View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by CiteSeerX

ESTIMATING CHANNEL-INDUCED DISTORTION IN H.264/AVC VIDEO WITHOUT BITSTREAM INFORMATION

G. Valenzise, S. Magni, M. Tagliasacchi, S. Tubaro

Dipartimento di Elettronica e Informazione
Politecnico di Milano, Italy

ABSTRACT

No-reference video quality monitoring algorithms typically assume the availability of the encoded bitstream in order to assess the quality of the received signal at the decoder side. In some situations this is not possible, e.g. because the bitstream is encrypted or processed by third party decoders. Thus no-reference quality monitoring must be carried out in a *blind* way, i.e. using only pixel-domain data output by the decoder. In this paper we target this scenario for the specific case of distortion introduced by channel losses. We estimate the missing coding parameters, as well as the channel error pattern, and feed them into a no-reference quality monitoring system which produces accurate estimates of the MSE distortion. The results produced by the proposed method are well correlated (linear correlation coefficient larger than 0.8 over a wide range of packet loss rates) with the distortion computed in full-reference mode.

1. INTRODUCTION

No-reference video quality monitoring aims at assessing the visual quality of a video content without the availability of the original signal. This is particularly useful for a broad class of applications where the end user does not have access to the original video, such as video on demand, peer-to-peer video sharing or video streaming. The received video may have a lower quality with respect to the original one for two reasons. On one hand, video contents need to be lossy coded (with some quantization distortion) in order to be transmitted over a band-limited channel [1, 2]. On the other hand, the communication channel may be subject to packet losses [3] or jitter [4]. While both aspects are equally important in determining the perceived quality, in this paper we focus on channel-induced distortion, and in particular on errors due to packet drops whose effect propagates along the decoded video.

Conventional no-reference methods that deal with channel-induced distortion assume the deterministic knowledge, at the receiver side, of the actual pattern of channel errors, as well as the availability of the corrupted bitstream [5, 6, 7]. In this way, coding parameters such as motion vectors, prediction residuals and coding modes (Inter, Intra, Skip, etc.), can be readily extracted. In [8] we proposed NORM (No-Reference video quality Monitoring) to estimate, with a macroblock resolution, the MSE distortion due to channel losses. The NORM algorithm extracts the coding parameters mentioned above and uses them to produce accurate estimates (correlation coefficient greater than 0.8) of the MSE for a broad range of packet loss rates. It is well known that, for the case of distortion due to compression, the mean square error is not in general a good indicator of the perceived video quality. Conversely, metrics based on the MSE such as the PSNR have been shown to be particularly effective (correlation coefficient between PSNR and differential mean opinion scores about 0.95) to assess the quality of video corrupted by channel losses [9].

In some circumstances, the bitstream may be unavailable, e.g. because it is encrypted and/or processed by third party decoders and only the pixel values of the decoded video sequence can be used. In this case, the no-reference quality monitoring task is *blind*, in the sense that both the coding parameters and the map of pixels that have been lost must be estimated from the pixel values *at the decoder side*. In this paper, we specifically target this scenario by extending our previous work on NORM to the case where neither the bitstream (with the related coding parameters) nor the actual error pattern are available. An illustrative example of this situation is given in Figure 1, where a video signal, X, is first coded through a H.264/AVC compliant [10] encoder, and the resulting bitstream b is transmitted over an error-prone network. The noisy channel drops packets with some unknown packet loss rate (PLR), thus the received bitstream $\tilde{b}$ may differ from the original b. A H.264/AVC decoder processes the corrupted bitstream, possibly applying an error concealment strategy as in [11] to partially alleviate the effect of packet losses, and produces a reconstructed video $\tilde{X}$ in the pixel domain. This decoded video $\tilde{X}$ is all the information we postulate to have in order to produce an estimate of the mean square error distortion, $\widehat{MSE}$, between the error-free decoded video $\hat{X}$ and the noisy one $\tilde{X}$, as in the NORM setting. The distortion introduced by lossy coding, indeed, can be approximately considered to be uncorrelated with channel-induced distortion [12], so the two terms can be summed up in order to obtain the overall distortion with respect to X.

978-1-4244-6960-4/10/$26.00 ©2010 IEEE

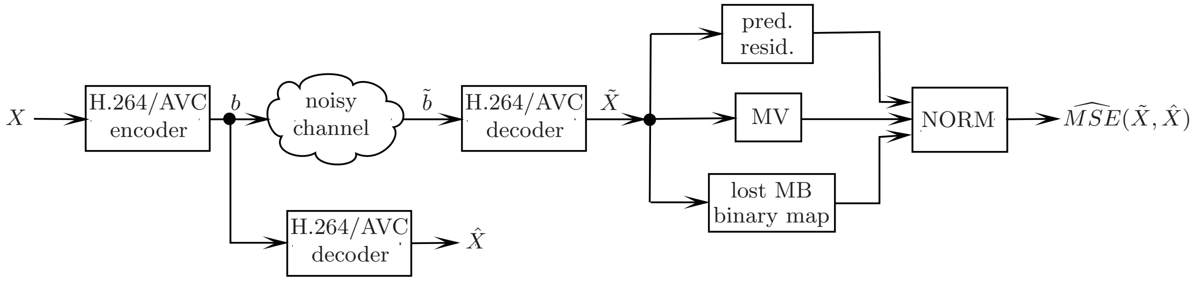


Fig. 1. Overview of the blind no-reference quality assessment system. We estimate the missing parameters from the corrupted decoded video $\tilde{X}$ and use them as input to NORM. The results is a macroblock-level map of MSE distortion between the noisy decoded $\tilde{X}$ and the video reconstructed at the encoder $\hat{X}$.

The rest of this paper is organized as follows. Section 2 reviews the basics of the NORM algorithm; the estimation of the coding parameters that are used as input to NORM is detailed in Section 3. Section 4 presents the results of the proposed blind method, while Section 5 concludes the paper.

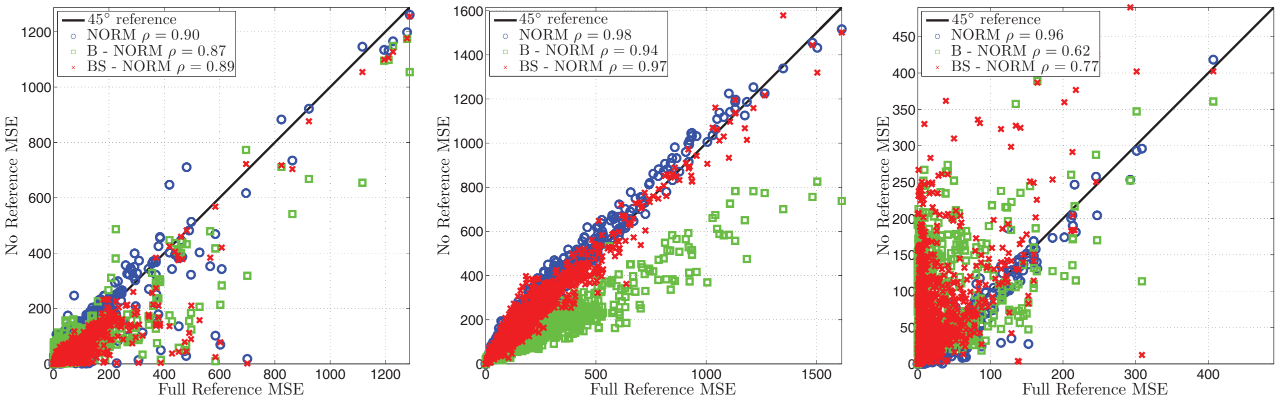
2. BACKGROUND

According to the H.264/AVC video standard [10], each frame of a video sequence is partitioned into non-overlapping regions of pixels called MBs. Each macroblock can be coded exploiting either the spatial redundancy (INTRA coding) or the temporal redundancy (INTER coding). Furthermore, some MBs may be coded using the SKIP mode, meaning that their predictor is found at the decoder without the need of transmitting any prediction residual. Coded data relative to macroblocks are gathered into slices, then packetized and transmitted through a noisy channel, that drops packets according to a given PLR. At the receiver side, macroblocks belonging to lost packets cannot be decoded. Therefore the decoder tries to partially recover lost data by means of an error concealment algorithm. Lost data cannot be perfectly recovered and channel distortion is inevitably introduced. Moreover, the inter-macroblock coding allows channel errors in previously decoded frames to propagate along the decoded video sequence, affecting also those macroblocks for which data have been correctly received, thus introducing temporal error propagation.

The NORM algorithm [8] receives as input a H.264/AVC compliant bitstream that has been transmitted over a noisy channel. The received bitstream is processed by the H.264/ AVC decoder, which applies its own embedded concealment strategy over lost data. The decoded frame, together with the received/concealed motion vectors, prediction residuals and coding modes are fed into NORM, which provides an estimate of the channel induced distortion $\widehat{MSE}_n^i$, for the ith macroblock in frame n. Also, NORM needs to know the pattern of channel errors, which consists of a binary map of the macroblocks that have been lost during transmission. The NORM algorithm specifically considers four types of error propagation deriving, respectively, from spatial prediction, spatial concealment, temporal prediction and temporal concealment. The distortion due to the temporal propagation of errors is modeled as a weighted sum of the distortion already found for the pixels used as predictors in the reference frame(s). The distortion due to the action of the spatial concealment is related to the loss of high frequency content of the lost MB, caused by the spatial interpolation performed during concealment. NORM estimates this loss by comparing the interpolated block with the one obtained with a simple zero-motion temporal concealment, which typically preserves the high frequency content of the original block. In the case of temporal concealment, the distortion is due to both the loss of the original motion vectors and to the lack of prediction residuals. Both terms are explicitly considered by NORM. As for the effect of spatial prediction, it is shown in [8] that the overall impact on distortion is negligible. The same consideration holds for the smoothing effect introduced by the deblocking filter.

3. BLIND ESTIMATION OF NORM PARAMETERS

As discussed in the previous section, the NORM algorithm requires as input the (possibly corrupted and concealed) decoded frame, the motion vectors, the prediction residuals and a map of the macroblocks belonging to slices that have been lost (see Figure 1). These parameters are not available in our setting, so they need to be estimated. We assume an error concealment strategy for inter-macroblocks as the one described in [11]: we choose the motion vector for the lost MB that minimizes the side match distortion, i.e. the absolute difference between boundary pixels of the current sample area and of neighboring blocks. The pixels pointed by this motion vector are then copied in the missing MB. As for intra-macroblocks concealment, we adopt zero-motion copy, which typically achieves better results than spatial interpolation as observed in [8]. The estimation of coding parameters is detailed in the following.



(a) *Foreman* (b) *Mobile* (c) *News*

Fig. 2. Frame level scatter plots with their respective correlation coefficients ρ for PLR $= 3\%$.

3.1. Motion vectors (MV) and prediction residuals

We find motion vectors by performing motion estimation on the decoded sequence. Any motion estimation algorithm can be used for this purpose. We set a number of reference frames k on which the search is carried out, as it is not known which is the exact number of reference frames used by the encoder. Larger values of k provide a better estimation, but they clearly entail a larger computational cost. We use $k = 5$ in our experiments. Prediction residuals can be readily computed once MVs have been found. Together with the prediction residuals, for each frame of $M \times N$ pixels we build a $(M/B) \times (N/B)$ map $\mathbf{E}$ of prediction residual energies, whose ith entry gives the MSE distortion between the ith $B \times B$ macroblock in the current frame and its respective predictor in the reference frame.

3.2. Map of lost macroblocks

In H.264/AVC video, macroblocks are divided into independent slices in such a way that each slice is contained in just one packet to be transmitted. Therefore, when a packet is lost, so are all the macroblocks of that slice. The NORM algorithm needs as input a binary map, asserting for each MB whether it has been received correctly or not. The accuracy of such a map is crucial to achieve a satisfactory distortion estimation, as it is used by NORM to determine the temporal propagation of the errors along frames. We propose two methods to estimate the lost MB binary map.

3.2.1. Blind binary map estimation

Even though the bitstream information about the lost MB is not available, we can estimate it by exploiting knowledge about how the concealment works, without the need of further assumptions on coding parameters such as macroblock-level coding modes. As mentioned above, we assume motion-compensated temporal concealment for inter-coded macroblocks, and zero-motion copy for intra-coded macroblocks. Thus, lost and concealed MBs must have a predictor, in a previous reference frame, with zero MSE or, equivalently, $\mathbf{E}_i \approx 0$ for concealed macroblocks. In practice, the residue energy is never exactly zero, e.g. because the deblocking filter is enabled. Therefore we deem as lost macroblocks the ones for which $\mathbf{E}_i \leq \tau$, where τ is a threshold empirically set to 0.3 in our experiments. The resulting $(M/B) \times (N/B)$ binary map $\mathbf{BM}$ is thus obtained from $\mathbf{E}$ by the simple thresholding:

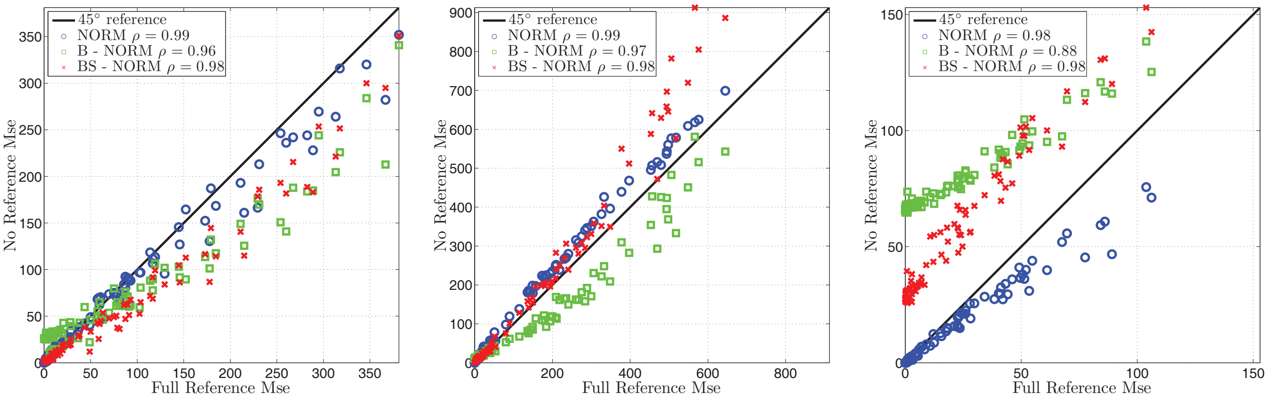
$$\mathbf{BM}_i = \begin{cases} 1 & \mathbf{E}_i \leq \tau \\ 0 & \text{otherwise.} \end{cases} \qquad (1)$$

In the following, we refer to this blind MSE estimation approach as B-NORM.

3.2.2. Blind binary map estimation with prior knowledge on slice structure

The main drawback of the previous approach is that it tends to overestimate the number of lost macroblocks. Actually, also skipped macroblocks satisfy (1) and may be incorrectly labeled as concealed. Indeed, the slice structure of H.264/AVC imposes topological constraints over the displacement of concealed MBs. In fact, it is not possible that inside one slice only a subset of macroblocks have been lost, while it may happen that only some of them have been coded using the SKIP mode. If we assume knowledge of the slice structure[1], we can leverage these constraints to improve the accuracy of the binary map $\mathbf{BM}$ by implementing a simple voting mechanism. Let $\mathcal{S}_j$ be the set of indexes of the macroblocks belonging to slice

[1] In general, it is not possible to infer the slice structure working purely on pixels. In our experiments we have considered a raster-scan slice structure (each row of MBs forms a slice) which is one of the most commonly used slicing schemes.



(a) *Foreman*

(b) *Mobile*

(c) *News*

Fig. 3. Sequence level scatter plots with their respective correlation coefficients ρ for all the tested PLRs.

j, and let $\mathbf{BM}(\mathcal{S}_j)$ denote the binary values of the lost MB map over the support $\mathcal{S}_j$. We can produce a refined estimate $\mathbf{BM}_s$ of the binary map in (1) as follows:

$$\mathbf{BM}_s(\mathcal{S}_j) = \begin{cases} 1 & \|\mathbf{BM}(\mathcal{S}_j)\|_0 \geq \sigma \\ 0 & \text{otherwise}, \end{cases} \qquad (2)$$

where the ℓ_0 norm $\|\cdot\|_0$ simply counts the number of nonzeros elements of $\mathbf{BM}(\mathcal{S}_j)$, and σ is a an empirically-set threshold. In the following, we refer to this algorithm using information about the slicing structure as BS-NORM, to distinguish it from the basic B-NORM method described in Section 3.2.1.

4. EXPERIMENTAL RESULTS

We tested our blind no-reference method with three CIF resolution video sequences: *Foreman*, *Mobile & calendar* and *News*. The video sequences have been coded with a fixed quantization parameter for I and P slices (QP $=$ 36), with a frame rate of 30 Hz, using the H.264/AVC reference software encoder (version JM12.3 [13]) with the main profile. The adopted error concealment is the one implemented in the reference decoder [11], but for Intra frame spatial concealment has been substituted with zero-motion temporal concealment.

Each coded frame is partitioned into slices, where each slice contains a horizontal row of macroblocks. Each coded slice is then packetized according to the real-time transfer protocol (RTP) specifications [14]. The simulated error-prone channel drops coded packets according to a packet loss rate (PLR) in the range [0.1 10]. The error patterns have been generated using a two-state Gilbert's model [15] with average burst length of three packets.

We simulated the transmission of the test sequences over 15 channel realizations for each considered PLR value. For each realization, we measured the $\widehat{MSE}$ distortion averaged at the frame or sequence level, estimated with the non-blind NORM method (which can use the received bitstream), and with our proposed blind no-reference approach. As for our method, we tested the two different techniques to estimate the binary map of lost macroblocks proposed in Section 3.2: the totally blind approach B-NORM and the blind approach with a-priori information on the slice structure BS-NORM, with a threshold $\sigma = 5$ macroblocks.

Figure 2 shows the scatter plot between the true full-reference MSE, and the no-reference one computed with the three methods described above, for a PLR equal to 3%. Clearly, the non-blind NORM approach is the one that performs better, because it has access to the original motion vectors and prediction residuals and, specifically, to the true map of channel errors. As pointed out in Section 3.2, an accurate estimation of $\mathbf{BM}$ is crucial for the quality of the MSE estimation, since NORM uses this information to temporally propagate errors along frames. Indeed, the main source of error of B- and BS-NORM with respect to NORM is due to a wrong estimation of $\mathbf{BM}$, which is rather frequent when the number of SKIP macroblocks is very large compared to the number of concealed MBs. This is particularly evident at low PLRs, and for sequences which have a very simple motion, and thus a high fraction of blocks coded as skipped (e.g. *News*). This phenomenon can be better observed in Figure 3, which shows the scatter plots for the three sequences when the MSE is computed at the sequence level. Both the B- and BS-NORM methods fail to provide accurate estimates for the *News* sequence, while for the first two video sequences the estimated distortion matches pretty well the ground-truth data.

This is confirmed by extensive simulations performed for each PLR, with the distortion computed both at the frame or sequence level. We quantified the quality of the estimation methods by measuring the Pearson correlation coefficient between the estimated and the actual MSE distortion. The results are reported in Tables 1-3. As observed before, the pro-

	B - NORM		BS - NORM		NORM	
PLR [%]	Frame	Seq	Frame	Seq	Frame	Seq
0.1	0.74	0.89	0.97	0.97	0.98	0.98
0.4	0.60	0.92	0.94	0.95	0.97	0.97
1	0.83	0.88	0.96	0.97	0.96	0.97
3	0.87	0.87	0.89	0.90	0.90	0.95
5	0.88	0.94	0.86	0.91	0.88	0.94
10	0.93	0.93	0.92	0.94	0.93	0.93
All PLRs	0.91	0.96	0.92	0.98	0.93	0.99

Table 1. Correlation coefficient between $\widehat{MSE}$ and the true MSE at the frame and sequence level for the *Foreman* video sequence.

	B - NORM		BS - NORM		NORM	
PLR [%]	Frame	Seq	Frame	Seq	Frame	Seq
0.1	0.88	0.93	0.96	0.97	0.99	0.99
0.4	0.88	0.91	0.95	0.96	0.98	0.98
1	0.88	0.92	0.93	0.93	0.97	0.97
3	0.94	0.95	0.97	0.98	0.98	0.98
5	0.89	0.89	0.95	0.95	0.98	0.99
10	0.77	0.84	0.81	0.87	0.98	0.98
All PLRs	0.90	0.97	0.93	0.98	0.99	0.99

Table 2. Correlation coefficient between $\widehat{MSE}$ and the true MSE at the frame and sequence level for the *Mobile* video sequence.

	B - NORM		BS - NORM		NORM	
PLR [%]	Frame	Seq	Frame	Seq	Frame	Seq
0.1	0.55	0.71	0.63	0.71	0.99	0.99
0.4	0.54	0.70	0.62	0.74	0.97	0.97
1	0.57	0.75	0.68	0.88	0.98	0.98
3	0.62	0.88	0.77	0.88	0.96	0.97
5	0.69	0.90	0.72	0.81	0.93	0.95
10	0.68	0.87	0.73	0.91	0.91	0.92
All PLRs	0.67	0.88	0.72	0.98	0.96	0.98

Table 3. Correlation coefficient between $\widehat{MSE}$ and the true MSE at the frame and sequence level for the *News* video sequence.

posed blind methods perform satisfactorily for the *Foreman* and *Mobile* sequences, while they have some problems for very static videos such as *News*. In fact, sequences with little motion tend to have a large number of skipped MBs in static/background regions of the frame. Consequently, the tests (1)-(2) produce a pessimistic estimate of the binary map, i.e. there is an extra amount of false positives that cause an overestimation of the MSE, as can be seen from the scatter plot of Figure 3(c). A possible solution to alleviate this problem might be to use a background subtraction algorithm to label MBs which are likely to be part of the background (and thus to be coded as skipped MBs). Indeed, losses of these MBs are quite easy to conceal, so labeling them as correctly received should not deteriorate considerably the performance of NORM. From Tables 1-3, it can be seen also that performance tend to increase with the PLR. Again, this has actually a simple explanation in terms of skipped macroblocks. In fact, if the PLR increases, so does the average number of concealed MBs in a frame. Thus the fraction of macroblocks with zero prediction error residual due to concealment rather than SKIP coding mode gets more significant. This implies a higher voting weight of concealed MBs in (2), if the BS-NORM approach is used.

5. CONCLUSIONS

In this paper we present a blind no-reference video quality monitoring system to estimate channel-induced distortion without the availability of the coded bitstream. We build on the NORM system, which can provide an accurate macroblock-level distortion map taking as input network (binary map of lost MBs) and coding (motion vectors, prediction residuals) parameters. This fine-granularity estimation is particularly beneficial as it enables more sophisticated error pooling strategies, which can be used to compute a wide variety of perceptual metrics that leverage localized distortion information. In our blind setting, these parameters are not available and need to be estimated in order to be fed into NORM. We observe that the key aspect governing the estimation performance is the binary map of lost MBs, as it is used by NORM to temporally propagate errors. We devise two methods to estimate it. In B-NORM we simply deem as concealed the macroblocks having zero prediction error residual. This may lead to incorrectly label skipped macroblocks as lost. If further information about slice organization is available, we propose a simple voting mechanism (BS-NORM) to enhance the correlation of the estimated results with the full-reference mode. Our results show that, even in a blind context, our estimates are well correlated (correlation coefficient larger than 0.8 at the sequence level for PLR $\geq 1\%$) with the true distortion. However, we observe that for static sequences, the high number of skipped macroblocks deteriorates the quality of estimation. As a future work, we aim at improving the binary map of lost MBs by modeling motion dependencies between pixels and background/foreground segmentation in order to handle static regions of the frame.

6. REFERENCES

[1] T. Brandao and M. P. Queluz, "No-reference PSNR estimation algorithm for H.264 encoded video sequences," in *EURASIP European Signal Processing Conference*, Lausanne, Switzerland, August 2008.

[2] A. Ichigaya, M. Kurozumi, N. Hara, Y. Nishida, and E. Nakasu, "A method of estimating coding PSNR using quantized DCT coefficients," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 2, pp. 251–259, 2006.

[3] Y. Wang, Z. Wu, and J.M. Boyce, "Modeling of transmission-loss-induced distortion in decoded video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 6, pp. 716, 2006.

[4] M. Claypool and J. Tanner, "The effects of jitter on the peceptual quality of video," in *Proc. of the seventh ACM Int. Conf. on Multimedia*, New York, NY, USA, October 1999, pp. 115–118.

[5] A. R. Reibman, V. A. Vaishmpayan, and Y. Sermadevi, "Quality monitoring of video over a packet network," *IEEE Trans. Multimedia*, vol. 6, no. 2, pp. 327–334, April 2004.

[6] T. Yamada, Y. Miyamoto, and M. Serizawa, "No-reference video quality estimation based on error-concealment effectiveness," in *IEEE Packet Video*, Lausanne, Switzerland, November 2007.

[7] S. Tao, J. Apostolopoulos, and R. Guérin, "Real-time monitoring of video quality in IP networks," *IEEE/ACM Trans. Networking*, vol. 16, no. 5, pp. 1052–1065, 2008.

[8] M. Naccari, M. Tagliasacchi, and S. Tubaro, "No-reference video quality monitoring for H.264/AVC coded video," *IEEE Trans. Multimedia*, vol. 11, no. 5, pp. 932–946, August 2009.

[9] M. Naccari, M. Tagliasacchi, and S. Tubaro, "Subjective evaluation of a no-reference video quality monitoring algorithm for H.264/AVC video over a noisy channel," in *Proc. Int. Conf. Image Processing*, Cairo, Egypt, November 2009.

[10] ITU-T, *Information Technology - Coding of Audio-visual Objects - Part 10: Advanced Video Coding*, May 2003, ISO/IEC International Standard 14496-10:2003.

[11] G. J. Sullivan, T. Wiegand, and K.-P. Lim, "Joint model reference encoding methods and decoding concealment methods," Tech. Rep. JVT-I049, Joint Video Team (JVT), September 2003.

[12] Z. He, J. Cai, and C.W. Chen, "Joint source channel rate-distortion analysis for adaptive mode selection and rate control in wireless video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 6, pp. 511–523, 2002.

[13] Joint Video Team (JVT), "H.264/AVC reference software version JM14.2," downloadable at http://iphome.hhi.de/suehring/tml/download/.

[14] S. Wenger, "H. 264/AVC over IP," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 645–656, 2003.

[15] E.N. Gilbert et al., "Capacity of a burst-noise channel," *Bell Syst. Tech. J*, vol. 39, no. 9, pp. 1253–1265, 1960.