

AN OBJECTIVE VIDEO QUALITY METRIC BASED ON SPATIOTEMPORAL DISTORTION

Junyong You^{1*}, Miska M. Hannuksela², Moncef Gabbouj¹

¹ Tampere University of Technology, Tampere, Finland; ² Nokia Research Center, Tampere, Finland
Junyong.You@ieee.org, Miska.Hannuksela@nokia.com, Moncef.Gabbouj@tut.fi

ABSTRACT

This paper proposes an objective video quality metric based on an analysis of spatial and temporal distortions. Spatial quality features extracted from the spatiotemporal region of reference and distorted videos are used to express the spatial distortion. Temporal distortion, caused by frame freezing resulting from a packet loss, is derived from the spatial distortion before and after the frozen frames. The overall quality is predicted according to the weighted combination of qualities over all the temporal regions. The experimental results with respect to the subjective measurements demonstrate the fast computation and promising performance of the proposed model compared with existing methods.

Index Terms— Objective video quality metric, spatial distortion analysis, temporal distortion analysis, PSNR

1. INTRODUCTION

Lossy video compression introduces coding artifacts and visual distortions, such as blockiness and blurring. In addition, packets can be lost when transmitting coded streams through limited-bandwidth channels, which results into different types of degradations, such as a picture freeze, depending on the error concealment in the decoder implementation. Although subjective evaluation is considered to reflect the human perception in the most accurate way, it is time-consuming and cannot be done in real time. Moreover, the widely used metrics, namely MSE and PSNR, were found not to be credible for measuring the perceived quality because they do not take the characteristics of human visual system (HVS) into account [1]. An accurate objective quality metric can be used in improving the compression ratio of the video coding scheme and evaluating the performance of a video communication system among other things.

A number of objective methods for video quality measurement have been proposed [2] since the foundation of the Video Quality Experts Group (VQEG). Based on the availability of reference information, the metrics are generally classified into three categories: full-reference (FR), reduced-reference (RR), and no-reference (NR). FR methods use the reference to predict the quality of the distorted video. NR metrics follow the approach of judging perceptual quality based only on the distorted video and are readily applicable in communication systems due to the fact that the original reference signal is not required at the receiver. RR metrics are a compromise between NR and FR methods, in which

the features extracted from the reference and the distorted videos are compared. Some FR metrics, such as perceptual evaluation of video quality (PEVQ) in ITU recommendation J.247 [3], are based on the comparison of distortion features between the reference and the distorted videos, and these kinds of metrics can be converted into RR metrics. Other FR methods compute the difference between the reference images and the degradation either in the pixel or in the transform domain. The perceptual distortion metric (PDM) [4] is a typical example of these methods. In addition, the computing unit of some metrics, such as the method proposed by NTIA (National Telecommunications and Information Administration) [5] and the structural similarity metric (SSIM) [6], is an image block. In other metrics, such as Yonsei method [7], the whole image is used. The above metrics mainly aim at predicting the quality caused by video coding, while in some metrics the influence of frozen frames caused by packet losses is also taken into account. Reibman *et al.* analyzed the attributes of packet loss impairment and measured them by a few selected metrics, such as SSIM and MSE. The authors emphasized on analyzing the characteristics of the video itself, including the camera motion and proximity to scene cuts, and predicting the packet loss visibility according to these attributes [8].

The proposed method belongs to the FR category, while it can also be applied as an RR scheme because this metric is based on the comparison of the spatial distortion parameters between the reference and distorted videos. Additionally, the temporal distortion caused by frame freezing is derived from the spatial distortion. Furthermore, the computation of the proposed distortion parameters is block-based and can therefore be applied in a rate-distortion-optimized video coding scheme.

Most of the existing video coding schemes are based on a block-based transform resulting into blockiness being a common visual artifact. In low bit rate video coding, pictures usually suffer from blurring at edges due to loss of high frequency transform components. In this study, we present a fast and effective method to detect the most prominent distortion features, namely blurring, blockiness, and temporal distortion caused by frozen frames, which are modeled based on the human perception. The main ideas of calculating the spatial quality features come from the NTIA model, while the proposed metric is much faster. Furthermore, we take the temporal distortion into account. The experimental results demonstrate the promising performance of the proposed method when compared to the respective subjective evaluation results.

The rest of this paper is organized as follows. The construction of the proposed perceptual quality metric is presented in Section 2. Section 3 gives the experimental results illustrating the performance of the metric. Finally, conclusions are drawn in Section 4.

* Junyong You is currently with Norwegian University of Science and Technology.

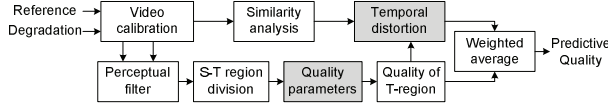


Fig. 1 Flowchart of the perception based video quality metric.

2. PERCEPTION BASED VIDEO QUALITY METRIC

Fig. 1 gives the main steps of the proposed metric. The main steps are briefly introduced below and described in depth in the following sub-sections. Before computing the spatial and temporal distortions, the metric first performs a calibration to adjust the potential spatiotemporal shift between the reference and distorted videos. In this paper, a full-reference calibration method described in NTIA report [5] was used to tune the spatiotemporal shift.

Quality degradations caused by video compression are analyzed with the bottom row of the blocks in Fig. 1. Perceptual filtering is used to enhance edges and to reduce noise. Then, a temporal region is selected to contain a certain number of adjacent frames and spatiotemporal (S-T) regions are formed, where one S-T region is a group of image blocks located in the same position within the temporal region. Spatial quality parameters are computed for each S-T region and then the quality of the respective temporal region is derived from the spatial quality of the S-T regions having the greatest distortion.

Quality degradations caused by packet losses are considered to be only frame freezes in this study. They are analyzed by segmenting the video sequence into shots in the similarity analysis block and deriving a temporal distortion measure by weighting the spatial distortion before and after the freeze period as a function of the freeze duration within particular shots. Finally, an overall quality for the entire video sequence is calculated based on the weighted average of all temporal regions. Unless stated otherwise, the coefficient values in the formulas given in the following sub-sections were derived by optimizing the performance for the training data as described in Section 3.

2.1. Spatial distortion analysis

As human vision is sensitive to image edges, two transposed perceptual filters are applied for enhancing the horizontal and vertical pixel differences. The size of the two filters is reduced to 5×5 from the original 13×13 in [5], and they have filter weights as:

$$\begin{cases} W_1(i, j) = 0.079 \cdot j / \exp(0.125 \cdot j^2) \\ W_2(j, i) = W_1(i, j) \end{cases}, \quad i, j = -2, -1, 0, 1, 2 \quad (1)$$

Each filter produces the same gain as a Sobel filter. The spatial distortion parameters are calculated based on the divided S-T regions. Since the degraded video has been calibrated, for each S-T region in the degradation there is an S-T region spanning the identical spatial and temporal position in the reference. S-T region sizes are described by the number of pixels horizontally and vertically, as well as the number of frames. In this study, we use identical horizontal and vertical extents which are determined according to the spatial complexity information defined as:

$$SI = \max_{time} \{std_{space}[W(F(n))]\} \quad (2)$$

where $W(F(n))$ denotes the filtered result of the video frame (luminance plane) F at time n by the perceptual filters W_1 and W_2 , and $time$ is a temporal segment, which is segmented with a fixed or variable duration for computing the spatial and temporal information in this segment. It is noted that the definition of SI is different from its original definition in [9] as we use the perceptual

filters instead of the Sobel filter in this paper. The temporal size of S-T region is determined by the temporal activity information defined as [9]:

$$TI = \max_{time} \{std_{space}[F(n) - F(n-1)]\} \quad (3)$$

We give some typical values of S-T region sizes according to the values of the spatial complexity and temporal activity information in the next section.

In this paper, we use the luminance plane to compute the spatial quality parameters. The following equations of the quality parameters and comparison functions are derived from those presented in [5] but some constants were further tuned according to the training set presented in Section 3. For the luminance plane of the reference or the distorted video, the filter responses of horizontal (W_1) and vertical (W_2) filters are denoted as H and V , respectively, which can be converted into polar coordinates as:

$$\begin{cases} R = \sqrt{H^2 + V^2} \\ \theta = \text{atan}(H/V) \end{cases} \quad (4)$$

Let a given S-T region be $S \times S$ pixels within K frames. An $(S \cdot K) \times S$ matrix is generated by vertically concatenating the K matrices of the $S \times S$ pixels of the S-T region. The first quality feature in this S-T region is defined on the generated matrix as follows:

$$F_{SI} = \max\{std(R), 9\} \quad (5)$$

This feature is sensitive to changes in the overall amount of spatial activity within a given S-T region. For example, localized blurring produces a reduction in the amount of spatial activity, whereas noise produces an increase. The second quality feature is sensitive to changes in the angular distribution or orientation of spatial activity, defined as:

$$F_{HV} = \frac{\max\{\text{mean}(\overline{HV}), 3\}}{\max\{\text{mean}(HV), 3\}} \quad (6)$$

where $HV = \begin{cases} R, & R \geq 20, m\frac{\pi}{2} - 0.225 < \theta < m\frac{\pi}{2} + 0.225 \quad (m = 0, 1, 2, 3) \\ 0, & \text{otherwise} \end{cases}$

and $\overline{HV} = \begin{cases} R, & R \geq 20, m\frac{\pi}{2} + 0.225 \leq \theta \leq (m+1)\frac{\pi}{2} - 0.225 \quad (m = 0, 1, 2, 3) \\ 0, & \text{otherwise} \end{cases}$.

If the horizontal and vertical edges suffer more blurring than diagonal edges, then F_{HV} of the degraded video will be smaller than F_{HV} of the reference. On the other hand, if erroneous horizontal or vertical edges are introduced, say in the form of blockiness or tiling distortion, then F_{HV} of the degraded video will be higher than that of the reference video.

For a given S-T region, the two features above, F_{SI} and F_{HV} defined in Eq. (5) and (6), are calculated for the reference and degraded videos. Then, three spatial distortion parameters, i.e. loss and gain of spatial activities, are derived based on the following comparison functions on these two features. Here, D and R denote the degraded and reference videos, respectively.

$$\begin{cases} HV_loss = \min\left\{\frac{F_{HV}(D) - F_{HV}(R)}{F_{HV}(R)}, 0\right\} \\ SI_loss = \min\left\{\frac{F_{SI}(D) - F_{SI}(R)}{F_{SI}(R)}, 0\right\} \\ HV_gain = \max\left\{\log_{10}\left(\frac{F_{HV}(D)}{F_{HV}(R)}\right), 0\right\} \end{cases} \quad (7)$$

According to our experience with subjective evaluation, localized impairments tend to draw the focus of viewers, making

the worst part of the picture the predominant factor in the subjective quality decision. Thus, it is proposed to average the worst 5% of the distortions observed over the S-T regions as suggested in [5]. Thus, the distortion of a temporal region is calculated according to the integration of the smallest 5% loss and biggest 5% gain parameter values of all the S-T regions in this temporal region (T) as follows:

$$SQ(T) = 0.4327 \cdot \{\max\{\text{mean}(S5\%_HV_loss)^2, 0.06\} - 0.06\} - 0.3269 \cdot \text{mean}(S5\%_SI_loss) + 0.2058 \cdot \text{mean}(B5\%_HV_gain) \quad (8)$$

where $\text{mean}(S5\%_SI_loss)$ denotes the mean of the smallest 5% values of SI_loss , similarly to HV_loss , while $\text{mean}(B5\%_HV_gain)$ denotes the mean of the biggest 5% values of HV_gain . In this paper, the larger SQ value corresponds to the worse quality.

2.2. Temporal distortion analysis and the overall quality

Most existing methods combine a number of quality features, many of which are spatial quality features, including the frame freeze. However, such a combination fails to discriminate the real effect of temporal frame freeze from that of other spatial features. We therefore propose to consider the effect of temporal frame freeze independently. For example, NTT model in ITU J.247 uses the duration of the freeze period as a direct feature in integrating into the overall quality; and Yonsei method increases the mean square error (MSE) using the ratio of total duration and the freeze duration as a weight. In this study, we take only the influence of the frame freeze into account, because it is a common error concealment strategy when a packet loss occurs [10]. We think that the perceived quality SQ caused by frozen pictures should be a combination of three factors: the freeze length FL , the semantic importance of the frozen frames, and the qualities of a temporal region before and another temporal region after the frozen frames, SQ_1 and SQ_2 , which can be computed by Eq. (8). The selection of the three factors is based on the assumption that viewers usually evaluate the quality during freezing according to the qualities before and after the frozen frames, whilst the importance and the duration of this segment definitely influence the perceived quality. Generally speaking, SQ is proportional to the importance and the freeze length, SQ_1 and SQ_2 of the segment. However, calculating the semantic importance of the video would make the computation more complex, so we use similarity analysis instead of importance analysis in this paper. We detect the shot boundaries during and immediately preceding and following the freeze period. Based on the shot boundaries, the freeze period is divided into three parts: frames belonging to the previous shot, the shots fully covered by the freeze period, and the frames of the subsequent shot, and the respective durations are denoted as L_1 , L , and L_2 . Then, the quality of the frozen frames in the degraded video is calculated as a linear combination resulting from the above discussion:

$$SQ = (1 + \frac{FL}{TL}) \cdot (\frac{L_1}{FL} \cdot SQ_1 + \frac{L}{FL} \cdot \frac{SQ_1 + SQ_2}{2} + \frac{L_2}{FL} \cdot SQ_2) \quad (9)$$

where TL denotes the duration of the whole sequence. For a practical system, we can transmit the computed quality features F_{HV} and F_{SI} of reference video to the receiver while the original video is unavailable, so SQ_1 and SQ_2 are computed based on the comparison between the reference and distortion features.

The overall quality of the whole video sequence is derived by combining the qualities of all temporal regions and frozen periods. For a long sequence with various scenes, the contribution of each temporal region to the overall quality is different from each other. It is a challenging task to find the best temporal combination for the overall quality based on the qualities of all temporal segments,

keeping in mind that the semantic importance and the quality of every segment are the most important factors. Most existing metrics use either direct averaging, such as the Minkowski summation in PDM, or the average of the segments with the greatest distortions, as in the NTIA model. In this study, all temporal regions contribute to the overall quality, while weighted averaging is used in computing the overall quality and the weights are derived from the respective distortions.

$$VQ = \frac{1}{A} \cdot \sum_{i=1}^N W_i \cdot SQ^2(T_i), \quad A = \sum SQ(T_i) \quad (10)$$

$$W_i = \begin{cases} 0.5624, & \text{if } SQ(T_i) \geq 0.6943 \\ 0.3207, & \text{if } 0.3187 \leq SQ(T_i) < 0.6943 \\ 0.1169, & \text{else} \end{cases}$$

where T_i denotes the i -th temporal region or the frozen segment, $SQ(T_i)$ is the spatial quality of T_i defined in Eq. (8) for a temporal region or in Eq. (9) for a frozen segment, and N is the amount of all temporal regions and frozen segments. The weights and the corresponding thresholds were derived by solving a non-linear conditional optimization on the training data, in which the root mean square error between the subjective quality scores and metric results was minimized by selecting the appropriate weights and thresholds.

For most distorted videos, the quality values computed by Eq. (10) fall in the interval $[0, 1]$. To prevent negative quality values, VQ is clipped at a lower threshold 0; on the other hand, VQ is also clipped at a higher threshold 1 to limit VQ values for excessively distorted videos, i.e.

$$VQ = \min\{\max\{VQ, 0\}, 1\} \quad (11)$$

3. EXPERIMENTAL RESULTS

To evaluate the performance of the proposed quality metric, a total of 362 video clips and the corresponding subjective measurements were evaluated, which included 320 VQEG FR phase 1 test clips [2], 30 Temporal Scalability test clips, and 12 Mobile test clips. The Temporal Scalability test was a single stimulus measurement to compare the performance of different temporal scalability parameters in the H.264/AVC codec. Four different video contents were employed with two resolutions: VGA and QVGA. 40 subjects participated in the test, about half of which had previous experience on visual quality assessment. The Mobile test was a double stimulus continuous quality scale measurement which employed three different contents with QCIF resolution to test the performance of H.264/AVC codec at four bit rates: 24, 32, 40, 48 kbps. 28 experienced subjects participated in the test. Because the different subjective experiments usually have different rating scales, different testing conditions, and many other test variables that change from one laboratory to another, it is difficult to compare or combine results of two or more subjective experiments directly. Pinson *et al.* proposed an objective method for combining multiple subjective data sets which can map the multiple subjective data sets onto a single common scale using the iterated nested least

TABLE I SELECTION OF S-T REGION SIZE

	SI	$[0, 3.563]$	$(3.563, 5.942]$	$(5.942, +\infty)$
Spatial size (pixels • pixels)		32 • 32	16 • 16	8 • 8
TI		$[0, 29.35]$	$(29.35, 51.67]$	$(51.67, +\infty)$
Temporal size (frames)		18	12	6

squares algorithm (INLSA) [11]. We used the INLSA to map all the subjective results to the range of [0, 1] in this paper.

The sizes of the S-T regions, especially the spatial region size, have a great influence on the computing speed. Experiments demonstrated that appropriate selection of S-T region sizes can reduce the computing time greatly while keeping the accuracy reasonably high. Based on the training with adequate sequences, we suggest using the interval thresholds of SI and TI in Table I to determine the S-T region sizes.

When constructing the proposed metric, half of the original uncompressed video scenarios and their corresponding distorted clips were used in training the coefficients and thresholds, and then the metric performance was evaluated with respect to the subjective measurements on the remaining sequences. To compare the performance of the proposed metrics, the results of the Yonsei method and NTIA general model were also computed. The Yonsei method measures the edge degradation by computing PSNR around the edge pixels, and it is comparatively simple and can be used in real time applications. NTIA general model computes several quality features, including spatial, chrominance, contrast, and temporal information, based on the fixed S-T region sizes, which is thought to be a comprehensive and accurate method. Furthermore, as a traditionally and widely used metric, PSNR is also taken as a benchmark. In the experiments, we used the NTIA CVQM software and our implementation of the Yonsei metric.

After getting the predictive qualities on the remaining clips, a non-linear regression was fitted first to the subjective MOS and metric results using a logistic function, as depicted in [2]. Pearson linear correlation coefficient and the root mean square error (RMSE) between the fitted results and the subjective MOS were selected to evaluate the prediction accuracy of the objective quality metrics. The quality model should predict a change in predictive quality that has the same sign as the change in subjective MOS, so Spearman rank-order correlation coefficient was also taken as another evaluation metric. Moreover, the computing time in the same experimental condition was selected to evaluate the computation complexity of these three metrics. Because the durations and image sizes of the test clips are different from each other, while VQEG FR phase I sequences take up a large proportion in our experimental data, we took the proportion of the average computing time regarding the time for computing PSNR for the VQEG sequences as the reference while excluding the time for video calibration. Table II gives the evaluation results on these quality metrics. In addition, Fig. 2 depicts the scatter plots of subjective MOS versus the metric results for these four metrics.

Based on the experimental results, the proposed quality metric performed better than the Yonsei method and PSNR, presumably because it takes into account the characteristics of HVS. The superior performance was also confirmed by a significance test on RMSE for each metric under 2-way repeated measures ANOVA. Moreover, since we use the simplified quality features and adaptive S-T region sizes, the computation of the proposed metric is faster than NTIA model, while the performance is satisfactory.

4. CONCLUSIONS

A fast and effective objective video quality metric was proposed in this paper. The spatial distortion parameters were characterized and modeled, and the temporal distortion caused by a frame freeze resulting from a packet loss was derived based on the spatial distortion parameters. The overall quality of the whole sequence was measured by the weighted average of the qualities of all the

TABLE II EVALUATION RESULTS OF THE METRICS

Criteria	Proposed	Yonsei	NTIA	PSNR
Pearson	0.843	0.814	0.886	0.701
RMSE	0.096	0.109	0.081	0.137
Spearman	0.821	0.783	0.865	0.651
Time proportion	18.3	12.7	36.1	1.0

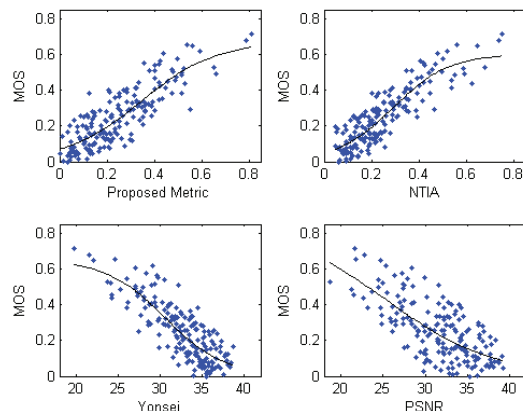


Fig. 2 Scatter plots of subjective MOS (after mapping into [0, 1] using INLSA) versus the metric results (before fitting), the black line denotes the fitting curve.

temporal regions. The evaluation results with respect to the subjective measurements demonstrated that the performance of the proposed metric is promising and can be used in assessing the quality magnitude in real time measurement systems.

REFERENCES

- [1] B. Girod, "What's wrong with mean-square error," in *Digital Images and Human Vision*, A. B. Watson, Cambridge, MA: MIT Press, pp. 207-220, 1993.
- [2] ITU-T Tutorial, "Objective perceptual assessment of video quality: full reference television," *ITU-T Telecommunication Standardization Bureau*, 2004.
- [3] ITU-T J.247, "Objective perceptual multimedia video quality measurement in the presence of a full reference," Aug. 2008.
- [4] S. Winkler, *Digital video quality: vision models and metrics*, John Wiley & Sons, 2005.
- [5] S. Wolf, and M. Pinson, "Video quality measurement techniques," *NTIA Report 02-392*, Jun. 2002.
- [6] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal Processing: Image Communication*, vol. 19, no. 2, pp. 121-132, Feb. 2004.
- [7] C. Lee, S. Cho, J. Choe, et al., "Objective video quality assessment," *Optical Engineering*, vol. 45, no. 1, pp. 017004-1-017004-11, Jan. 2006.
- [8] A. R. Reibman, and D. Poole, "Predicting packet-loss visibility using scene characteristics," *Packet Video*, Nov. 2007.
- [9] ITU-T P.910, "Subjective video quality assessment methods for multimedia applications," Sep. 1999.
- [10] Y-K Wang, M. M. Hannuksela, V. Varsa, A. Hourunranta, M. Gabbouj, "The error concealment feature in H.26L test model," *ICIP2002*, pp. II-729-II-732, 2002.
- [11] M. Pinson, and S. Wolf, "An Objective Method for Combining Multiple Subjective Data Sets," *SPIE Video Commun. and Image Processing Conf.*, Lugano, Switzerland, Jul. 2003.