 1 =“Cb”, 2 =“Cr”)
$M[c]$	Number of active pixels per line in component plane ($M[0] = M, M[1] = M[2] = Mc$)
i	Horizontal index in an interlaced frame or a frame of which the fields have been separated ($i = 0, \ldots, M(c) - 1$); $i = 0$ corresponds to the left of the display
j	Vertical index in a frame or a frame of which the fields have been separated ($i = 0, 1, \ldots, N - 1$); $j = 0$ corresponds to the top of the display
t	Frame index ($t = 0, 1, \ldots, P - 1$)
$x[c, i, j, t]$	4:2:2 reference video signal
$y[c, i, j, t]$	Spatio-temporally aligned and amplification corrected version of the 4:2:2 degraded video signal (e.g. as output by NIAR, see Appendix A)
$x1[c, i, j, t]$	4:4:4 resampled version of $x[c, i, j, t]$
$y1[c, i, j, t]$	4:4:4 resampled version of $y[c, i, j, t]$
$x2[c, i, j, t]$	Horizontally filtered version of $x1[c, I, j, t]$
$y2[c, i, j, t]$	Horizontally filtered version of $y1[c, I, j, t]$
$x3[c, i, j, t]$	Version of $x2[c, i, j, t]$ in which the top half of the frame is filled by the top field (line $j = 0, 2, \ldots, N/2 - 2$) of $x2[c, i, j, t]$ and the bottom half is filled by the bottom field (line $j = 1, 3, \ldots, N/2 - 1$) of $x2[c, i, j, t]$
$y3[c, i, j, t]$	Version of $y2[c, i, j, t]$ in which the top half of the frame is filled by the top field (line $j = 0, 2, \ldots, N/2 - 2$) of $y2[c, i, j, t]$ and the bottom half is filled by the bottom field (line $j = 1, 3, \ldots, N/2 - 1$) of $y2[c, i, j, t]$
$x_{\mathrm{hor}}[i, j, t]$	Approximation of the horizontal derivative of $x3[0, i, j, t]$
$y_{\mathrm{hor}}[i, j, t]$	Approximation of the horizontal derivative of $y3[0, i, j, t]$
$x_{\mathrm{vert}}[i, j, t]$	Approximation of the vertical derivative of $x3[0, i, j, t]$
$y_{\mathrm{vert}}[i, j, t]$	Approximation of the vertical derivative of $y3[0, i, j, t]$
$x_{\mathrm{edge}}[i, j, t]$	Edginess of $x3[0, i, j, t]$
$y_{\mathrm{edge}}[i, j, t]$	Edginess of $y3[0, i, j, t]$
$x'_{\mathrm{edge}}[i, j, t]$	Dilated edginess of $x3[0, i, j, t]$
$y'_{\mathrm{edge}}[i, j, t]$	Dilated edginess of $y3[0, i, j, t]$
$\mathrm{dev}[i, j, t]$	Maximum deviation of the reference filtered luminance $x3[0, i, j, t]$ and the degraded filtered luminance $y3[0, i, j, t]$ from 100
$e[i, j, t]$	Normalized relative change of edginess (a function of $x_{\mathrm{edge}}[i, j, t]$, $y_{\mathrm{edge}}[i, j, t]$, $\mathrm{dev}[i, j, t]$)
$e1[i, j, t]$	Version of $e[i, j, t]$ that has been clipped in between -40 and $+40$
$w[i, j]$	Sinusoidal weighting function that puts a heavier weighting at the center of the fields
TOP	Area that contains the top field (in the top half of the frame after separation of the fields), excluding the margins
BOT	Area that contains the bottom field (in the bottom half of the frame after separation of the fields), excluding the margins
$e1_{\mathrm{top}}[t]$	Weighted Lebesgue-7 norm of top half (i.e. top field) of $e1[i, j, t]$
$e1_{\mathrm{bot}}[t]$	Weighted Lebesgue-7 norm of bottom half (i.e. bottom field) of $e1[i, j, t]$
$e[t]$	Average of $e1_{\mathrm{top}}[t]$ and $e1_{\mathrm{bot}}[t]$
E	Edginess indicator on the sequence level
E'	Clipped version of E
$\mathrm{sat}_x[i, j, t]$	Color saturation of $x3[c, i, j, t]$
$\mathrm{sat}_y[i, j, t]$	Color saturation of $y3[c, i, j, t]$
$\mathrm{sat}[i, j, t]$	Maximum of $\mathrm{sat}_x[i, j, t]$ and $\mathrm{sat}_y[i, j, t]$
$n[c, i, j, t]$	Normalized color error ($c = 1, 2$)
$n_{\mathrm{top}}[t]$	Normalized color indicator for the top half (i.e. top field) of frame t
$n_{\mathrm{bot}}[t]$	Normalized color indicator for the bottom half (i.e. bottom field) of frame t
$n[t]$	Normalized color indicator for frame t
N	Normalized color error on the sequence level
$d[t]$	Decorrelation indicator
D	Decorrelation indicator on the sequence level
DMOS'_P	Clipped DMOS_P

$$x2[0,i,j,t] = \frac{x1[0,i-1,j,t] + 2x1[0,i,j,t] + x1[0,i+1,j,t]}{4},$$

$$x2[0,M-1,j,t] = x1[0,M-1,j,t],$$

$$y2[0,0,j,t] = y1[0,0,j,t],$$

$$y2[0,i,j,t] = \frac{y1[0,i-1,j,t] + 2y1[0,i,j,t] + y1[0,i+1,j,t]}{4},$$

$$y2[0,M-1,j,t] = y1[0,M-1,j,t].$$

A.2. Separation of the fields in a frame

PVQM contains a field-based perceptual model. In order to facilitate the computations, the field containing the top line of a frame (“the top field”) is put into the top half of the array that first stored the frame. The other field (“the bottom field”) is stored in the bottom half of the array that previously stored the frame ($j = 0, 1, \ldots, N/2-1$).

$$x3[c,i,j,t] = x2[c,i,2j,t],$$

$$x3\left[c,i,\frac{N}{2}+j,t\right] = x2[c,i,1+2j,t],$$

$$y3[c,i,j,t] = y2[c,i,2j,t],$$

$$y3\left[c,i,\frac{N}{2}+j,t\right] = y2[c,i,1+2j,t].$$

Note. In the model no reference is made to “the previous field” only to “the same field of the previous frame”. Therefore, it is immaterial whether the top fields or the bottom fields are temporally earlier.

A.3. Determination of the edginess of the luminance

A.3.1. Determination of edginess

The edginess is computed as an approximation to the local gradient of the luminance signal. Since the arrays are organized in fields, the length of the vertical filter is only half that of the horizontal filter ($i = 1,2,\ldots,M-2, j = 1,2,\ldots,N/2-2$ or $j = N/2+1, N/2+2,\ldots,N-1$, otherwise x_{edge} and y_{edge} are zero),

$$x_{\text{hor}}[i,j,t] = \frac{x3[i+2,j,t] + x3[i+1,j,t] - x3[i-1],j,t] - x3[i-2,j,t]}{2},$$

$$x_{\text{vert}}[i,j,t] = x3[i,j+1,t] - x3[i,j-1,t],$$

$$x_{\text{edge}}[i,j,t] = \sqrt{x_{\text{hor}}[i,j,t]^2 + x_{\text{vert}}[i,j,t]^2},$$

$$y_{\text{hor}}[i,j,t] = \frac{y3[i+2,j,t] + y3[i+1,j,t] - y3[i-1],j,t] - y3[i-2,j,t]}{2},$$

$$y_{\text{vert}}[i,j,t] = y3[i,j+1,t] - y3[i,j-1,t],$$

$$y_{\text{edge}}[i,j,t] = \sqrt{y_{\text{hor}}[i,j,t]^2 + y_{\text{vert}}[i,j,t]^2}.$$

A.3.2. Dilation of the edginess

The edginess signals x_{edge} and y_{edge} are dilated to allow for the fact that single pixel errors are more visible then otherwise predicted.

$$x'_{\text{edge}}[i,j,t] = \max\{x_{\text{edge}}[i',j',t] | i' = i-1, i, i+1,\ \ j' = j-1, j, j+1\},$$

$$y'_{\text{edge}}[i,j,t] = \max\{y_{\text{edge}}[i',j',t] | i' = i-1, i, i+1,\ \ j' = j-1, j, j+1\}.$$

A.3.3. Determination of maximum deviation from mid-range luminance

Subjective experiments are normally carried out with a background illumination that corresponds to the mid-range gray level of the display. The HVS of the test persons adapts to this background illumination. As a result, the visual quality of dark and light areas in the picture is of lesser importance to the quality judgement. In order to reflect this effect, the deviation of the luminance from 100 is computed for both the reference video signal and the degraded video signal. It is the maximum of these deviations that is used in the edginess frame indicator.

$$\text{dev}[i,j,t] = \max\{\text{abs}(x3[0,i,j,t] - 100), \text{abs}(y3[0.i.j,t - 100)\}.$$

A.3.4. Determination of normalized change in edginess

If for a particular pair of reference and degraded sequences, wherever x'_{edge} is nonzero, y'_{edge} is in general smaller than x'_{edge}, this is perceived as a loss of sharpness. Some reference sequences have an overall higher edginess than others. It turns out that the relative decrease of y'_{edge} with respect to x'_{edge} is a better indicator of loss of sharpness than the absolute difference. The indicator computed in this section not only indicates loss of sharpness ($y'_{\text{edge}}[i,j,t]$ minus $x'_{\text{edge}}[i,j,t]$ negative). Also, the introduction of sharpness is registered as a distortion ($y'_{\text{edge}}[i,j,t]$ minus $x'_{\text{edge}}[i,j,t]$ positive). The relativeness of the effect also manifests itself locally for introduced edginess. The introduction of edginess in areas with a lot of edginess is less disturbing than the introduction of sharpness where little edginess is originally present.

$$e[i,j,t] = \frac{y'_{\text{edge}}[i,j,t] - x'_{\text{edge}}[i,j,t]}{x'_{\text{edge}}[i,j,t] + 80 + \text{dev}[i,j,t]}.$$

A.3.5. Clipping of the normalized change in edginess

The normalized change in edginess e is locally clipped to the range $[-40, 40]$,

$$e1[i,j,t] = \begin{cases} -40 & \text{when } e[i,j,t] < -40, \\ 40 & \text{when } e[i,j,t] > 40, \\ e[i,j,t] & \text{otherwise.} \end{cases}$$

A.3.6. Aggregation over top and bottom field

The normalized change in edginess is aggregated over top and bottom field separately using heavier weighting of the changes that have occurred in the central part of the display using a sinusoidal weighting function. The weighting function has two peaks, one in the center of each field.

$$w[i,j] = \sin\left(\pi\frac{i}{M}\right)\text{abs}\left(\sin\left(\pi\frac{j}{N}\right)\right).$$

The aggregation of the relative change in edginess e is performed using a Lebesgue-7 weight with w as measure ($t = 0, 1, \ldots, P-1$).

$$\mathrm{TOP} = \left\{(i,j)|i = L, \ldots, M-R-1, j = \frac{T}{2}, \ldots, \frac{N-B}{2} - 1\right\},$$

$$e1_{\mathrm{top}}[t] = \sqrt[7]{\frac{\sum_{(i,j)\in \mathrm{TOP}} \mathrm{abs}(e1[i,j,t])^7 w[i,j]}{\sum_{(i,j)\in \mathrm{TOP}} w[i,j]}},$$

$$\mathrm{BOT} = \left\{(i,j)|i = L, \ldots, M-R-1, j = \frac{N+T}{2}, \ldots, N - \frac{B}{2} - 1\right\},$$

$$e1_{\mathrm{bot}}[t] = \sqrt[7]{\frac{\sum_{(i,j)\in \mathrm{BOT}} \mathrm{abs}(e1[i,j,t])^7 w[i,j]}{\sum_{(i,j)\in \mathrm{BOT}} w[i,j]}}.$$

A.3.7. Averaging of the field indicators

The outcome of the Lebesgue-7 weights in the previous paragraph can be dominated by relatively small areas of the image with gross errors. In order for a frame t to have a large $e[t]$ indicator, both fields must have large Lebesgue-7 weights.

$$e[t] = \frac{e1_{\mathrm{top}}[t] + e1_{\mathrm{bot}}[t]}{2}.$$

A.4. Determination of the temporal decorrelation indicator

The decorrelation frame level indicator ($t = 1, \ldots, P-1$) is defined by

$$d[t] = 1 - \frac{\sum_{(i,j)\in \mathrm{TOP}\cup\mathrm{BOT}} x3[0,i,j,t]x3[0,i,j,t-1]}{\sum_{(i,j)\in \mathrm{TOP}\cup\mathrm{BOT}} x3[0,i,j,t]^2 \sum_{(i,j)\in \mathrm{TOP}\cup\mathrm{BOT}} x3[0,i,j,t-1]^2}.$$

The advantage of a correlation (or inner product) between the current original frame and the previous original frame instead of a root mean square error is that it is independent of the overall brightness of the sequence.

A.5. Determination of the chrominance indicator

In the Cb, Cr domain, the HVS is more sensitive to areas than to edges. Furthermore, the HVS is much more sensitive to the quality of the Cr component than that of the Cb component. On the sequence level, the Cr indicator dominates the contribution to the DMOS_P unless the error in the Cb component is twice bigger than that in the Cr component.

A.5.1. Determination of the maximum color saturation

The HVS is less sensitive to numerical errors in the color difference components that occur in areas that have saturated colors. This is even more true in bright areas. To reflect this, the color saturation is computed as follows. As for the deviation signal, the maximum of the color saturation of the reference signal and the degraded video signal is taken.

$$\mathrm{sat}_x[i,j,t] = \sqrt{(x3[1,i,j,t] - 128)^2 + (x3[2,i,j,t] - 128)^2},$$

$$\mathrm{sat}_y[i,j,t] = \sqrt{(y3[1,i,j,t]-128)^2 + (y3[2,i,j,t]-128)^2},$$

$$\mathrm{sat}[i,j,t] = \max\{\mathrm{sat}_x[i,j,t], \mathrm{sat}_y[i,j,t]\}.$$

A.5.2. Determination of the normalized color error

For each pixel, the normalized color error is determined ($c = 1, 2$).

$$n[c,i,j,t] = \frac{\mathrm{abs}(y3[c,i,j,t] - x3[c,i,j,t])}{25 + 0.3\mathrm{sat}[i,j,t]}.$$

A.5.3. Aggregation over top and bottom field

The normalized color error is averaged over the fields with a heavier weight on errors in the center of the fields ($t = 0, 1, \ldots, P-1$).

$$n_{\mathrm{top}}[c,t] = \sqrt[2]{\frac{\sum_{(i,j)\in \mathrm{TOP}} n[c,i,j,t]^2 w[i,j]}{\sum_{(i,j)\in \mathrm{TOP}} w[i,j]}},$$

$$n_{\mathrm{bot}}[c,t] = \sqrt[2]{\frac{\sum_{(i,j)\in \mathrm{BOT}} n[c,i,j,t]^2 w[i,j]}{\sum_{(i,j)\in \mathrm{BOT}} w[i,j]}}.$$

For instance, a ITU-T H.263 video codec typically makes the output frames progressive by repetition of one of the two input fields. In that case, the color errors of the field that has not been coded can be very large whereas the edginess indicator has already accounted for the progressiveness of the frames. Therefore, the minimum of both field indicators is taken as frame indicator $n[c,t]$.

A.6. Aggregation of the frame indicators to the sequence level

A.6.1. Aggregation of the luminance edginess indicators

Transmission errors can be localized in time and still have a big impact on the DMOS. In order to reflect this, the edginess frame indicators are aggregated over time using a Lebesgue-7 weight.

$$E = \sqrt[7]{\frac{\sum_{t=0,1,\ldots,P-1} e[t]^7}{P}}.$$

For the prediction of the DMOS_P, a deadzone is introduced around zero.

$$E' = \begin{cases} 0 & \text{if } E < 7, \\ E - 7 & \text{otherwise.} \end{cases}$$

A.6.2. Aggregation of the decorrelation indicators

The decorrelation indicator that compensates the edginess indicator and normalized color error indicator, is likewise aggregated using a Lebesgue-7 weight.

A.6.3. Aggregation of the color indicators

The normalized color errors are linearly averaged over time.

$$N[c] = \sqrt[7]{\frac{\sum_{t=1,2,\ldots,P-1} n[c,t]^7}{P-1}}.$$

A.6.4. Production of the predicted DMOS

The DMOS_P is predicted as a linear combination of indicators ($c = 2$ corresponds to the "red" color difference component),

$$\mathrm{DMOS}_P = 3.95E' + 0.74N[2] - 0.78D - 0.4,$$

$$\mathrm{DMOS}'_P = \begin{cases} 0 & \text{if } \mathrm{DMOS}_P < 0, \\ 85 & \text{if } \mathrm{DMOS}_P > 85, \\ \mathrm{DMOS}_P & \text{otherwise.} \end{cases}$$

References

[1] T. Alpert, V. Baroncini, D. Choi, L. Contin, R. Koenen, F. Pereira, H. Peterson, Subjective evaluation of MPEG-4 video codec proposals: Methodological approach and test procedures, Signal Processing: Image Communication 9 (1997) 305–325.

[2] J.G. Beerends, Modelling cognitive effects that play a role in the perception of speech quality, in: DEGA, ITG, EURASIP (Eds.), Speech Quality Assessment, Bochum, November 1994, pp. 1–9.

[3] J.G. Beerends, J.A. Stemerdink, A perceptual speech quality measure based on a psychoacoustic sound representation, J. Audio Eng. Soc. 42 (March 1994) 115–123.

[4] S. Daly, The visible differences predictor: an algorithm for the assessment of image fidelity, in: A.B. Watson (Ed.), Digital Images and Human Vision, Part III, Measurement and Prediction of Visual Quality, MIT Press, London, England, 1993, Chapter 14, pp. 179–206.

[5] A.P. Hekstra, J.G. Beerends, R.H. Koenen, F.E. de Caluwe, A method, an arrangement, an ASIC and their use for objective assessment of video quality, European Patent EP0940996, March 1998.

[6] http://www.vqeg.org

[7] ISO-IEC, Coding of moving pictures and associated audio, International Standard 13818, 1995.

[8] ITU-R, Encoding parameters of digital television for studios, Recommendation 601, 1992.

[9] ITU-T, Video coding for low bit rate communication, Recommendation H.263, March 1996.

[10] ITU-T Study Group 12, Review of validation tests for objective speech quality measures, Document COM 12-74, March 1996.

[11] ITU-T, Objective quality measurement of telephone-band (300–3400 Hz) speech codecs, Recommendation P.861, August 1996.

[12] ITU-T Study Group 12, Final report from VQEG on the validation of objective models of video quality assessment, Geneva, Temporary Document 8 (WP2/12), May 2000.

[13] ITU-T Rec. P.862, Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs, International Telecommunication Union, Geneva, Switzerland, February 2001.

[14] S. Lubin, The use of psychophysical data and models in the analysis of display system performance, in: A.B. Watson (Ed.), Digital Images and Human Vision, Part III, Measurement and Prediction of Visual Quality, MIT Press, London, England, 1993, Chapter 13, pp. 161–178.

[15] MPEG AOE Group, MPEG-4 testing and evaluation procedures document, ISO/IEC JTC1/SC29/WG11, Document N999, Tokyo MPEG Meeting, July 1995.

[16] L.A. Olzak, J.P. Thomas, Seeing spatial patterns, in: L. Kaufman, K.R. Boff, J.P. Thomas (Eds.), Handbook of Perception and Human Performance, Section II, Basic Sensory processes I, Wiley, New York, 1986, Chapter 7.

[17] A.W. Rix, M.P. Hollier, The perceptual analysis measurement system for robust end-to-end speech quality assessment, IEEE ICASSP, June 2000.