Perceived Coding Distortion Assessment for Streaming Video

Fuzheng Yang  
State Key Lab. of Integrated Service Networks, Xidian University  
Email: fzhyang@mail.xidian.edu.cn

Shuai Wan  
School of Electronic and Information, Northwestern Polytechnical University  
Email: swan@nwpu.edu.cn

Yilin Chang  
State Key Lab. of Integrated Service Networks, Xidian University  
Email: ylchang@xidian.edu.cn

Qingpeng Xie  
HuaWei Technologies Cooperation LTD.  
Email: alicxie@huawei.com

Abstract—For applications involving video streaming, full decoding is usually not acceptable for quality assessment. To address the inherent challenges, an efficient method for coding distortion assessment is proposed in this paper. Building on empirical analysis, the proposed method employs a linear model to assess the coding distortion using the quantization scale. Furthermore, the characteristics of the human visual system are exploited by taking into account the spatial and temporal masking. To estimate the required spatial and temporal complexities in absence of sufficient information, a rate-distortion model is theoretically derived to formulate their relationship with the coding bit-rate. Extensive experimental results have demonstrated the effectiveness of the proposed method for quality assessment with respect to perceived coding distortion.

Index Terms—coding distortion, video quality assessment, streaming video

I. INTRODUCTION

Rapidly evolving technologies in networks and multimedia communications have increased the prevalence of video streaming during the last two decades. For the underlying applications, real-time processing is usually preferable. Therefore there is a crucial requirement for quality assessment for streaming video that the perceived video quality should be assessed in an accurate and efficient way. It is obviously that conventional methods for objective quality assessment cannot fulfill such a requirement since for most of them the time-consuming full decoding is required, not to mention those based on subjective assessment where human assessors should be involved.

Considering the lossy nature of video coding and characteristics of practical networks, quality degradation or distortion in streaming video consists of two main parts: coding distortion caused by quantization and distortion caused by packet loss. A variety of objective methods for quality assessment has been designed for streaming video, with focuses on assessing the coding distortion only [1, 2] or on evaluating the effect of packet loss on video quality [3, 4]. Indeed, coding distortion assessment is fundamental in video quality assessment, since it can be used to assess the quality in absence of channel errors. When channel errors are present, coding distortion assessment still serves as a critical part for quality assessment for streaming video. In this paper, coding distortion assessment for streaming video is addressed.

Without full decoding of the streaming video, very little information could be directly utilized, which makes the related assessment of coding distortion rather difficult. By analyzing or partial decoding the bit stream, a few methods have been reported to estimate the peak signal to noise ratio (PSNR) of the streaming video through using quantization scale and information about DCT coefficients [1,2]. Traditionally, PSNR has been widely used to represent the quality of a video sequence. However it does not always correlate well with the perceived quality and therefore is rarely employed in actual video quality assessment. Alternatively, video quality assessment is usually devoted to objectively estimate the perceived quality to emulate the detection by human vision. For that purpose, [5] proposed to simply assess the perceived quality using the average bit-rate of the encoded video stream, whereas only the bit-rate itself obviously cannot provide sufficient information for effective assessment of video quality.

Quantization is the essential reason for coding distortion, and therefore the relationship between the quantization scale and the perceived coding distortion is initially analyzed in this paper. Moreover, the perception of coding distortion by the human visual system (HVS) is not only affected by the quantization scale, but also significantly dependent on the video content. Considering...
this fact, the spatial complexity and temporal complexity are also taking into account in the underlying coding distortion assessment. It is noticeable that full decoding is usually not preferable in practice when performing quality assessment for streaming video. Therefore we propose to estimate the related complexities using the compressed bit-rate. The necessary information, i.e. the quantization scale and the compressed bit-rate, can be easily obtained through partial decoding of the header information and simple analysis on the video stream when only the video bitstream and packet headers are accessible in practical video streaming services.

The reminder of this paper is organized as follows. Section 2 initially provides extensive empirical analysis on the relationship between the quantization scale and the perceived video quality, based on which a linear model is proposed for coding distortion assessment. Then the characteristics of video content are taken into account in the assessment. Through theoretical derivation, the spatial complexity and temporal complexity are estimated using the compressed bit-rate and further incorporated in the initially proposed model for coding distortion assessment. To validate the performance of the proposed method, experimental results are reported in section 3. Section 4 closes this paper with concluding remarks.

II. ASSESSMENT OF PERCEIVED CODING DISTORTION

A. Perceived Coding Distortion and Quantization Scale

Quantization is the central reason for coding distortion and therefore the perceived coding distortion is closely tied to the quantization scale. However, their relationship has not been well studied in the literature. To evaluate the underlying relationship, extensive experiments have been conducted and the corresponding results suggest a linear relationship between them. As shown in Fig. 1, the perceived video quality measured by the subjective mean opinion score (MOS) can be well estimated using a linear model of the quantization parameter (QP) for all tested sequences. However, from the observation we can also see that the related parameters (i.e. slopes and intercepts) of the linear model are quite different for different video content. Such differences in model parameters comply with the fact that apart from the quantization scale, the perception of the HVS also significantly depends on other factors such as video content. Following this observation, the influence of video content on the perceived quality is further incorporated into the assessment of coding distortion in terms of using its two main factors: the spatial complexity and temporal complexity.

Generally speaking, the spatial complexity can be captured by the mean squared error or variance of a frame, and the temporal complexity can be well represented by an average value of the maximal motion offset of each frame such as motion vectors. However, as discussed in the previous section, those values are not available for such cases as quality assessment for streaming video due to lack of full decoding. An accurate estimation of these two factors is then needed by which the unfeasible full decoding can be avoided.

B. Estimation of Spatial and Temporal Complexity using Bit-Rate

For frames coded using intra modes, i.e. I frames, the spatial complexity can be measured by the variance of the pixel values. It is obvious that a high spatial complexity in a frame usually leads to a high value of the variance, and a

![Figure 1: the relationship between MOS and quantization parameter. (a) “News”; (b) “Football”; (c) “Mobile”.

© 2010 ACADEMY PUBLISHER
low spatial complexity is usually corresponding to a small variance. For P frames, on the other hand, the variance of the pixel values in a residual frame after motion-compensated prediction (MCP) reflects the temporal complexity to some extent. That is, the more complex the motion is, the more residual will be retained after MCP, and the higher the corresponding variance will be. Therefore the spatial complexity and temporal complexity can be captured by the variance of a frame and a residual frame, respectively. However, these values of the variance are only available after full decoding and accurate estimation using available information is needed for them to be used in practice.

Given a video bitstream, the most straightforward information which could be obtained without sorting to decoding is the bit-rate for coding each frame. Such information will be readily available after a simple analysis on the headers embedded in the bitstream. In this paper a rate-distortion model is theoretically derived to form the relationship between the variance and the bit-rate, building on which the variance of pixels in a frame or a residual frame is estimated for complexity measurement. The proposed model is described in the rest of this subsection.

Assuming that the DCT coefficients of the video signal are independent and identically-distributed, they can be individually approximated by a Laplacian probability distribution [6]. The well-known RD function of a Laplacian distributed source under the magnitude error criterion is given by

\[ R(D) = \ln \frac{1}{\alpha D} - \frac{1}{\alpha D} - \frac{1}{2} \left( \frac{1}{\alpha D} - 1 \right)^2 + R_s(D) \]

(1)

where \( \alpha \) is a positive constant representing the Laplacian distribution parameter. The Taylor expansion of Equation (1) is given by:

\[ R(D) = \ln x_0 + \frac{1}{x_0} \left( \frac{1}{\alpha D} - x_0 \right) - \frac{1}{2} \left( \frac{1}{\alpha D} - x_0 \right)^2 + R_s(D) \]

\[ \approx \left( \ln x_0 - 1 \right) + \frac{1}{\alpha x_0} D^{-1} + R_s(D), \]

(2)

with \( x_0 > 1 \) to satisfy convergence of the Taylor series. Since the magnitude error criterion is used as the distortion measure in Equation (2), \( D \) can be expressed as:

\[ D = Q / 4, \]

(3)

where \( Q \) is the quantisation step [7]. Inserting Equation (3) into (2), a R-Q model can be formulated as follows:

\[ R(Q) = a_0 + \frac{a_1}{\alpha} Q^{-1}. \]

(4)

Using Equation (4), the R-Q function of the \( i \)-th DCT coefficient in a transform block is given by

\[ R_i = a_0 + \frac{a_i}{\alpha} Q_i^{-1}. \]

(5)

For a Laplacian distributed source, the relationship between the standard deviation \( \sigma \) and the parameter of the Laplacian distribution can be expressed as:

\[ \alpha = \frac{2}{\sigma}. \]

(6)

For a detailed derivation of Equation (6) the reader is referred to [8]. Inserting Equation (6) into (5) the average number of bits for the \( i \)-th DCT coefficient can be obtained as:

\[ R_i = b_0 + b_i \sigma_i Q_i^{-1}, \]

(7)

where \( \sigma_i \) is the standard deviation of the Laplacian distribution for the current DCT coefficient. Unfortunately, it is very difficult to determine the standard deviation or the variance for a given DCT coefficient in realistic application scenarios. However, previous studies suggest that the distribution of the values of residual pixels after motion-compensation can be also modelled by a Laplacian distribution [9]. Based on this observation the variance of the DCT coefficient can be estimated by:

\[ \sigma_i^2 = k_i \sigma_i^2, \]

(8)

where \( k_i \) is a parameter related to the DCT transform matrix, and \( \sigma_i^2 \) is the variance of the pixel values over the current block before the DCT transform. The quantization steps in a block can be expressed as \( Q_i = q_i \cdot W_i \), where \( q_i \) is a quantization scaling factor and \( \{ W_i, i = 0, \ldots, N - 1 \} \) is the corresponding weighting matrix. Then the total average number of bits \( \bar{R} \) in a transformed block is:

\[ \bar{R} = \frac{1}{N} \sum_{i=0}^{N-1} R_i = c_0 + c_1 \sigma_i q_i^{-1}. \]

(9)

Clearly, Equation (9) gives the relationship between the rate, quantization factor, and variance of the pixel values over the block of concern.

C. Perceived Coding Distortion Assessment

Following the discussions in the previous sections, the perceived coding distortion is measured by the quality of the video without channel errors. The video quality can be assessed by a linear model of the quantization scale, while taking into account the spatial complexity and temporal complexity estimated by the bit-rate. The proposed method for perceived coding distortion assessment can be carried out in 3 steps as described in the following.

(a) Initially rate the quality for each frame with respect to the quantization scale. Specifically, the initial score of the quality of the \( n \)-th frame \( (Q_{fr,n}) \) is obtained using the following linear model:

\[ Q_{fr,n} = a_0 + \frac{a_1}{\alpha} Q^{-1}. \]
\[ Q_{F,a} = a_1 \cdot q_a + b_1, \]  

where \( q_a \) is the quantization scale employed in coding the \( n^{th} \) frame. Here \( a_1 \) and \( b_1 \) in Equation (10) are empirical parameters.

(b) The characteristics of the HVS are further incorporated in the quality assessment by making use of temporal masking and spatial masking as proposed in our previous work [10]. The score of \( Q_{F,a} \) is then modified as

\[ Q'_{F,a} = Q_{F,a} \left( 1 + \left( \frac{\sigma_{S,a}}{a_2} \right)^{b_2} \right) \left( 1 + \left( \frac{\sigma_{T,a}}{a_3} \right)^{b_3} \right), \]  

where \( \sigma_{S,a} \) and \( \sigma_{T,a} \) are the spatial complexity and temporal complexity of the \( n^{th} \) frame, respectively. The parameters in Equation (11), i.e. \( a_2, b_2, a_3, b_3 \), are obtained by the experiments.

For intra-coded frames, the spatial complexity is computed using Equation (9) as:

\[ \sigma_{S,a} = (a_1 \cdot R_n + b_1) \cdot q_n, \]  

where \( R_n \) is the bit-rate for coding the \( n^{th} \) frame, and \( a_1, b_1 \) parameters are obtained by the experiments.

Similarly, the temporal complexity is given as follows for inter-coded frames:

\[ \sigma_{T,a} = (a_3 \cdot R_n + b_3) \cdot q_n \]  

where parameters \( a_3 \) and \( b_3 \) are also obtained by the experiments.

It is noticeable that the spatial complexity for inter-coded frames and the temporal complexity for intra-coded frames are not directly obtainable without knowing further information other than the bit-rate. Since the spatial complexity and temporal complexity do not change much in adjacent frames, in the proposed method \( \sigma_{T,a} \) for an inter-coded frame is approximated using the corresponding value of one of its neighbour inter-coded frames, and \( \sigma_{S,a} \) for an inter-coded frame approximated by the spatial complexity of the nearest inter-coded frame.

(c) Average the scores of each frame to obtain the quality for the video sequence as:

\[ Q = \frac{1}{N} \sum_{n=1}^{N} Q_{F,a}, \]  

where \( Q \) is the video quality measuring coding distortion and \( N \) is the number of frames in the video sequence.

III. EXPERIMENTAL RESULTS

The performance of the proposed method has been extensively evaluated using the MPEG-4 video codec [11]. Standard test sequences in QCIF format (176 x 144) were used. For all considered sequences, the coding bit-rates were set to 80 kbps, 120 kbps, and 240 kbps respectively. For the sake of conciseness the results reported in this paper include only six test sequences: “Football”, “Mobile”, “News”, “Foreman”, “Claire”, and “Susie”. The frame rate of 15fps was employed in all experiments.

The scores obtained using the proposed method were compared to the subjective scores. The subjective quality scores with respect to coding distortion were obtained using subjective tests, where the absolute category scale (ACR) was used in all the tests. The guidelines specified by the VQEG in [12] were followed for the subjective tests, involving 25 non-expert viewers. The viewers have evaluated the video quality using a slider device and a continuous grading scale marked with “Excellent”, “Good”, “Fair”, “Poor” and “Bad” respectively. The subjective scores are therefore quantized on a scale of [1..5].

For comparison, the objective scores obtained using the proposed method were quantized to [1..5] in terms of uniform quantization. The underlying parameters were tuned according to the experiments. The resulting values are \( a_1 = -0.119 \), \( b_1 = 5.056 \), \( a_2 = 10000 \), \( b_2 = 1.8 \), \( a_3 = 1800 \), \( b_3 = 1.3 \), \( a_4 = 2080 \), \( b_4 = 0.002 \), \( a_5 = 1170 \), \( b_5 = 0.003 \). These parameters were set fixed for all carried experiments. However, they may need to be adjusted in quality assessment for videos generated by other codecs.

<table>
<thead>
<tr>
<th>RMSE</th>
<th>PCC</th>
<th>SCC</th>
<th>OR</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.046</td>
<td>0.972</td>
<td>0.961</td>
<td>0.11</td>
</tr>
</tbody>
</table>

Figure 2. Scatter plot of MOS vs objective prediction.

Four metrics for performance evaluation suggested by the VQEG were used to evaluate the performance of the proposed video quality metric [12]. That is, the root-mean-squared error (RMSE), Pearson correlation coefficient (PCC), Spearman rank order correlation coefficient (SCC) and outlier ratio (OR) were computed between the fitted objective data and the corresponding
subjective data [10]. From Table I, we observe that the objective scores obtained through the proposed method are consistent with the subjective assessment for video sequences.

The scatter plot of the objective scores obtained by the proposed method and the scaled subjective scores is shown in Fig. 2, from which the same conclusion can be drawn that using the proposed method the perceived coding distortion can be accurately measured.

IV. CONCLUSIONS

In this paper an objective method is proposed to measure the perceived coding distortion in streaming video. The relationship between the perceived quality and the quantization scale is investigated and a linear model is proposed. Taking into account the characteristics of the HVS, the spatial and temporal masking are considered in quality assessment. Due to lack of sufficient information, the required spatial complexity and temporal complexity are estimated using the bit-rate through a proposed RD model. Extensive experimental results have demonstrated the effectiveness of the proposed method for quality assessment with respect to perceived coding distortion.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the supports provided by the Natural Science Basic Research Plan in Shaanxi Province of China (Program No. SJ08F03), the National Science Foundation of China (60772134), and the 111 Project (B08038). The supports by HuaWei Technologies Co. Ltd. for subjective quality assessment are also gratefully acknowledged.

REFERENCES


Fuzheng Yang received the B.E degree in Telecommunication Engineering, the ME degree and the Ph.D. in Communication and Information System from Xidian University, Xi’an, China, in 2000, 2003 and 2005, respectively. He became a lecturer and an Associate Professor in Multimedia Communication in Xidian University in 2005 and 2006, respectively. During 2006-2007, he served as a visiting scholar and postdoctoral researcher in Department of Electronic Engineering in Queen Mary, University of London. His research interests include video quality assessment, video coding and multimedia communication.

Shuai Wan received the B.E degree in Telecommunication Engineering and the ME degree in Communication and Information System from Xidian University, Xi’an, China, in 2001 and 2004, respectively, and obtained the Ph.D. in Electronic Engineering from Queen Mary, University of London in 2007. She is now an Associate Professor in Northwestern Polytechnical University, Xi’an, China. Her research interests include scalable/multiview video coding, video quality assessment and hyperspectral image compression.