Vision Model Based Impairment Metric to Evaluate Blocking Artifacts in Digital Video

Zhenghua Yu, *Member IEEE*, Hong Ren Wu, Stefan Winkler, and Tao Chen, *Student Member IEEE*

Z. Yu was with the School of Computer Science and Software Engineering, Monash University, Clayton Campus, VIC 3800, Australia. He is now with Motorola Australian Research Centre, Locked Bag 5028, Botany NSW 1455, Australia. (Email: zhyu@ieee.org).

H. R. Wu is with the School of Computer Science and Software Engineering, Monash University, Clayton Campus, VIC 3800, Australia (Email: hrw@mail.csse.monash.edu.au). **Correspondence** to Associate Professor Dr. H. R. Wu.

S. Winkler was with the Signal Processing Laboratory, Swiss Federal Institute of Technology, 1015 Lausanne, Switzerland. He is now with Genimedia, EPFL - PSE, 1015 Lausanne, Switzerland. (Email: swinkler@genimedia.com).

T. Chen was with the School of Computer Science and Software Engineering, Monash University, Clayton Campus, VIC 3800, Australia. He is now with Sarnoff Corporation, Princeton NJ 08543-5300, USA. (Email: tchen@sarnoff.com).

Abstract

In this paper, investigations are conducted to simplify and refine a vision model based video quality metric without compromising its prediction accuracy. Unlike other vision model based quality metrics, the proposed metric is parameterized using subjective quality assessment data recently provided by the Video Quality Experts Group (VQEG). The quality metric is able to generate a perceptual distortion map for each and every video frame. A Perceptual Blocking Distortion Metric (PBDM) is introduced which utilizes this simplified quality metric. The PBDM is formulated based on the observation that blocking artifacts are noticeable only in certain regions of a picture. A method to segment blocking dominant regions is devised, and perceptual distortions in these regions are summed up to form an objective measure of blocking artifacts. Subjective and objective tests are conducted, and the performance of the PBDM is assessed by a number of measures such as the Spearman rank order correlation, the Pearson correlation and the average absolute error. The results show a strong correlation between the objective blocking ratings and the mean opinion scores on blocking artifacts.

Keywords

Blocking artifacts, digital video quality and impairment metrics, vision model

I. Introduction

International standardization activities in the area of digital video coding have resulted in a series of international video and associated audio coding standards, and have led to a proliferation of applications in visual communications, digital television, multimedia computing, etc. Various video coding standards have adopted a motion compensated hybrid of temporal DPCM[1] and block DCT[2] (hybrid MC/DPCM/DCT for short) algorithms [1]. The coding algorithm exploits statistical and psychovisual redundancies of input video sequences to achieve a low bit rate, which may cause visible coding distortions in reconstructed sequences due to the lossy coding nature [2, 3]. In order to evaluate, monitor and improve the coding system performance, it is imperative to develop quantitative digital video quality/impairment metrics.

In the evaluation of video systems, a widely used objective quality metric is the Peak Signal-to-Noise Ratio (PSNR). However, it is widely acknowledged that PSNR does not always correlate well with perceived picture quality [4]. The limitation of PSNR becomes

[1] Differential Pulse Code Modulation

[2] Discrete Cosine Transform

more apparent when applied to the evaluation of digital video quality, where distortions such as blocking and ringing are highly structured and differ fundamentally from those in analog video systems. There has been an obvious lack of accurate and widely accepted objective quality/impairment metrics. The only reliable assessment methods have been subjective experiments, which are formally defined in ITU-R Recommendation BT.500 [9]. However, subjective testing is very expensive and time-consuming.

People have long been investigating objective quality assessment methods [5–8]. The first quantitative measure of video quality was proposed, to the best of our knowledge, by Lukas and Budrikis in 1982 [5]. This field of research has become very active recently, as reflected by the number of quality metrics proposed [11–22] as well as the VQEG activities, which represent international efforts towards the standardization of objective video quality metrics [23].

Amid the advances of video quality metric research, a very important branch is the work on vision model based metrics (see e.g. [24]). Human vision research has provided crucial information on the structure and the working mechanisms of the Human Visual System (HVS), which has been adopted to design quality metrics [13–16]. Most current psychovisual quality metrics are based on multichannel vision models [25].

Digital video coding distortions have been well understood and classified [2,3], e.g. blocking, ringing, blurring, etc. For many applications it is highly desirable not only to compute an overall distortion measure, but also to identify the type of occurring distortions and to quantify the quality degradation caused by each type. This allows a more detailed analysis and tuning of the system performance. Among digital video coding distortions, blocking artifacts are of particular importance. They are directly linked to block based coding algorithms and thus represent a major type of distortion in most existing compression systems [26].

Although blocking impairment metrics have already been investigated by a number of researchers, their focus so far has been on still images [27–29]. A blocking impairment metric for video sequences is proposed in [30], but it acts only as a building block for a single-ended quality metric rather than a stand-alone blocking impairment metric, and it is not based on a vision model. In general, current achievements in vision research and quality metrics have not yet been reflected in the quantification of digital video blocking artifacts.

While various kinds of distortions contribute to the degradation of video quality in practical systems, certain types of distortions dominate depending on the compression scheme as well as the pre- or post-filtering processes. It has also been observed that blocking artifacts only dominate in certain regions of coded sequences [31]. It can be expected that distortions in these regions affect the perceived blockiness of the sequence. Therefore, summing up distortions in these blocking dominant regions could serve as a measure of perceived blockiness. Based on this observation, a perceptual blocking distortion metric for block based transform coded digital video is proposed. In this metric, blocking dominant regions are segmented after the spatio-temporal decomposition, and perceptual distortions in these regions are summed up to form an objective measure of blocking artifacts. The decomposition and the distortion measurement are based on a vision model which takes into account the multichannel structure of the HVS, spatio-temporal contrast sensitivity and pattern masking.

The paper is organized as follows. Previous research on vision model based quality metrics is reviewed, and a spatio-temporal distortion metric is introduced in Section II. The model used in this paper is a simplified and refined model, which can be traced back to the work of [10, 13, 14]. Section III presents the proposed blocking distortion metric, while details of the method to segment blocking dominant regions are discussed in Section IV. Section V describes the methods used in subjective and objective tests and presents experimental results. Section VI concludes the paper.

II. A Vision Model Based Video Quality Metric

Modeling human vision has long been a challenging research topic [24,32], as the human visual system is extremely complex. In this paper, most attention is focused on the pattern sensitivity aspect of human vision, upon which a quality metric is built.

Psychovisual experiments to investigate the sensitivity of the human visual system to spatial and temporal patterns can be classified into detection and discrimination tests [33]. In the former, the threshold for detecting the presence of a particular pattern is examined, while in the latter the threshold for detecting differences between two test patterns is studied. Depending on the strength of the stimulus, these experiments can be further categorized into threshold tests and suprathreshold tests [34].

Typical psychovisual experiments use simple synthetic patterns as stimuli. In most experiments only a small group of assessors are employed, therefore the data are usually sparse, and it is often hard to compare results obtained by different researchers under different conditions. These problems have motivated the recent Modelfest efforts to provide a reliable and general data set for common psychovisual stimuli [35].

It is important to note that most experiments in vision research are threshold experiments, while in the application considered here, i.e. assessing natural video scenes over a wide quality range, both threshold vision and suprathreshold vision are involved. There is also a significant difference between the tasks of subjects in the respective subjective tests. In vision research, assessors typically are asked for qualitative judgements (yes/no), while in quality assessment they need to rate the stimuli quantitatively using quality scales. In recognition of the above-mentioned differences and the need for reliable subjective data, the metric devised in this paper adopts the basic architecture of the HVS as revealed by threshold experiments, while its vision model is optimized for video quality assessment with different parameters. This parameterization approach differs from other vision model based quality metrics as will be explained later.

Another consideration in the design of a metric is its computational complexity. Vision models are computationally very expensive. However, from a practical point of view, it is desirable to have a low-complexity video quality metric. Therefore, studies are carried out to reduce the computational complexity of the proposed quality metric without significantly compromising its performance.

The quality metric used in this paper is based on the Teo and Heeger model [10], which was proposed mainly to explain pattern masking. The major building blocks of the model are a steerable pyramid decomposition [36], contrast gain control, detection and pooling [10]. Temporal filters and Contrast Sensitivity Function (CSF) filters were added to the model by van den Branden Lambrecht in the NVFM (Normalization Video Fidelity Metric) [13]. The model was extended to the PDM (Perceptual Distortion Metric) for color video by Winkler [14]. In the PDM, the contrast gain control stage was changed to Watson and Solomon's model [37]. The color space conversion, the perceptual decomposition and the contrast gain control characteristics of human vision are discussed in the following subsections.

A. Color Perception

Various investigations on how color information is coded in the HVS have been conducted [38,39]. Psychological and physiological experiments have revealed that there exist three main color channels [38]. Furthermore, some pairs of hue can coexist in a single color sensation while others cannot, which has led to the development of the theory of opponent colors. Winkler's PDM adopts a pattern-color separable opponent colors space [39], whose three principal color components are Black-White (B-W), Red-Green (R-G) and Blue-Yellow (B-Y).

Human vision has the highest acuity in the black-white pathway, which is one of the properties used in color TV system design, where a broader bandwidth is allotted to the luminance component. Winkler has conducted research on the influence of the choice of color space on the performance of vision models [40]. It has been shown that it is possible for the vision model to work on the luminance (Y) component only without a dramatic degradation in prediction accuracy [40]. Therefore, the metric introduced in this paper uses only the luminance component. The advantage of this approach is that the computational load is reduced by as high as two thirds. If the prediction accuracy is of higher priority, the metric can be extended to a three-channel color space as done in [14].

B. Spatio-temporal Contrast Sensitivity

It is well known that the visual system processes information by contrast rather than absolute light level. There exist several mathematical contrast definitions [41,42]. Generally speaking, contrast is proportional to the relative magnitude of the stimulus luminance. Subjective experiments have been designed to measure contrast sensitivity, i.e. the inverse of the contrast threshold, to different test patterns (e.g. with varying spatial/temporal frequencies).

Most early research on contrast sensitivity resulted in single channel vision models, such as the pioneering work by Schade [43]. While these single channel models perform well with simple patterns, they fail in more complex mixture pattern experiments. In recognition of the limitations of single channel models, and motivated by other experimental findings such as pattern adaptation, multi-channel vision models have been proposed and proven to be an important theoretical tool in vision science [44, 45]. Consequently multi-channel decompositions have been used in a number of most recent HVS-based image and video quality metrics [13–16].

Psychophysical studies have shown evidence of the existence of several channels tuned to different temporal frequencies, spatial frequencies and orientations [32]. This multi-channel system can be simulated by filter banks using digital signal processing techniques. In many existing quality metrics, the visual system is modeled by spatio-temporally separable filter banks, with which a reasonably good approximation has been achieved [13, 14].

It is generally believed that there exist two temporal mechanisms, one being transient and the other sustained [46]. Consequently, two temporal filters have been employed in [13, 14]. However, most visual detail information is carried by the sustained channel. This is partially the reason why some early single channel models still work well under certain simple test conditions. Objective tests conducted with the PDM revealed that the majority of distortions existed in the sustained channel. Therefore only the sustained temporal channel is used in this paper. As will be demonstrated in the following sections, it is still possible to achieve a high prediction accuracy with a single temporal filter, with the advantage that the computational complexity is halved.

Compared with temporal mechanisms, more research efforts have been put into modeling the spatial mechanisms of the HVS. For the achromatic visual pathway, experiments have revealed that there are several spatial frequency and orientation tuned channels with frequency bandwidths of one to two octaves and orientation bandwidths between 30 and 60 degrees [47, 48], suggesting 4 to 7 spatial frequency bands.

A number of spatial decomposition filters have been employed in existing vision models, e.g. the Laplacian pyramid [15] or the DCT [16]. Two other decomposition transforms have been suggested by Teo and Heeger, i.e., a hex-QMF filter bank and a steerable pyramid transform [36], where the latter is preferred because it has the advantage of being rotation-invariant and self-inverting while minimizing the amount of aliasing in the subbands [10, 49]. Therefore the steerable pyramid transform is adopted in this paper. Fig. 1 illustrates an example of the partitioning of the spatial frequency plane by the steerable pyramid transform. It shows four orientation bands (tuned to 0, 45, 90 and 135 degrees) at three different scales, one (isotropic) low-pass band and one (isotropic) high-pass band.

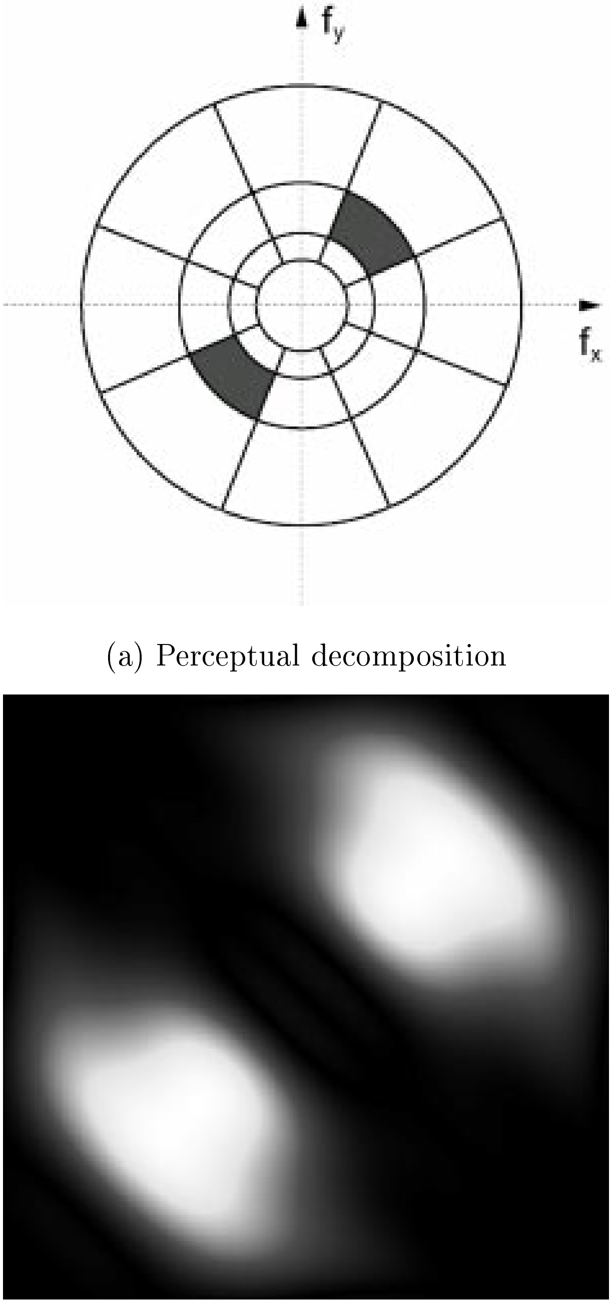
In vision research, the sensitivity for the detection of spatial patterns at various spatial frequencies is characterized by spatial CSFs. When applied in quality metrics, the CSF can be simulated by digital filtering with the respective filter response. Since the spatial filter bank already decomposes the sequences into several frequency levels, CSF filtering can be implemented by multiplying every subband with the proper CSF coefficient [14]. Overall, this constitutes a coarse, but very fast approximation of the desired CSF filter response.

When implementing the steerable pyramid at image boundaries, care should be taken when pixels outside of the image are needed for filtering. There are some investigations in how to extend the pixel value to cover areas outside the boundaries for wavelet transforms [50]. In this paper, a simple approach is adopted using only the central region of the frame and discarding boundary pixels for calculations in the subsequent stages. A similar approach has been adopted in the PDM, but larger parts of the boundaries are cropped in the proposed metric so that the boundaries of the lowest spatial-frequency subbands (which have been downsampled several times and therefore have the lowest spatial resolution) are properly manipulated.

C. Contrast Gain Control

Another essential element in vision modeling is pattern masking. When a target pattern and a masking pattern are mixed, the perception of the target pattern will differ depending on the characteristics of the masking pattern. A number of prominent contrast gain control models of pattern masking have been proposed by Foley [51], Teo and Heeger [10], and Watson and Solomon [37]. The latter two are based on multi-channel models, and they are "image-driven" [37], which explains their popularity in video quality metrics [13, 14].

The contrast gain control stage in a multi-channel vision model follows the spatio-temporal decomposition and the CSF filtering. In this paper Teo and Heeger's contrast



(a) Perceptual decomposition

(b) Filter tuned to 45 degrees

Fig. 1. Illustration of the partition of the spatial frequency plane by the steerable pyramid transform.

gain control model is used [10], which consists of an excitatory-inhibitory stage and a normalization stage. In the excitatory visual pathway, the input coefficient values are locally squared. In the inhibitory visual pathway pooling is performed over the squared subbands. The ratio of the excitatory and inhibitory energies is calculated in the normalization stage. Let $A(j, f, \theta, x, y)$ be a coefficient after the steerable pyramid decomposition and the CSF filtering (cf. Subsection II-B), where j, f, θ, x, y represent frame number, spatial frequency, orientation, horizontal and vertical spatial location, respectively [10]. The squared and normalized output $R(j, f, \theta, x, y)$ is then computed as:

$$R(j, f, \theta, x, y) = \sum_i k_i \frac{(A(j, f, \theta, x, y))^2}{\sum_\phi (A(j, f, \phi, x, y))^2 + \gamma_i^2} \;, \tag{1}$$

where k_i is an overall scaling constant, γ_i^2 is a saturation constant, $i = 1, 2, 3, 4$, and $\phi = 0, 45, 90, 135$ degrees. With this formula, masking over orientation subbands at the same spatial frequency level is considered. Because an exponent of 2 is used in both the excitatory and the inhibitory paths, four contrast channels (i.e. four pairs of k_i and γ_i) are needed to overcome the rapid saturation of each channel [49]. The dynamic range of the acceptable input is limited by the value of γ_i, while a large value of $A(j, f, \theta, x, y)$ will saturate the output of the contrast gain control model.

When implementing the contrast gain control model together with the aforementioned steerable pyramid transform, the output of the steerable pyramid will be the input to the contrast gain control model. For the band-pass bands, the input of the contrast gain control model is well within the desired dynamic range of the model. This is not the case for the low-pass band, however, which has passed through several stages of pyramid decomposition and down-sampling. Because the DC component is not separated from the coefficients before the pyramid decomposition, the DC energy accumulates in the low-pass band. After the decomposition, coefficients in that subband become so large that they are far beyond the desired dynamic range of the contrast gain control model. Therefore, the mean value is subtracted from each pixel of the respective frame in the spatial domain before the decomposition, in order to prevent the accumulation of the DC energy into the low-pass band.

D. A Vision Model Based Video Quality Metric

Following the discussions in the above subsections, a vision model based video quality metric is proposed. The architecture of the metric is illustrated in Fig. 2. The metric takes two video sequences, one original and one processed, as inputs. In the case of interlaced sequences, odd fields of the luminance (Y) component are extracted and passed to the temporal filter. The temporal filter was designed by van den Branden Lambrecht [13] to fit the frequency response of the low pass filter proposed in [46]. The filter is physically implemented by a first order Infinite Impulse Response (IIR) filter, which has a low delay and low storage requirement. After the temporal filtering, the mean pixel value of every frame is calculated for both the original and the processed sequences and is then subtracted from each pixel coefficient. This stage is followed by the steerable pyramid decomposition [36]. Input sequences are decomposed into six frequency levels including an isotropic low-pass (LP) level, four band-pass (BP) levels with four orientations each centered around 0^o, 45^o, 90^o and 135^o, and an isotropic high-pass (HP) level. After the decomposition, the output of each subband is multiplied by the corresponding CSF coefficient. The purpose of this weighting operation is to match the overall gain of the metric with the contrast sensitivity curve of the HVS. In the subsequent contrast gain control stage the responses in every subband are squared and normalized according to Eq. 1. Up to this stage the original sequence and the processed sequence are handled independently. The detection and pooling stage then integrates the data from the different bands of both sequences according to a summation rule. This stage simulates the integration process of the visual cortex. In the metric, it is modeled by a squared error norm of the difference between the sensor outputs of the original sequence $R_o(j, f, \theta, x, y)$ and the sensor outputs of the processed sequence $R_p(j, f, \theta, x, y)$ using:

$$\Delta R = \frac{\sum_{j,f,\theta,x,y} |R_o(j, f, \theta, x, y) - R_p(j, f, \theta, x, y)|^2}{N}, \qquad (2)$$

where ΔR is a measure of perceptual distortion, N is the number of frames, and the other notations follow Eq. 1. In this stage, the perceptual distortion map of each frame can also be generated after summing over spatial frequency and orientation subbands. This distortion map is of the same resolution as the original frame. Each pixel value represents

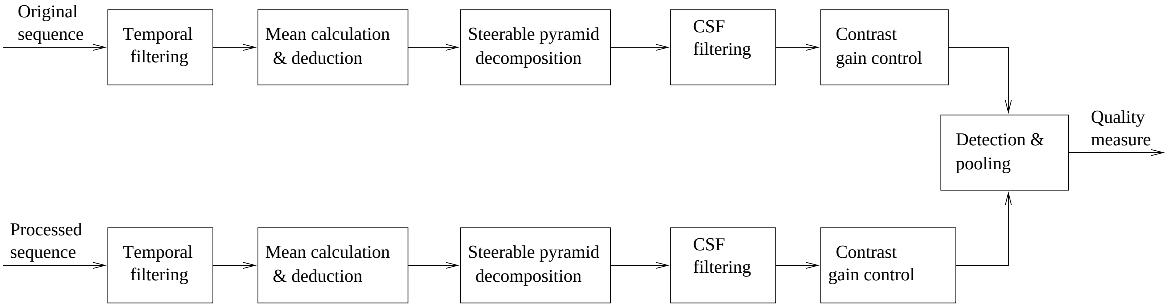


Fig. 2. Block diagram of the proposed quality metric.

the perceptual distortion at that spatial location.

E. Relationship with Other Metrics

The proposed metric is based on the NVFM and the PDM, therefore there are many similarities among these three metrics. They have a similar architecture, consisting of temporal filtering, subband decomposition, CSF filtering, contrast gain control and pooling stages. Major differences are listed in Table I. Because the proposed metric does not require color space conversion, only has one temporal channel, and discards the largest boundary region, its computational complexity is much lower than that of the other two. With the parameterization approach that will be discussed in the following subsection, the performance of the proposed metric is still comparable to the other metrics.

F. Metric Parameterization

The Video Quality Experts Group (VQEG) has conducted extensive subjective tests on video quality using both 50Hz and 60Hz video sequences [23]. The tests in this paper are conducted using 60Hz sequences only. Nevertheless the approach can be easily extended to the design of metrics for 50Hz sequences using a different temporal filter. Ten 60Hz scenes and sixteen Hypothetical Reference Circuits (HRCs) were used in the VQEG test. A subset of the VQEG 60Hz subjective test data has been used to parameterize the proposed metric.

The task is to obtain six CSF coefficients (c_i) for six frequency levels, and four pairs of k_i and γ_i. A search algorithm has been devised, which is adapted to the architecture of the PC cluster computing facility used. Seven HRCs and ten scenes were used in

TABLE I

Comparison between the NVFM, the PDM and the Proposed Metric.

Models	NVFM [13]	PDM [14]	Proposed
Number of color channels	1	3	1
Number of temporal filters	2	2 for luminance, 1 for chrominance	1
Max. order of temporal filter	2	4	1
Mean deduction	No	No	Yes
Contrast gain control model	Teo and Heeger [10]	Watson and Solomon [37]	Teo and Heeger [10]
Unused boundary pixels	0	14	56
Model parameterization	Threshold vision	Threshold vision	VQEG

the optimization. At first, these sequences were passed through the steerable pyramid decomposition without temporal filtering, and the subband coefficients of frames 30 and 120 were stored for the optimization. With the objective metric, the summed distortion of a frame can be obtained by applying CSF filtering, contrast gain control and pooling on the stored subband coefficients of that frame. Therefore, for any set of c_i, k_i and γ_i coefficients, a collection of objective scores can be obtained which are associated with the HRCs/scenes used. The subjective scores of these HRCs/scenes were taken from the VQEG subjective data. The Spearman rank order correlation was used as the cost function to be maximized. It is a measure for the agreement between the rank orders of objective and subjective scores [53]. Given the bivariate sample (MOS, VQR) (see Subsection II-G), the mutual dependence of MOS and VQR can be assessed by the Spearman correlation coefficient r_s:

$$r_s = 1 - \frac{6 \sum_l D_l^2}{N(N^2-1)}, \tag{3}$$

where D_l is the difference between the ranked pairs, l is the index of the data sample and N is the number of data samples. The optimization problem can be formulated as follows. Let c_i $(i = 1, 2, ..., 6)$, k_i $(i = 1, 2, 3, 4)$, and γ_i $(i = 1, 2, 3, 4)$ represent the model coefficients in the objective model, and $\mathbf{C}, \mathbf{K}, \mathbf{\Gamma}$ be the vector representation of the corresponding set of coefficients. As an example, $\mathbf{K} = [k_1, k_2, k_3, k_4]^T$. The problem is to find the optimal $\mathbf{C}_{opt}$, $\mathbf{K}_{opt}$, and $\mathbf{\Gamma}_{opt}$ that maximize the Spearman rank order correlation $r_s(\mathbf{C}, \mathbf{K}, \mathbf{\Gamma})$. The optimization algorithm can be described as follows using pseudocode.

Algorithm 1: Model optimization

```
(1) initialize $\mathbf{C}$ and $\mathbf{\Gamma}$ to those of the NVFM [13], $\mathbf{K}$ to $[1, 1, 1, 1]^T$ and $current\_r_s = 0$
(2) **repeat**
(3)     let $previous\_r_s = current\_r_s$
(4)     fix $\mathbf{C}, \mathbf{K}$ and search for optimal $\mathbf{\Gamma}$ using *Algorithm* 2 (see below)
(5)     fix $\mathbf{C}, \mathbf{\Gamma}$ to the optimal sets from (4) and search for optimal $\mathbf{K}$ using *Algorithm* 2
(6)     fix $\mathbf{K}, \mathbf{\Gamma}$ to the optimal sets from (5), search for optimal $\mathbf{C}$ using *Algorithm* 2, and record the highest $r_s$ as $current\_r_s$
(7) **until** $current\_r_s - previous\_r_s < T$
(8) output the $\mathbf{C}, \mathbf{K}, \mathbf{\Gamma}$ associated with the highest $r_s$ as $\mathbf{C}_{opt}$, $\mathbf{K}_{opt}$, $\mathbf{\Gamma}_{opt}$
```

As part of *Algorithm* 1, a search algorithm has been devised to find an optimal set of coefficients while the other two are fixed. For each coefficient being optimized and at each iteration, the algorithm selects from up to four candidate data points for the optimal coefficient. As an example, let c_i be a coefficient to be determined. The candidate coefficient set can be denoted as $\{c_{ij} : j = 1, 2, 3, 4\}$, where c_{ij} is the jth candidate for coefficient c_i, and $c_{i1} < c_{i2} < c_{i3} < c_{i4}$. The candidate set is constructed around a seed coefficient[3] with a known search step size. A coarse search is conducted in the first iteration, i.e. a large search step size $\Delta c_{ij,j+1} = \|c_{ij} - c_{ij+1}\|$ is used. If a locally maximal coefficient vector $\tilde{\mathbf{C}}$ has been found, a smaller search step will be used to search around $\tilde{\mathbf{C}}$ in the next iteration until the improvement in correlation is below the threshold T. The locally maximal coefficient vector $\tilde{\mathbf{C}}$ is defined as the coefficient vector in the candidate

[3] The initial seed coefficient is obtained from *Algorithm* 1 and passed to *Algorithm* 2, e.g., the initial Γ in Step (4) of *Algorithm* 1 is a seed coefficient vector.

set that meets the following two conditions:

a) it yields the highest correlation, and

b) none of its component coefficients $\tilde{c}_{ij}$ is at the boundary of the candidate set (i.e. $j \neq 1$ and $j \neq 4$ for coefficient c_i).

After one iteration, human interaction is required to review the search step size, to decrease the step size when needed, and to reduce the number of candidates for a particular coefficient if the previous iteration demonstrates that varying this coefficient has less influence on the correlation than other coefficients. Sometimes condition (a) could be met, but some component coefficients would fail to satisfy condition (b). In this case only these component coefficients will be varied in the next iteration, while the other component coefficients are fixed at the previous optimal value.

The search algorithm can be summarized as follows.

Algorithm 2: Search for an optimal coefficient vector among a candidate set.

```
initialize the seed coefficient vector $\mathbf{S}$ (passed from $Algorithm$ 1), search step size $\Delta c$ and set $current\_r_s = 0$.
repeat
  let $previous\_r_s = current\_r_s$
  repeat
    construct the candidate coefficient set $\mathcal{C}$ around $\mathbf{S}$ with step size $\Delta c$
    compute the corresponding $r_s$ for every vector $\mathbf{C}$ in $\mathcal{C}$
    find the vector $\tilde{\mathbf{C}}$ which yields the highest $r_s$ (denoted as $\tilde{r}_s$)
    let $\mathbf{S} = \tilde{\mathbf{C}}$
  until $\tilde{\mathbf{C}}$ is a locally maximal coefficient vector
  let $current\_r_s = \tilde{r}_s$ and $\Delta c = \Delta c/2$
until $current\_r_s - previous\_r_s < T$
output $\tilde{\mathbf{C}}$ as the optimal coefficient vector
```

T was set to 0.0001 in the experiments. There were three to four iterations with *Algorithm* 2 before the optimization converged, and two iterations with *Algorithm* 1. Clearly the coefficients obtained are not globally optimized. However, a high correlation can still be achieved with this method. The optimized coefficients are reported in Table

II. Among the variable coefficients, k_i has little influence on the correlation. As a result, all k_i coefficients were fixed at 1.

TABLE II

Metric Coefficients

c_i (CSF)	0.41, 1.25, 1.2, 0.4097, 0.083, 0.001
k_i	1, 1, 1, 1
γ_i	4.4817, 7.3891, 12.1825, 54.5982

G. Metric Performance

The performance of the proposed quality metric has been tested using the 60Hz VQEG sequences. Evaluation metrics used include:

1. The Spearman rank order correlation (Spearman for short) between the objective Video Quality Rating (VQR) and the Mean Opinion Score (MOS), which is metric 3 in the VQEG objective test plan [54].

2. The Pearson correlation (Pearson-Logistic for short) between the VQR and the MOS with non-linear mapping of the objective model outputs (metric 2 from the VQEG objective test plan [54]). First, the VQR is mapped to the predicted $MOSp$ by a 4-parameter logistic curve, minimizing $(MOS - MOSp(VQR))^2$:

$$MOSp(VQR) = \frac{y_{max} - y_{min}}{1 + e^{-\frac{VQR - \overline{x}}{|\beta|}}} + y_{min}, \tag{4}$$

where $y_{min}, y_{max}, \overline{x}, \beta$ are four parameters to be determined by an optimization procedure. The purpose of the nonlinear regression is to remove any nonlinearities due to the subjective rating process and to facilitate comparison of the models in a common analysis space [54]. Then the Pearson correlation between the $MOSp(VQR)$ and the MOS is calculated by:

$$r_P = \frac{\sum_i (MOSp_i - \overline{MOSp})(MOS_i - \overline{MOS})}{\sqrt{\sum_i (MOSp_i - \overline{MOSp})^2}\sqrt{\sum_i (MOS_i - \overline{MOS})^2}} \tag{5}$$

where $i = 1, 2, ...N$ and N is the number of test trials.

Table III compares correlations of PSNR (VQEG Proponent P0), the original PDM (VQEG Proponent P5), the PDM with two other color spaces, namely CIE $L^*a^*b^*$ and

Y only [40], and the proposed metric, all calculated for the 60Hz VQEG sequences. The PSNR is the benchmark in objective video quality assessment, while the PDM achieves the highest correlation among all VQEG proponents for the 60Hz scenes. Furthermore, the PDM has been tested with different color spaces, with the finding that CIE $L^*a^*b^*$ and CIE $L^*u^*v^*$ outperform other color spaces [40]. It is evident from the table that a high correlation is achieved with the proposed quality metric.

TABLE III

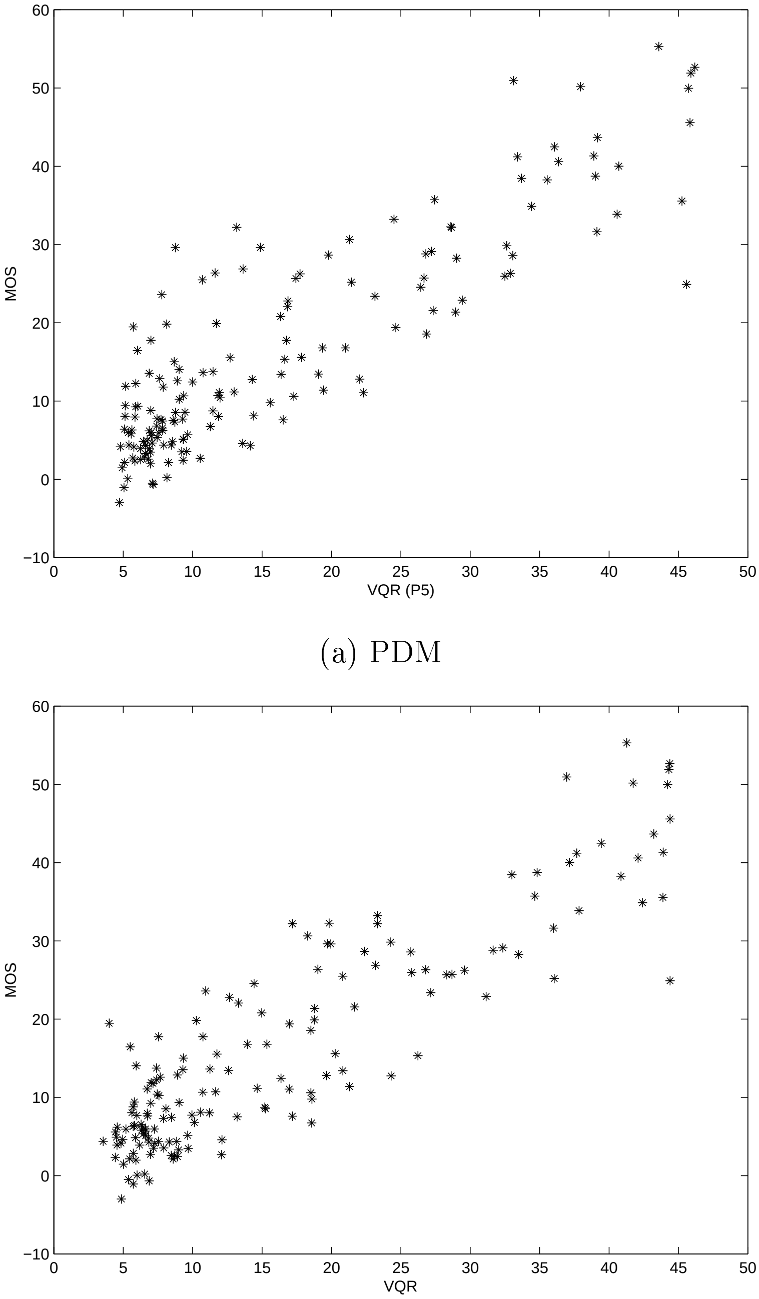
Correlations

Metrics	Spearman	Pearson-Logistic		
		correlation	upper bound	lower bound
PSNR	0.723	0.767	0.825	0.693
Original PDM	0.805	0.875	0.908	0.832
PDM with Y channel	0.753	0.834	0.874	0.782
PDM with $L^*a^*b^*$ color space	0.827	0.891	0.919	0.858
Proposed	0.823	0.892	0.920	0.854

Fig. 3 illustrates the scatter plot of MOS versus the objective score of the PDM and the proposed metric after the logistic fit. These two metrics have similar scatter plots; the major improvement of the proposed metric is that some outliers have been removed. Actually the Spearman rank order correlation between the PDM and the proposed metric is 0.9, while the Pearson correlation between the two metrics after the logistic fit is 0.936. As discussed in Subsection II-E, these two metrics use different temporal filters and different contrast gain control models, the proposed metric only has the luminance channel opposed to the three color channels in the PDM, and the proposed metric has much lower computational complexity than the PDM.

III. Perceptual Blocking Distortion Metric

Blocking artifacts are defined as discontinuities across block boundaries [2]. They are among the most common artifacts in compressed video content today. For accurate perceptual measurements of these artifacts, it is important to recognize that different distortions



(a) PDM

(b) Proposed quality metric

Fig. 3. Scatter plot of VQR versus MOS.

are predominant in different regions of digitally compressed images [31]. For example, ringing artifacts mostly occur around high contrast edges of objects, while blocking artifacts are more noticeable in smooth regions.

Based on these observations, it is assumed that only distortions in the blocking dominant regions will contribute to the measure of blocking artifacts. In the video quality metric proposed in Subsection II-D, a perceptual distortion map can be generated that represents the perceptual distortion at every spatial location. Together with blocking dominant region segmentation, a good measure of blocking artifacts can be obtained by calculating the cumulative perceptual distortion in these regions.

The Perceptual Blocking Distortion Metric (PBDM) presented here integrates the above-mentioned quality metric with a novel blocking dominant region segmentation algorithm. As illustrated in Fig. 4, the PBDM consists of the following stages: temporal filtering, mean calculation and deduction, steerable pyramid decomposition, blocking region segmentation, CSF filtering, contrast gain control and pooling. The PBDM is heavily based on the proposed video quality metric, with two major differences: a blocking dominant region segmentation stage has been added, and at the pooling stage, the sum of the differences between outputs from the original and the processed sequences is calculated over spatial frequency and orientation subbands according to the blocking region map, and then averaged by the number of frames N, in order to obtain the blocking distortion d. Mathematically, this summation can be expressed as

$$d = \frac{\sum_{j,f,\theta,x,y} |R_o(j,f,\theta,x,y) - R_p(j,f,\theta,x,y)|^2}{N}, \forall(x,y) \in \mathcal{B} \tag{6}$$

where $\mathcal{B}$ denotes the set of coefficients in the blocking dominant region, and the other notations follow Eq. 2. After the pooling, d needs to be converted to the Objective Blocking Rating (OBR) on a scale between 1 and 5, corresponding to the five-grade impairment scale [9]. The conversion is performed with the following formula:

$$OBR = \begin{cases} 5 - d^{0.6} & \text{if } d < 4^{1/0.6}, \\ 1 & \text{otherwise.} \end{cases} \tag{7}$$

The saturation threshold $4^{1/0.6}$ is a precaution in case there is a very large d. The exponent 0.6 was derived through experiments to best fit MOS with the OBR. The blocking

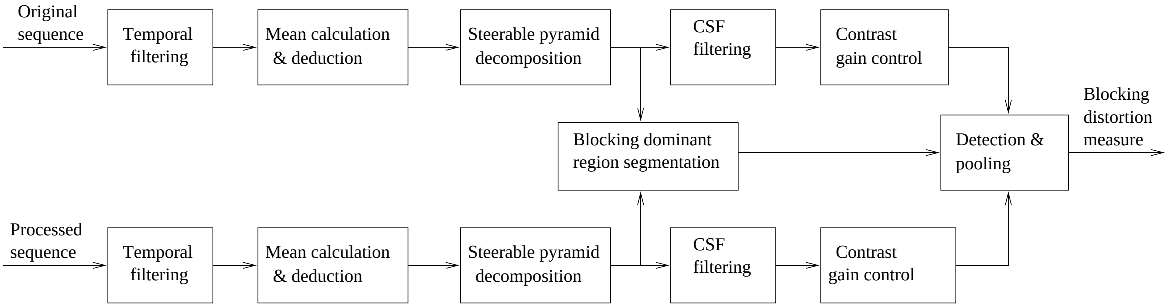

Original sequence
Temporal filtering
Mean calculation & deduction
Steerable pyramid decomposition
CSF filtering
Contrast gain control
Blocking dominant region segmentation
Detection & pooling
Blocking distortion measure
Processed sequence
Temporal filtering
Mean calculation & deduction
Steerable pyramid decomposition
CSF filtering
Contrast gain control

Fig. 4. Block diagram of the perceptual blocking distortion measure.

dominant region segmentation algorithm will be discussed in detail in Section IV.

IV. Blocking Dominant Region Segmentation

It is of great interest to investigate the waveform of blocking artifacts after the spatio-temporal decomposition, since the input frames are decomposed into several frequency and orientation tuned subbands. It can be expected that in some subbands certain types of distortion are more pronounced than they were before the decomposition. For blocking artifacts, the major form of appearance is a sharp edge at block boundaries (block edge), therefore there will be a lot of high-frequency components after the spatial decomposition. Analyzing the subband images, it is observed that blockiness mostly manifests itself in the high-pass subband of the processed sequence, where it appears as additional vertical and horizontal edges after the spatio-temporal decomposition. Therefore, detection of blocking edges could be accomplished by the detection of this kind of waveform, together with the condition that the edges are persistent horizontally or vertically.

The segmentation algorithm consists of the following seven steps. Fig. 6 is used as an example to demonstrate how the algorithm works. It shows one frame of the "Claire" sequence before the decomposition (a), the high-pass subband of that frame after the decomposition (b), a magnified part of the images (c,d) and images after each stage of the segmentation algorithm.

A. Vertical and Horizontal Edge Detection

Edges are detected by fitting pixel coefficients to the waveform shape illustrated in Fig. 5. An edge is composed of several vertically or horizontally adjacent edge points.

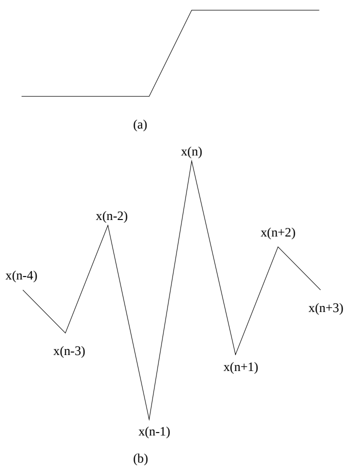


Fig. 5. Gray level of part of one scan line before and after decomposition (illustration only)

The detection algorithm consists of two stages. In the first stage, horizontal and vertical edge points are detected separately. The detection is performed on the one-dimensional waveform. In the second stage, characteristics of six consecutive data points are considered because of the most common 8 × 8 block size. If a certain amount of edge points exist within these six points, combined with other conditions, an edge is found. The two extra data points at the ends of the six-pixel edge may not show up as typical edge points, because after decomposition they are also affected by the neighboring blocks.

Take vertical edge detection as an example and denote $(m, n-1)$ as the coordinate where an edge is to be located. If a vertical edge is to be detected, the edge point detection will first be executed horizontally. At the beginning, the edge data point counter is reset to zero. If a local minimum lies next to a local maximum, it is counted as an edge point. Mathematically, let $x(m, n)$ represent the pixel value at position (m, n). In the mth row, the notation $x(m, n)$ can be rewritten as $x_m(n)$. A typical 1-D waveform is illustrated in Fig. 5(b). If $x_m(n-1)$ and $x_m(n)$ are a pair of local maximum and minimum, then $x_m(n-1)$ is marked as an edge point, the sign of $d_m(n-1, n) = x_m(n) - x_m(n-1)$ is stored, and the counter of edge data points gets increment. The same procedure is applied to the other five data points which lie vertically below $x(m, n-1)$.

In the second stage, the properties of six vertically neighboring data points are considered. The absolute sums $\hat{d}_i = \sum_{j=m}^{m+5} |d_j(i, i+1)|$, $i \in [n-3, n+1]$ are calculated. If the following conditions are met, a vertical edge is found:

- The edge data point counter is larger than four.
- The number of sign changes of $d_j(n-1,n)$ between two vertically adjacent edge data points $(j,n-1)$ and (j,n) is less than two.
- The following inequalities are satisfied:

$$\begin{aligned}
\hat{d}_{n-1} &> T_1, \\
\hat{d}_{n-1} &> \hat{d}_n + T_2, \\
\hat{d}_{n-1} &> \hat{d}_{n-2} + T_2, \\
\hat{d}_{n-2} &> T_1/2, \\
\hat{d}_n &> T_1/2, \\
\hat{d}_{n-3} &< \min(\hat{d}_{n-2}/2, T_1/2 + T_2), \text{ and} \\
\hat{d}_{n+1} &< \min(\hat{d}_n/2, T_1/2 + T_2),
\end{aligned}$$

where T_1 and T_2 are positive thresholds to be determined by experiments.

If a vertical edge is found, the six data points and the other two vertically neighboring points at two ends are marked as vertical edge points. A similar procedure is applied to horizontal edges. Both the reference and the processed sequences are subject to edge detection, and an edge indicator map is generated afterwards. Taking the "Claire" sequence as an example, the detected horizontal and vertical edges in part of the processed and the reference frames are shown in Fig. 6(e,f).

B. Removal of Edges Coexisting in the Original and the Processed Sequences

Edges in the original sequence are due to the scene content rather than blocking artifacts. Therefore only additional edges in the processed sequence need to be retained for blocking region segmentation. If one data point is marked as a vertical edge point in the original sequence, then the data point at the corresponding spatial location in the processed sequence and the other four horizontally neighboring points will be removed from the vertical edge map. Horizontal edge points are processed analogously. Fig. 6(g) illustrates the edges in the example "Claire" frame after this stage.

C. Removal of Short Isolated Edges in the Processed Sequence

If no pixel in an edge appears as the crossing of a horizontal edge and a vertical edge, and the length of the edge is shorter than the block size (i.e. 8 pixels) in the processed sequence, then the edge is too small to be considered as a blocking artifact. It is thus considered as a short isolated edge and should be removed from the edge indicator map. The algorithm scans the indicator map from top to bottom and left to right and removes any isolated edges shorter than 8 pixels in length. The edges in the example "Claire" frame after this stage are shown in Fig. 6(h).

D. Adjacent Edge Removal

Two detected adjacent edges represent only one real edge. Therefore, one of them should be removed from the indicator map. The algorithm searches for any parallel adjacent horizontal or vertical edges and keeps only one edge (the left or the top) of the adjacent pairs in the indicator map. Fig. 6(i) shows the edges in the example "Claire" frame after the adjacent edge removal.

E. Generation of the Blocking Region Map

It is most likely that eight rows or columns surrounding an edge will show up as blocking artifacts. This assumption is the basis for the region generation algorithm. If there is one vertical edge point, this pixel and eight adjacent pixels on the left and right sides of the pixel are classified as part of the blocking region. Similar processing is applied to horizontal edges. The generated blocking region map for the example "Claire" frame is shown in Fig. 6(j).

F. Ringing Region Detection

Strong ringing artifacts will occur along high-contrast edges of objects in encoded scenes, i.e. these edges will induce strong oscillations in the high-pass band of the original sequence. Based on this property, ringing dominant regions can be determined. The detection is performed on the original sequence. For a pixel $x(m,n)$, the inter-pixel differences of its surrounding 5×5 block are calculated. If $\sum_{j=-2}^{2}\sum_{i=-2}^{1}(x(m+j,n+i)-x(m+j,n+i+1))^2+\sum_{j=-2}^{1}\sum_{i=-2}^{2}(x(m+j,n+i)-x(m+j+1,n+i))^2>T_3$, then a strong oscillation region is located, and the 3×3 block surrounding $x(m, n)$ will be marked as a ringing region. The detected ringing regions in the "Claire" frame are shown in Fig. 6(k).

G. Exclusion of Ringing Regions from the Blocking Region Map

The major distortions in the detected ringing regions will be the reconstruction errors of edges, which appear as ringing. Distortions in these regions should not be considered as blocking artifacts, even though the additional edges may still exist in the regions. Therefore, regions dominated by ringing distortion are excluded from the blocking region map.

After these seven steps the final blocking region map is obtained, which represents the regions where blocking artifacts are dominant. Using the "Claire" sequence as an example, the magnified part of the blocking region map is illustrated in Fig. 6(l), which is consistent with subjective observation.

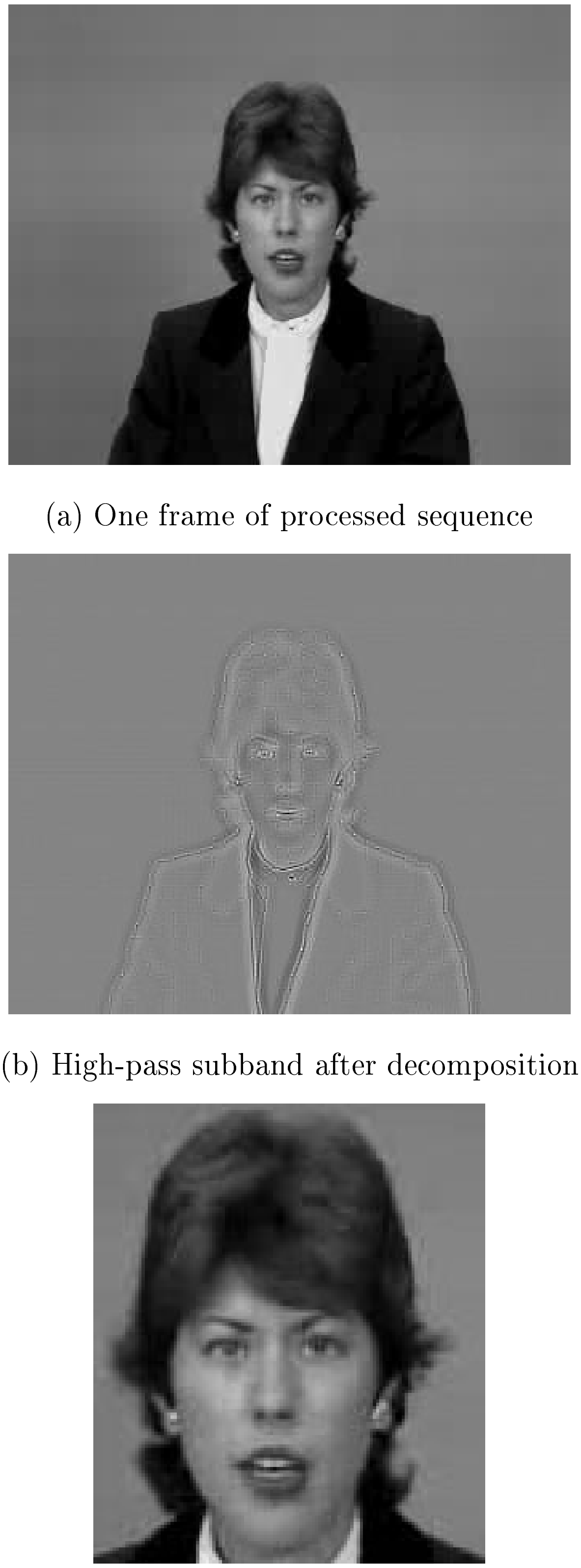
Three thresholds for blocking region segmentation, T_1, T_2 and T_3, were parameterized experimentally. Two video sequences, "Claire" (CIF) and "Car phone" (QCIF) were used for this purpose. The thresholds were adjusted so that blocking regions segmented by the algorithm described above coincided with subjective segmentation. The resulting values are $T_1 = 8$, $T_2 = 1$ and $T_3 = 300$.

V. Experimental Results

Both subjective and objective tests have been conducted. Correlations and prediction errors between subjective and objective data are used to evaluate the performance of the proposed blocking impairment metric.

A. Selection of Test Material

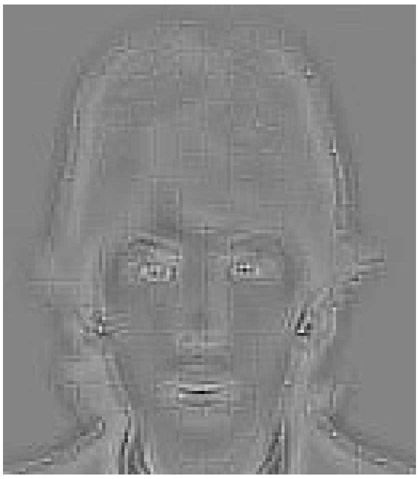
The test scenes were selected from two sources: the ANSI T1A1 data set and the VQEG data set. The video sequences are interlaced with a frame rate of 30 frames per second and a resolution of 720×480 pixels. The lengths of the ANSI and the VQEG sequences are 360 and 260 frames, respectively. Twenty-five test scenes are standardized in ANSI T1.801.01-1995 [55], which are grouped into five categories: (1) one person, mainly head and shoulders, (2) one person with graphics and/or more detail, (3) more than one



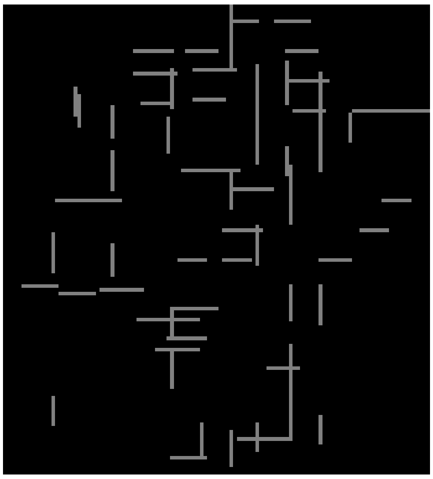
(a) One frame of processed sequence

(b) High-pass subband after decomposition

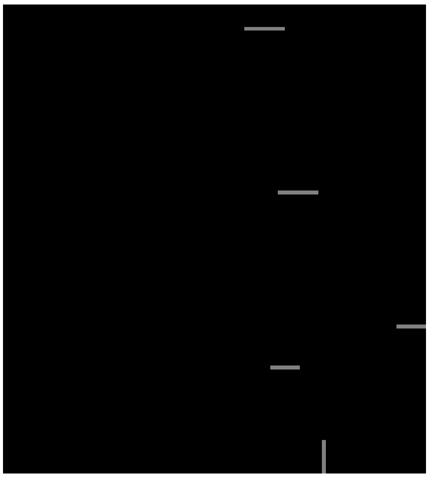
(c) Magnified part of the processed frame



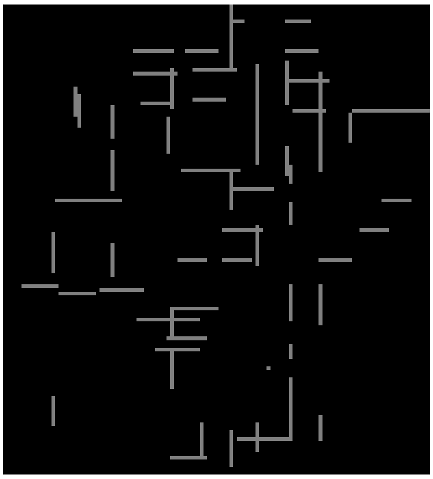
(d) Magnified part of the high-pass subband



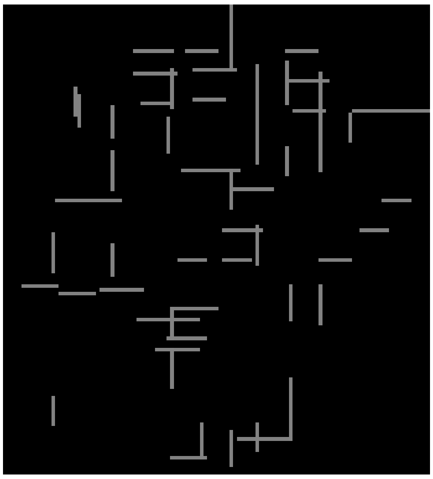
(e) Detected horizontal and vertical edges in the processed sequence



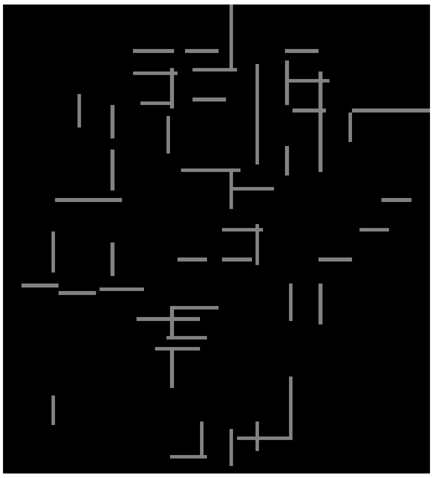
(f) Detected horizontal and vertical edges in the original sequence



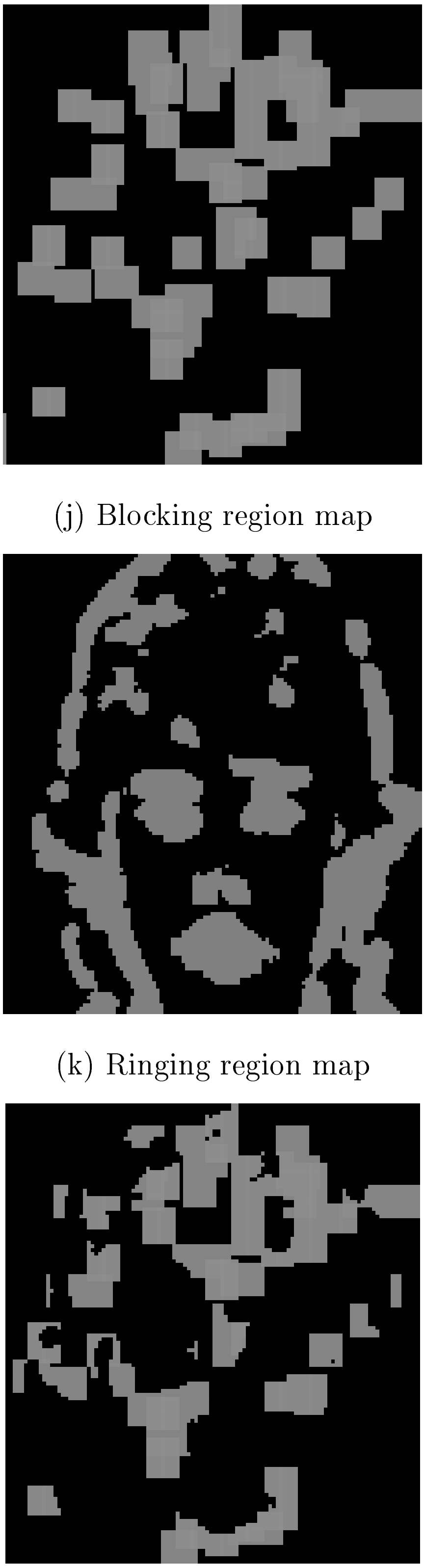
(g) After the removal of edges coexisting in the original and processed sequences



(h) After the removal of short edges



(i) After the removal of adjacent edges



(j) Blocking region map

(k) Ringing region map

(l) After the removal of ringing regions from the blocking region map

Fig. 6. One frame of the “Claire” sequence.

person, (4) graphics with pointing, and (5) high object and/or camera motion. One scene from each of the fivecategories was selected for the evaluation, namely "disgal", "smity1", "5row1", "inspec" and "ftball". The test scenes chosen from the VQEG data set were "Src13 Balloon-pops", "Src14 NewYork 2", "Src15 Mobile & Calendar", "Src16 Betes pas betes" and "Src18 Autumn leaves". Selected single frames from these scenes are shown in Fig. 7.

The following five HRCs were used for the ANSI sequences: [4]

- HRC31, MPEG-2 at 1.4 Mbps,
- HRC32, MPEG-2 at 2 Mbps,
- HRC33, MPEG-2 at 3 Mbps,
- HRC35, MPEG-2 at 768 kbps,
- HRC36, MPEG-2 at 5 Mbps.

The MPEG-2 encoder used is a software implementation of MPEG-2 Test Model 5 (TM5) [56]. Bit rates were selected such that the generated sequences covered the full range of impairments. None of these HRCs were used for parameterizing the PBDM.

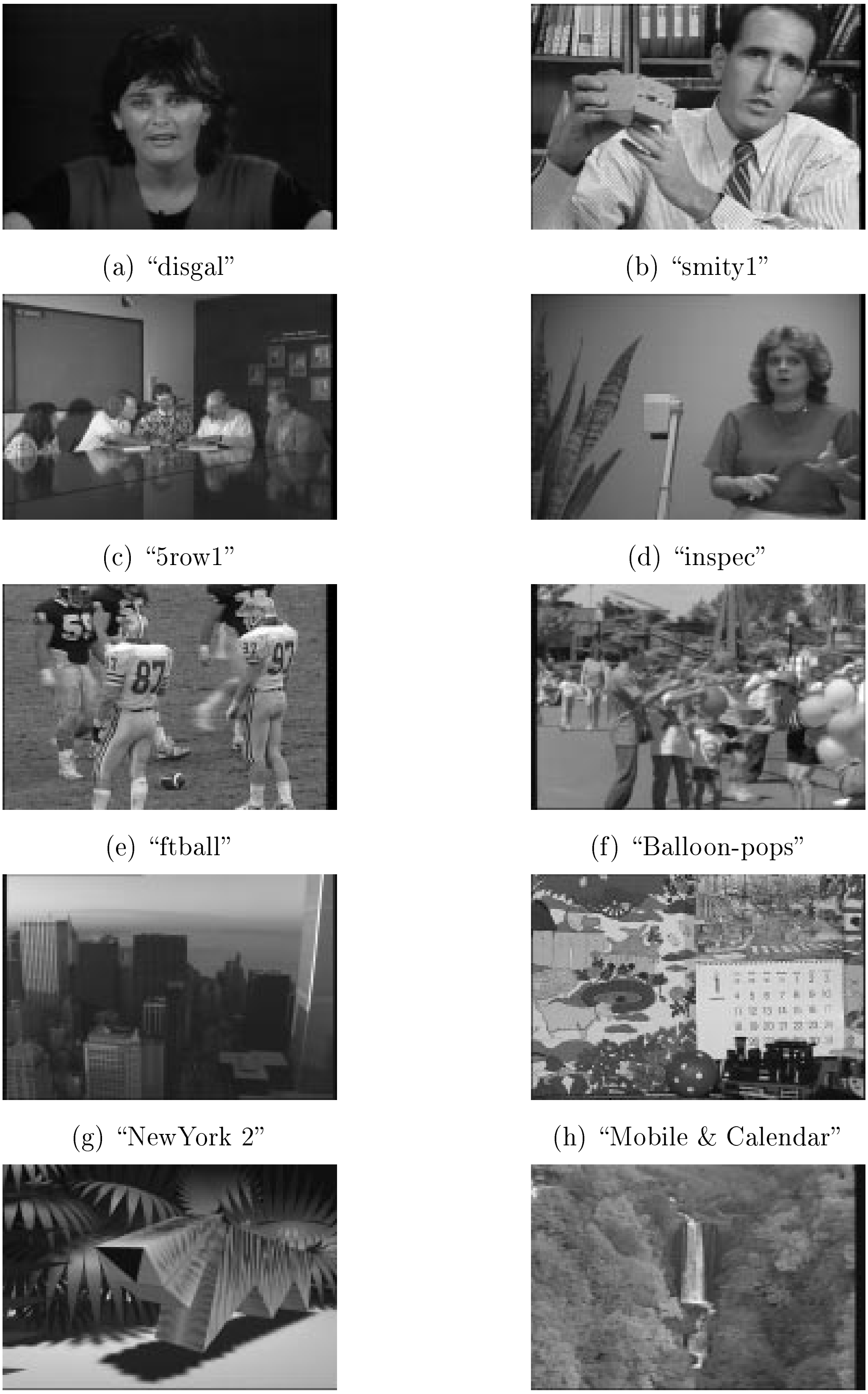
Two HRCs were applied to the VQEG scenes:

- HRC9, MPEG-2 at 3 Mbps, full resolution,
- HRC14, MPEG-2 at 2 Mbps, 3/4 horizontal resolution.

HRC14 was used in the parameter fit of the quality metric, but HRC9 was not. Unlike the ANSI sequences, the VQEG HRCs were created with commercial MPEG-2 codecs. In general these codecs are good at minimizing blocking artifacts, even at low bit rates. Most sequences suffer from blurring artifacts, and ringing artifacts are noticeable in some sequences. By incorporating VQEG sequences in the test sessions, it is possible to examine the ability of the proposed blocking metric to differentiate blocking artifacts from other distortions.

The test matrix of all scene/HRC combinations is shown in Table IV. A total of 30 different test combinations were used.

[4] The processed sequences are available from Dr. H. R. Wu of Monash University.



(a) “disgal”

(b) “smity1”

(c) “5row1”

(d) “inspec”

(e) “ftball”

(f) “Balloon-pops”

(g) “NewYork 2”

(h) “Mobile & Calendar”

(i) “Betes pas betes”

(j) “Autumn leaves”

Fig. 7. Frames of the test scenes.

TABLE IV

Test matrix

Scenes	TM5 HRCs					VQEG HRCs	
	HRC31	HRC32	HRC33	HRC35	HRC36	HRC9	HRC14
5row1	Yes	Yes	Yes	Yes			
Ftball	Yes	Yes	Yes		Yes		
Disgal	Yes	Yes	Yes	Yes			
Smity1	Yes	Yes	Yes	Yes			
Inspec	Yes	Yes	Yes	Yes			
Balloon-pops						Yes	Yes
New York 2						Yes	Yes
Mobile & calendar						Yes	Yes
Betes pas betes						Yes	Yes
Autumn leaves						Yes	Yes

B. Subjective Test Method

Both the VQEG [57] and the ANSI T1A1 [58] subjective test plans were the basis for designing the test. The following major changes were made:

- The assessors were asked to vote on the degree of blocking distortion only.
- The Double-Stimulus Impairment Scale variant II (DSIS-II) method as defined in ITU-R BT. 500 [9] was used instead of the DSCQS. The DSIS-II method presents two sequences to the participants, first the original, then the processed, and then the sequences are repeated in sequel, as illustrated in Fig. 8. The assessors are required to vote using a five-grade impairment scale: imperceptible (5); perceptible, but not annoying (4); slightly annoying (3); annoying (2); very annoying (1). Subjects vote by ticking the appropriate boxes on pre-designed voting forms.
- The panel consisted exclusively of expert viewers. Often non-expert viewers are used for subjective tests, because people do not necessarily need to be trained to distinguish between "good" and "bad" picture quality. In this test, however, the task is to assess

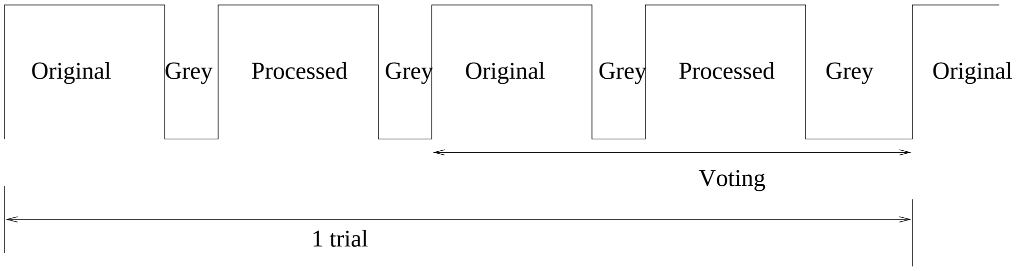


Fig. 8. Presentation structure of the test material.

blocking artifacts, which requires considerable experience. It is hard for non-expert assessors to differentiate blocking artifacts from other distortions. Therefore five expert viewers were used in the test.

The display used was a SONY BVM-20F1E color video monitor with an active display diagonal of 20 inches. The sequences were played back using a SONY D-1 digital video cassette recorder and converted to component analog video by an Abekas A26 D1 to analog converter. The viewing distance was 5 times screen height (the same as in the VQEG test). The test session was randomized so that no two consecutive trials would present the same video sequence or HRC. Before the test, instructions were given to the participants, followed by a training session.

C. Data Analysis

First, the MOS and the associated confidence intervals were calculated. Screening of subjective data was carried out according to ITU-R BT.500 [9]. No subjects had to be discarded as a result of the screening.

Three evaluation metrics were used to compare the Objective Blocking Ratings (OBR) with the MOS on blockiness: Spearman rank order correlation, Pearson correlation with logistic fit, and the average absolute error between the MOS and the OBR (E$|error|$) [22]. Currently there is no standard formula to scale PSNR into the range between 1 and 5, however the logistic fit partly serves this purpose by assuming a monotonic nonlinear relationship. Therefore, both the PSNR and the PBDM passed through the logistic fit before the average absolute error calculation.

Table V presents the evaluation results of the PBDM and PSNR. The 95% confidence bounds of the Pearson-Logistic metric were calculated using the method described in [59].

TABLE V

METRIC PERFORMANCE

Metric	Pearson-Logistic			Spearman	E$\lvert error\rvert$	Average standard deviation in the subjective test data
	Correlation	Upper bound	Lower bound			
PBDM (all)	0.961	0.982	0.918	0.910	0.296	0.455
PSNR (all)	0.489	0.726	0.149	0.517	1.038	
PBDM (TM5)	0.960	0.985	0.900	0.950	0.301	0.511
PSNR (TM5)	0.516	0.784	0.085	0.353	1.121	
PBDM (VQEG)	0.933	0.985	0.730	0.521	0.288	0.343
PSNR (VQEG)	0.531	0.873	0	0.806	0.872	

The average standard deviations of the subjective data are also shown. The calculations are not only based on the full data set (TM5+VQEG), but also grouped by encoder type (TM5 vs. VQEG). Fig. 9 also illustrates the scatter plot of the OBR versus the MOS. As shown experimentally, the PBDM achieves a very good agreement with the MOS, as reflected by the high correlations and low prediction errors.

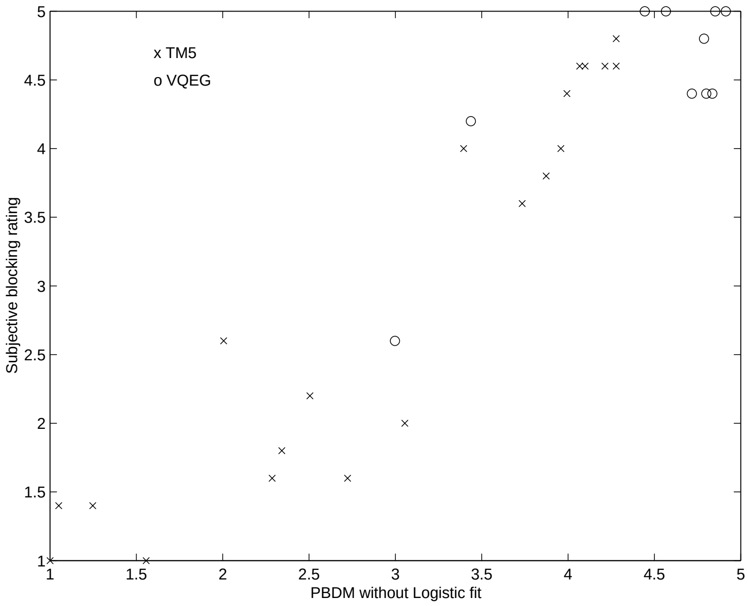
The influence of the codec type on the metric performance has also been analyzed. As mentioned before, the TM5 sequences are predominantly distorted by blocking artifacts, while most VQEG sequences exhibit only slight blocking artifacts. Comparing the PBDM performance on TM5 sequences vs. the full data set (TM5+VQEG), which was the major motivation to add VQEG sequences in the test, there are minor increases in Pearson correlation, slight decreases in Spearman correlation and in prediction error. Overall, adding VQEG sequences has little influence on the performance of the PBDM. Analyzing the PBDM performance on VQEG sequences alone, high Pearson correlation, low Spearman correlation and low prediction error are noticed, which seems contradictory at first. Further analysis of the data reveals that VQEG data points concentrate on a small area of the high quality end, where only one MOS data point is less than 4 (see Fig. 9). In this particular case, the subjective data and the objective data are so close (the average absolute prediction error 0.288 is less than the average standard deviation of the corresponding subjective test data 0.343) that a slight prediction error could cause a significant decrease in the Spearman rank correlation, which only considers the rank order of data. A faithful judgement of metric performance must consider both correlations and prediction errors, as well as the particular consequences. In practical applications the PBDM could serve as a good approximation of subjective ratings.

Among the VQEG sequences, there are some low quality ones that suffer from severe blurring artifacts. As designed, the PBDM is not distracted by the blurring artifacts and still rates these sequences as having few blocking artifacts. Although the PSNR performs well in VQEG tests [23], where the dominant distortion is blurring, it is unsuitable for measuring blocking artifacts, as shown in the experimental results with low correlations and high prediction errors. PSNR's performance on VQEG sequences is better than that on TM5 sequences. However, even among these sequences, the overall quality of some sequences (e.g. "Mobile & Calendar") is considerably poor. This results in low PSNR ratings, but there are few blocking artifacts, consequently the blocking subjective scores are still quite high. Since PSNR is not designed for blocking artifacts, it fails in this case, leading to low Pearson correlation and high prediction error. This becomes evident in Fig. 9(b) for the two VQEG data points with low PSNR ratings.

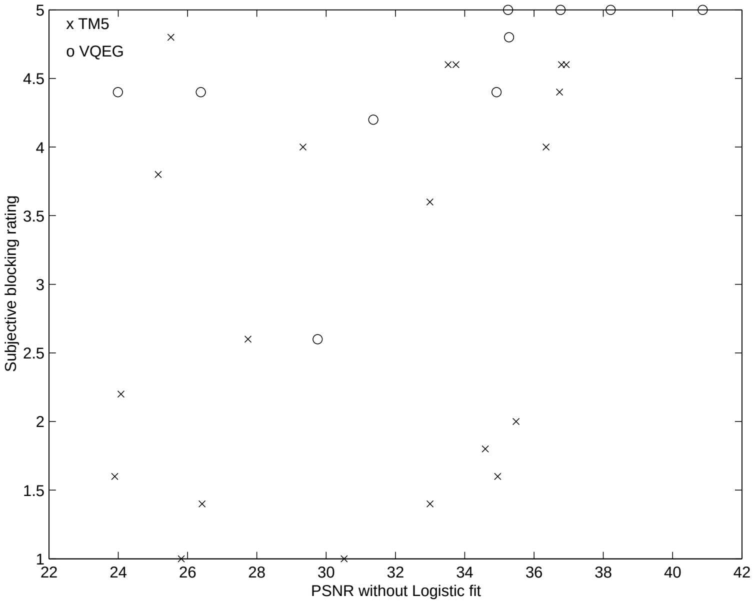
The examples shown in Fig. 10 demonstrate the performance of different metrics. The PSNR indicates similar quality for the two pictures. However, more severe blocking distortions are discerned in the "inspec" frame, where the PBDM yields results consistent with perceived quality.

VI. Conclusions

In this paper, vision model based video quality metrics have been reviewed, in particular the NVFM and the PDM. Investigations have been conducted to simplify and refine these quality metrics. It has been demonstrated through extensive simulations that with a novel optimization process a simplified quality metric can still achieve high correlations with subjective scores. Based on this simplified quality metric, a perceptual blocking distortion metric has been introduced. A segmentation algorithm has been devised to determine blocking dominant regions. The performance of the new perceptual blocking

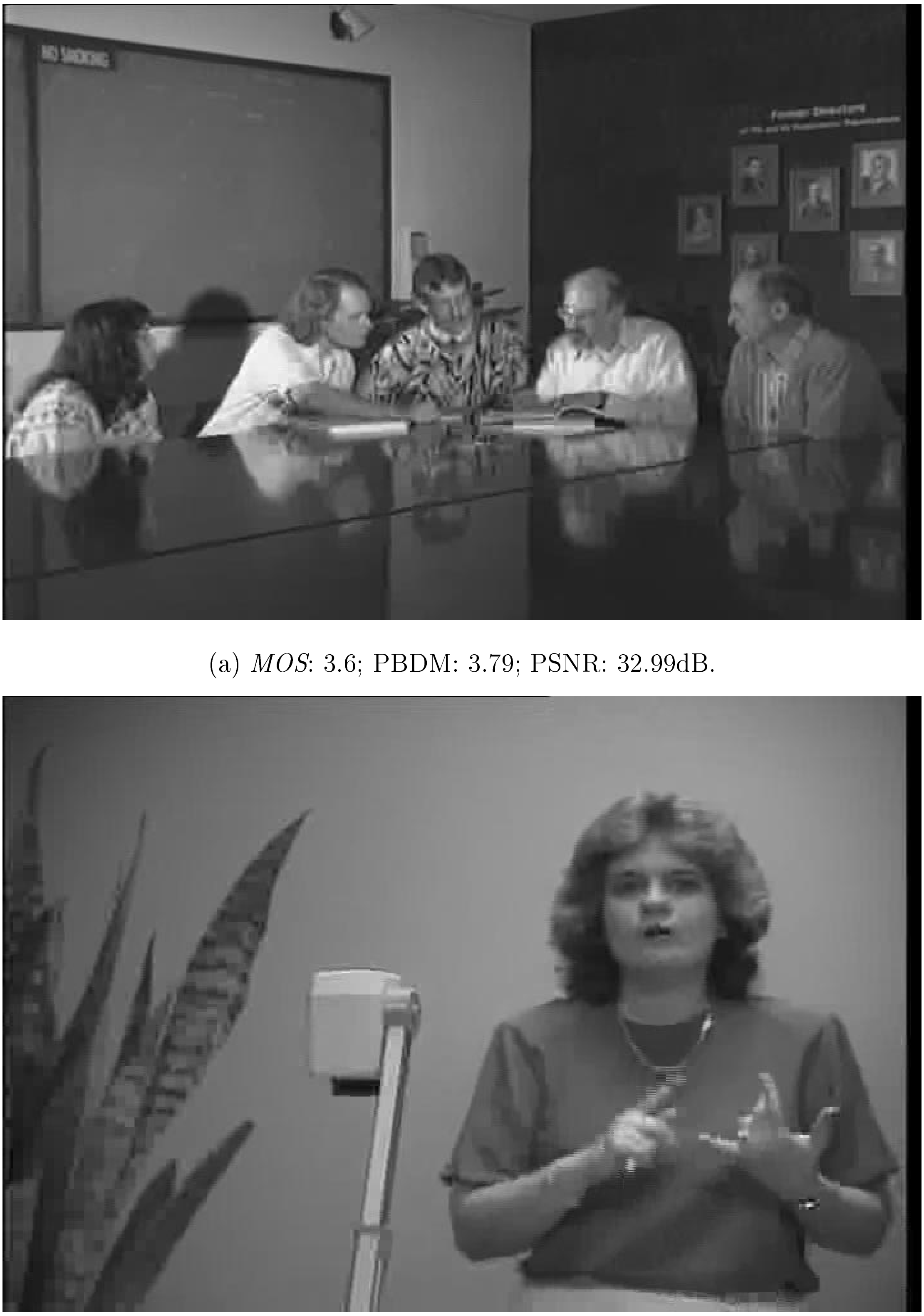


(a) PBDM without the logistic fit



(b) PSNR (in dB) without the logistic fit

Fig. 9. Scatter plots of objective ratings versus mean subjective ratings.



(a) *MOS*: 3.6; PBDM: 3.79; PSNR: 32.99dB.

(b) *MOS*: 1.4; PBDM: 1.22; PSNR: 33.00dB.

Fig. 10. Performance comparison of different metrics for (a) “5row1” and (b) “inspec” sequences.

distortion metric has been evaluated through subjective experiments, where the results show high correlations with subjective data and low prediction errors.

Acknowledgments

The work was supported by the Australian Research Council under Grant A49927209. The authors would like to thank Dr. Christian J. van den Branden Lambrecht of the EMC Media Solutions Group for providing the NVFM software, Stephen Wolf of the NTIA/ITS for offering the T1A1 data set, the VQEG for sharing the video sequences and the subjective test data, as well as the following colleagues and organization in Monash University: Damian Tan, Michael Gillies, Dr. Tim Furgeson and the Centre for Telecommunications and Information Engineering (CTIE), and the participants in the subjective tests. The authors also wish to thank the anonymous reviewers for their valuable comments.

References

[1] K. R. Rao and J. J. Huang, *Techniques and Standards for Image, Video and Audio Coding*, Prentice Hall, 1996.

[2] M. Yuen and H. R. Wu, "A survey of hybrid MC/DPCM/DCT video coding distortions," *Signal Processing*, vol. 70, pp. 247-278, 1998.

[3] ANSI T1.801.02-1995, "Digital transport of video teleconferencing/video telephony signals - Performance terms, definitions, and examples," American National Standard for Telecommunication, 1995.

[4] B. Girod, "What's wrong with mean-squared error," in *Digital Images and Human Vision*, A. B. Watson, Cambridge, MA: MIT Press, 1993, pp. 207-220.

[5] F. J. Lukas and Z. L. Budrikis, "Picture quality prediction based on a visual model," *IEEE Trans. Commun.*, vol. COM-30, pp. 1679-1692, Jul. 1982.

[6] C. R. Carlson and R. Cohen, "A simple psychophysical model for predicting the visibility of displayed information," in *Proceedings of the Society for Information Display*, vol. 21, pp. 229-245, 1980.

[7] D. J. Sakrison, "On the role of the observer and a distortion measure in image transmission," *IEEE Trans. Commun.*, vol. COM-25, no. 11, 1977.

[8] M. Miyahara, "Quality assessments for visual service," *IEEE Commun. Mag.*, vol. 26, no. 10, pp. 51-60, 1988.

[9] ITU-R Rec. BT. 500-9, "Methodology for the subjective assessment of the quality of television pictures," ITU, Geneva, Switzerland, 1998.

[10] P. T. Teo and D. J. Heeger, "Perceptual image distortion," in *Proc. First IEEE International Conference on Image Processing*, vol. 2, pp. 982-986, Nov. 1994.

[11] J. Lubin, "The use of psychophysical data and models in the analysis of display system performance," in *Digital Images and Human Vision*, A. B. Watson, Cambridge, MA: MIT Press, 1993, pp. 163-178.

[12] S. Daly, "The visible differences predictor: An algorithm for the assessment of image fidelity," in *Digital Images and Human Vision*, A. B. Watson, Cambridge, MA: MIT Press, 1993, pp. 179-206.

[13] C. J. van den Branden Lambrecht, *Perceptual Models and Architectures for Video Coding Applications*, Ph.D. Thesis, EPFL, 1996.

[14] S. Winkler, "A perceptual distortion metric for digital color video," in *Human Vision and Electronic Imaging, Proc. SPIE*, vol. 3644, pp. 175-184, 1999.

[15] Sarnoff Corp., "Sarnoff JND vision model algorithm description and testing," *VQEG Doc. jrg003*, Aug. 1997, available from *ftp.its.bldrdoc.gov.*

[16] A. B. Watson, "Toward a perceptual video quality metric," in *Human Vision and Electronic Imaging III, Proc. SPIE*, vol. 3299, pp. 139-147, 1998.

[17] KPN Research, "Objective measurement of video quality," *VQEG Doc. vqeg012*, Feb. 1997, available from *ftp.its.bldrdoc.gov.*

[18] S. Wolf and M. H. Pinson, "Spatial-temporal distortion metrics for in-service quality monitoring of any digital video system," in *Proc. SPIE International Symposium on Voice, Video, and Data Communications*, Boston, MA, Sept. 1999.

[19] T. Yamashita, M. Kameda and M. Miyahara, "An objective picture quality scale for video images (PQSvideo) - definition of distortion factors," in *Visual Communications and Image Processing 2000, Proc. SPIE*, vol. 4067, pp. 801-811, 2000.

[20] P. Bretillon et. al, "Quality meter and digital television applications," in *Visual Communications and Image Processing 2000, Proc. SPIE*, vol. 4067, pp. 780-790, 2000.

[21] N. Damera-Venkata et. al, "Image quality assessment based on a degradation model," *IEEE Trans. Image Proc.*, vol. 9, no. 4, pp. 636-650, 2000.

[22] M. Miyahara, K. Kotani and V. R. Algazi, "Objective picture quality scale (PQS) for image coding", *IEEE Trans. Commun.*, vol. 46, no. 9, pp. 1215-1226, 1998.

[23] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment," Mar. 2000, available from *ftp.its.bldrdoc.gov.*

[24] S. Winkler, "Issues in vision modeling for perceptual video quality assessment," *Signal Processing*, vol. 78, no. 2, pp. 231-252, 1999.

[25] Z. Yu and H. R. Wu, "Human visual system based objective digital video quality metrics," in *International Conference on Signal Processing 2000 of 16th IFIP World Computer Congress*, Vol. II, pp. 1088-1095, August 2000.

[26] H. R. Wu, "Analysis of video reconctruction artifacts and quality metrics," in *Proc. Australian Telecommunications Networks and Applications Conf.*, pp. 191-194, Dec. 1995.

[27] M. Miyahara and K. Kotani, "Block distortion in orthogonal transform coding - Analysis, minimization, and distortion measure," *IEEE Trans. Commun.*, vol. COM-33, no.1, pp. 90-96, Jan. 1985.

[28] S. A. Karunasekera and N. K. Kingsbury, "A distortion measure for blocking artifacts on images based on human visual sensitivity," *IEEE Trans. Image Proc.*, vol. 4, pp. 713-724, June 1995.

[29] H. R. Wu and M. Yuen, "A generalized block-edge impairment metric (GBIM) for video coding," *IEEE Signal Processing Letters*, vol. 4, no. 11, pp. 317-320, Nov. 1997.

[30] J. E. Caviedes, A. Drouot, A. Gesnot and L. Rouvellou, "Impairment metrics for digital video and their role in objective quality assessment," in *Visual Communications and Image Processing 2000, Proc. SPIE*, vol. 4067, pp. 791-800, 2000.

[31] I. Balasingham et. al, "Performance evaluation of different filter banks in the JPEG-2000 baseline system," in *Proc. IEEE Int'l. Conf. Image Processing*, vol. 2, pp. 569-573, 1998.

[32] B. A. Wandell, *Foundations of Vision*, Sinauer Associates, Inc., 1995.

[33] A. Bradlwy and I. Ohzawa, "A comparison of contrast detection and discrimination," *Vision Res.*, vol. 26, no. 6, pp. 991-997, 1986.

[34] R. F. Quick, Jr. and J. R. Hamberly, "The absence of a measurable 'critical band' at low suprathreshold contrasts," *Vision Res.*, vol. 16, pp. 351-355, 1976.

[35] T. Carney et. al, "The development of an image/threshold database for designing and testing human vision models," in *Human Vision, Visual Processing, and Digital Display IX*, *Proc. SPIE*, vol. 3644, pp. 542-551, 1999.

[36] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multi-scale transforms," *IEEE Trans. Inform. Theory*, vol. 3, pp. 587-607, 1992.

[37] A. B. Watson and J. A. Solomon, "A model of visual contrast gain control and pattern masking," *J. Opt. Soc. Am. A*, vol. 14, pp. 2379-2391, 1997.

[38] L. M. Hurvich and D. Jameson, "An opponent-process theory of color vision," *Psych. Rev.*, vol. 64, pp. 384-404, 1957.

[39] A. B. Poirson and B. A. Wandell, "Pattern-color seperable pathways predict sensitivity to simple colored patterns," *Vision Res.*, vol. 36, no. 4, pp. 515-526, 1996.

[40] S. Winkler, "Quality metric design: A closer look," in *Proc. SPIE Human Vision and Electronic Imaging Conference*, vol. 3959, San Jose, California, January 22-28, 2000.

[41] A. A. Michelson, *Studies in Optics*, Chicago, Ill.: U. Chicago Press, 1927.

[42] E. Peli, "Contrast in complex images," *J. Opt. Soc. Am. A.*, vol. 10, no. 10, pp. 2032-2040, Oct. 1990.

[43] O. H. Schade, "Optical and photoelectric analog of the eye," *J. Opt. Soc. Am.*, vol. 46, pp. 721-739, 1956.

[44] A. B. Watson, "Detection and recognition of simple spatial forms," in *Physical and Biological Processing of Images*, A. C. Slade, Springer-Verlag, 1983, pp. 100-114.

[45] H. R. Wilson and D. Regan, "Spatial-frequency adaptation and grating discrimination: Predictions of a line-element model," *J. Opt. Soc. Am. A*, vol. 1, pp. 1091-1096, 1984.

[46] R. F. Hess and R. J. Snowden, "Temporal properties of human visual filters: Number, shapes and spatial covariation," *Vision Res.*, vol. 32, no. 1, pp. 47-59, 1992.

[47] R. L. D. Valois et. al, "Spatial frequency selectivity of cells in macaque visual cortex," *Vision Res.*, vol. 22, no. 5, pp. 545-559, 1982.

[48] G. C. Philips and H. R. Wilson, "Orientation bandwidth of spatial mechanisms measured by masking," *J. Opt. Soc. Am. A*, vol. 1, no. 2, pp. 226-232, 1984.

[49] P. T. Teo and D. J. Heeger, "Perceptual image distortion," in *Proc. SPIE*, vol. 2179, pp. 127-141, 1994.

[50] I. Kharitonenko, X. Zhang, "Point-symmetric signal extension for tile-based image compression," in *Proc. IEEE ICASSP*, vol. 4, pp. 2067-2070, 2000.

[51] J. M. Foley, "Human luminance pattern-vision mechanisms: Masking experiments require a new model," *J. Opt. Soc. Am. A*, vol. 11, pp. 1710-1719, 1994.

[52] J. M. Foley and G. M. Boynton, "A new model of human luminance pattern vision mechanism: Analysis of the effects of pattern orientation, spatial phase, and temporal frequency," in *Computational Vision Based on Neurobiology*, T. A. Lawton, SPIE vol. 2054, 1994.

[53] L. Sachs, *Applied Statistics*, Springer-Verlag, 1984.

[54] Co-chair Objective Testgroup VQEG, KPN Research, "Evaluation of new methods for objective testing of video quality: objective test plan," Sept. 1998, available from *ftp.its.bldrdoc.gov.*

[55] ANSI T1.801.01-1995, "American national standard for telecommunication - Digital transport of video teleconferencing/video telephony signals - Video test scenes for subjective and objective performance assessment," American National Standards Institute, 1995.

[56] MPEG Software Simulation Group, mpeg2encode/mpeg2decode version 1.1, Jun. 1994, available from *ftp.netcom.com.*

[57] VQEG, "VQEG subjective test plan ver. 3," Jul. 1999, available from *ftp.its.bldrdoc.gov.*

[58] ANSI Accredited Standards Working Group T1A1 contribution number T1A1.5/94-118R1, "Subjective test plan (tenth and final draft)," Alliance for Telecommunications Industry Solutions, Washington DC, Oct. 3, 1993.

[59] G. W. Snedecor and W. G. Cochran, *Statistical Methods*, Iowa State University Press, 1989.

Zhenghua Yu(S'98-M'00) received his B.Eng. and M.Eng. degrees from Southeast University, Nanjing, China, in 1993 and 1996, respectively, and a Ph.D. degree from Shanghai Jiaotong University, Shanghai, China, in 1999, all in electrical engineering.

In August 2000, he joined the Visual Information Processing Lab of Motorola Australian Research Centre, Sydney, Australia, as a Senior Research Engineer. From August 1999 to August 2000, he was with the School of Computer Science and Software Engineering, Monash University, Melbourne, Australia, as a Research Fellow. During 1996 to 1999, he was a Research Assistant at Shanghai Jiaotong University and was a key contributor to the success of China's first HDTV functional prototype system. He also acted as an independent consultant to the industry for a short period of time. Dr. Yu's current research interests are visual communications and streaming over wireless networks and the Internet, multimedia systems, video quality assessment and digital television.

Hong Ren WU was born in Beijing, China in 1956. He received the BEng and MEng degrees from University of Science and Technology, Beijing (formerly Beijing University of Iron and Steel Technology), China, in 1982 and 1985, respectively. He received the PhD degree in electrical and computer engineering from the University of Wollongong, New South Wales, Australia, in 1990. From 1982 to 1985 he worked as an assistant lecturer in the

Department of Industrial Automation at the University of Science and Technology, Beijing. On receiving his PhD, Dr Wu joined Department of Robotics and Digital Technology, Chisholm Institute of Technology in April 1990. Joining Monash University in July 1990 after the amalgamation of Chisholm Institute of Technology and Monash University, he was on academic staff of the Department of Robotics and Digital Technology (1990–1996), and of the Department of Digital Systems (1996–1997). Since 1998, he has been with School of Computer Science and Software Engineering, at Monash University, where he is currently an associate professor in digital systems.

From 1987 to the present, Dr Wu has participated in and headed numerous research projects and industrial contracts in the fields of signal processing, image processing, and digital video coding, compression and transmission. He authored and co-authored over 130 papers in refereed international journals and conferences papers as well as commercial-in-confidence R&D reports in areas of digital signal processing, image processing, video coding and compression, circuit and systems, automatic control and DSP education. He was the co-chair of the 6th IEEE International Workshop on Intelligent Signal Processing and Communication Systems. He was a guest editor for the Special Issue on Multimedia Communication Services of the Circuits, Systems and Signal Processing journal, March 2001. His current research interests include DSP and fast DSP algorithms, image processing, audio/image/video coding and compression, multimedia signal processing and communications, digital image and video quality/impairment assessment and metrics, DSP industrial applications, and fast digital signal processors, hardware and systems.

Stefan Winkler received the M.Sc. degree in electrical engineering from the University of Technology in Vienna, Austria, in 1996, and the Ph.D. degree in electrical engineering from the Swiss Federal Institute of Technology in Lausanne in 2000 for work on vision modeling and video quality measurement. He also spent one year at the University of Illinois at Urbana-Champaign as a Fulbright student.

In 2001 he joined Genimedia, Switzerland, where he is responsible for the development of perceptual quality metrics for multimedia applications.

Tao Chen (S'99) received the BEng degree from Southeast University, Nanjing, China, and MEng degree from Nanyang Technological University, Singapore, in 1995 and 1999, respectively. He was conferred a PhD degree from Monash University, Melbourne, Australia in 2001.
Since January 2001, he has been with Sarnoff Corporation, Princeton, NJ, as a Member of Technical Staff. In Monash University, he was an associate investigator of several ARC (Australian Research Council) funded projects. His current research interests include image and video processing, noise and coding artifacts suppression, visual quality assessment, and applications of neural networks.