PERCEPTUAL TEMPORAL QUALITY METRIC FOR COMPRESSED VIDEO

Kai-Chieh Yang, Clark C. Guest, Pankaj K. Das

University of California, San Diego
Department of Electrical and Computer Engineering
E-mail: kcyang, cguest@ucsd.edu, das@cwc.ucsd.edu

Khaled El-Maleh
Qualcomm Inc.
kemaleh@qualcomm.com

ABSTRACT

A novel and effective temporal quality metric for compressed video is proposed in this paper. This metric is able to accurately estimate the perceived temporal degradation introduced by both consistent and inconsistent frame dropping. Unlike conventional temporal quality metrics using the frame rate as the measurement basis, the proposed approach uses the length of each frame dropping occasion as the estimation basis. This new measurement basis gives more insight into temporal quality analysis and more flexibility for designing the metric. A new concept of variable sensitivity for each group of frame dropping caused by temporal quality fluctuation is discussed and a corresponding mathematical model is proposed. Subjective experiments show that the sensitivity model provides significant improvement in the performance of temporal quality prediction. The final subjective results show that the quality score from proposed temporal quality metric are highly correlated with human perceived quality.

1. INTRODUCTION

With the fast pace of development in the wireless industry, video communication using the video phone, video streaming and video broadcasting is becoming more and more popular. In order to transmit such rich multimedia content via a relatively narrow band network, the size of the original video signal must be reduced by compression, and the viewer’s perceived video quality is degraded because of data loss during compression. In order to achieve the best balance between compression efficiency and humans perceived quality, a number of different video encoding standards have been established. The goal of decreasing video data size can be achieved by either truncating some high frequency signal or by skipping frames. The formal one introduce spatial distortion and the later one cause temporal quality degradation, such as motion jerkiness, and jittering. Research into quantifying jerkiness and jittering, which happen very frequently in real time video communication, i.e. video telephony and broadcasting, is still lacking.

When jerkiness occurs, viewers perceive regular frame freezing and discontinuous motion. Jitter is usually caused by unexpected frame skipping by the coder or frame loss during wireless transmission. In this case, viewers perceive irregular frame freezing and various impacts of discontinuous motion. In this paper, the temporal quality is generalized into two types, which are (1) Consistent frame dropping and (2) Inconsistent frame dropping. Consistent frame dropping covers uniform single/aggregated frame skipping/loss on the coder side and inconsistent frame dropping covers the case of variable size aggregated frame skipping/loss from both the coder and the channel. With same amount of frame loss, inconsistent frame dropping tends to have a greater impact on perceived temporal quality degradation than consistent frame dropping. Therefore, this paper proposes a novel objective temporal quality metric - Perceptual Temporal Quality Metric (PTQM). Several important features and contributions of the proposed metric are:

1. The proposed method accurately estimates humans perceived visual discomfort that is induced by temporal discontinuity under various combinations of scenes and different video motion activity.
2. Temporal degradation caused by both consistent and inconsistent frame loss can be accurately measured.
3. The impact of local temporal quality fluctuations is investigated and a mathematical model is proposed.
4. The proposed metric utilizes the length of grouped frame dropping as the basic measurement unit and it provides more important attributes for temporal quality analysis.

The video bitstream is used as input for PTQM and the time stamp information are used as indices of each frame. Each frame is referred as a “group-dropping”. The dropping severity of each group-dropping is estimated by the normalized difference of the index of the current and previous frame. Afterward, the dropping severity of each group-dropping is adjusted by motion and the perceptual importance. Finally, the temporal quality score is the linear combination of all adjusted dropping severity.

The remainder of this paper is organized as follows. Section 2 provides a literature survey of temporal quality related research. Section 3 introduces some important but not yet resolved issues of temporal quality assessment. Section 4 gives an elaborate explanation of the proposed metric and provides a detailed study of the impact of the temporal quality fluctuation. Section 5 shows the experimental results and their associated subjective scores. Conclusions are summarized in Section 6.

2. RELATED WORKS

Among objective metric designs, Feghali et al. [1] uses frame rate as the scaling factor to adjust PSNR and output a spatial-temporal quality score. In Refs. [2] and [3], a jerkiness metric based on frame rate and motion activity is proposed. This metric in Ref. [2] is further applied to guiding a new transcoder. In Ref. [4], the inter-frame correlation is used to determine the location of lost frames. Some post-processing is conducted on the number...
of lost frames to extract several indices, such as the duration of group dropping and the number of group dropping occurrences, etc. The final temporal quality score is determined based on the ad-hoc analysis of those indices. Pastrana-Vidal and Gicquel [5] proposed a no-reference objective metric for measuring the fluidity impairments in video service. This metric responds to their previous work [6] and takes the density of group dropping into account. Lost frames are detected by the inter-frame dissimilarity on the decoder side. After thresholding, noticeable fluidity breaks are obtained. Each fluidity break is weighted by a function of the pixel variation of the last frame at the end of the freeze and the first frame appearing after the freeze. This stage tries to map the fluidity break to different types of motion. Afterward, the fluidity break is further adjusted by a function of the fluidity break density. The paper claims that the contribution to temporal quality degradation of the fluidity breaks with higher number of occurrence is less significant. In another word, with the same amount of frame loss, the temporal quality with scattering fluidity break is better than aggregated fluidity break. Watanabe et al. [7] studied the temporal distortion with different combinations of small group dropping with a fixed amount of frame loss subjectively. Based on that, they tuned a logarithmic function specifically for different combinations of frame loss in each sequence. This work provides important evidence that same amount of frame loss within one sequence could lead to a different level of subjective temporal degradation through different combinations of aggregated frame loss. It shows the prediction accuracy of the logarithmic function can be improved by tuning parameters according to the duration of each grouped frame loss.

### 3. PROBLEM STATEMENT

Many of the previous works estimate the temporal quality based on average frame rate and motion activity. However, many practical observations show that estimating temporal quality based on frame rate is insufficient because it assumes the dropped frames are distributed through the whole sequence evenly. In fact, with the same frame rate, inconsistent frame dropping can introduce a dramatically different quality impact as compared to consistent frame dropping. Inconsistent frame dropping refers to the length of each frame dropping occasion is quite different. Consistent frame dropping generally refers to that dropping frames distribute evenly at many different positions within the video sequence. The inherent cognitive interpolative mechanism of humans causes very different responses to inconsistent frame dropping versus consistent frame dropping.

Most of previous researches assume (1) the non-dropped frames do not contribute to the temporal degradation at all and (2) the local temporal quality is independent of its neighboring group-droppings. Nevertheless, from subjective data, we have found that the sensitivity to each group-dropping not only varies with the number of frames lost but also the local temporal quality contrast to the neighboring group-dropping. Hence, we claim that each group-dropping should not be treated independently. The sensitivity to each group-dropping is related to the inter-group dropping dependency. In this paper, several experiments are conducted to investigate this problem. Based on the observations, a generic and accurate temporal quality metric is proposed.

4. SYSTEM DESCRIPTION

#### 4.1. Dropping Severity Estimator

In this paper, the assumption is that each frame has its own time stamp information from the bitstream after video data compression. If the time stamps for consecutive received frames show a gap, it is evident that one or more intervening frames have been dropped. The length of the gap, as determined by the time stamps, permits the determination of the number of consecutive frames that have been dropped, which is defined as the length of group-dropping. The length could be zero or greater; zero means no frame dropping occurs. Furthermore, the length of group-dropping is transformed into dropping severity \( s_{m,n} \) by

\[
s_{m,n} = \frac{1}{R-1} \left[ \frac{t_{m,n+1} - t_{m,n}}{T} - 1 \right],
\]

where \( s_{m,n} \in (0, 1) \), \( t_{m,n} \) is the time stamp associated with each frame, \( T \) is the time interval between frames, \( m \) is the index of scenes and \( n \) is the index of group-dropping. Also, the factor \( R - 1 \) is equivalent to the applicable maximum frame rate minus one frame, and is used to normalize the frame dropping severity value across different frame rates. We set the frame rate to 30fps in this paper. A sequence that plays at 1fps would have dropping severity equal to 1, and the sequence playing at full frame rate has a dropping severity of 0.

#### 4.2. Shot Boundary Determination

A video sequence may contain multiple scenes and each scene may be different in terms of the captured subjective matter. Therefore, one of the reasons for doing scene boundary detection is that video
sequences with similar content usually have similar motion activity. Another reason for scene boundary detection is that large displacements usually occur between the last frame of one scene and the first frame of the next scene; the motion activity is usually very high. However, since these kinds of motion do not contribute to temporal degradation at all, the motion activity caused by scene transitions should not be taken into account. 

In order to detect the location of scene boundaries from the bitstream, the approach from Ref. [8] is adopted.

4.3. Motion Activity Estimator

The input to the motion activity estimator are the motion vectors derived from the video bitstream. In general, a motion vector points from a video block in one frame to a substantially similar or identical video block in another frame, providing an indication of displacement. The motion activity estimator determines the general level of motion activity within a given frame based on the average magnitude of motion vectors of selected blocks in the frame. The motion activity of the nth group-dropping of the nth scene is

\[
ma_{m,n} = \frac{1}{N_{MB}} \sum_{i=1}^{N_{MB}} MA(v_{m,n,i})
\]

(2)

where

\[
MA(v_{m,n,i}) = \sqrt{v_{m,n,i,x}^2 + v_{m,n,i,y}^2}.
\]

(3)

\(N_{MB}\) is the number of total selected blocks for motion estimation, \(i\) is the block index and \(v_{m,n,i,x}, v_{m,n,i,y}\) are the motion vectors in the horizontal and vertical directions respectively. The motion activity is normalized as \(ma_{m,n} \in (1, 10)\).

The main discomfort from temporal degradation results from discontinuous movement while the video is played. However, not all motion vectors in a frame should be taken into account. Only a moving object can affect the level of temporal quality. Hence, estimating the motion activity by averaging the magnitude of motion vectors through all blocks will be misleading. Moreover, some macro blocks could be intra-coded because the motion of those macro blocks is too large and the motion estimator can not find the matched block within its search range. In that case, no motion vector information is available for those macro blocks, but they actually have large displacement. Based on the concerns stated above, the selected motion vectors must belong to one of the following categories:

1. Effective motion vectors: The motion vectors whose motion activity is larger than threshold \(T_{ma}\),

\[
v_1; \quad v_1 \subset V_{m,n}, \quad MA(v_1) > T_{ma},
\]

(4)

where \(V_{m,n}\) represents all motion vectors of frame \(n\).

2. Intra-coded motion vectors: For the macro blocks that are intra-coded because of the very high motion, the motion vector of the macro block will be assigned by a predefined value,

\[
v_j = (16, 16); \quad \text{mode}(j) = \text{Intra},
\]

(5)

where \(\text{mode}(j)\) is the coding mode of the \(j\)th macro block. Therefore, the selected motion vectors are \(v_{m,n,i} = v_1 \cup v_j\) and Eqn. (2) is applied to calculate the motion activity.

4.4. Mode Decision and Motion Mapping

Though the same amount dropping severity occurs, the viewer’s perceived discomfort may vary under different motion activity. For example, with the same number of dropped frames, a sequence A has a very static scene, i.e. mother and daughter, but sequence B has more high motion content, i.e. football. As a result, the introduced temporal negative impact on sequence B is much larger than sequence A. Therefore, an appropriate motion-mapping model is chosen to map the dropping severity \(s_{m,n}\) to \(s'_{m,n}\) and the model is formulated as

\[
s'_{m,n} = \gamma \left[ 1 - (1 - \alpha_{PTQM} ma_{m,n} - s_{m,n}) + 1 \right]
\]

(6)

and

\[
\gamma = \begin{cases} 1 & \text{if } ma_{m,n} > T_{mm} \\ 0 & \text{if } ma_{m,n} \leq T_{mm} \end{cases}
\]

(7)

where constant \(\alpha_{PTQM} = 11.5\) is obtained experimentally, \(\gamma\) is the factor that disables mapping for some low motion sequences whose motion activity is lower than \(T_{mm}\). The motion model is shown in Fig. (2).

In Fig. (2), each slice along the frame dropping severity axis represents a mapping function corresponding to a motion activity. As the motion activity increases, the slope of the mapping function increases accordingly, which means the same \(s_{m,n}\) is projected into higher \(s'_{m,n}\) with higher motion activity. Therefore, the same dropping severity is more noticeable for a high motion clip than for a low motion clip.

4.5. Temporal Fusion

Temporal quality degradation is mainly caused by frame loss during compression. Humans usually have higher tolerance to consistent frame dropping because of our inherent cognitive interpolation mechanism and the well preserved correlation of the remaining frames. When frames are dropped consistently, the content of the adjacent remaining frames is still very similar and the correlation is high. In addition, a human viewer has the inherent ability to interpolate the missing content and can get used to the down sampled frame rate easily. But in the inconsistent case, the frame loss does not occur in a fixed time frame and it can occur in groups. Compared with the first case, the human viewer has much less tolerance to this kind of negative impact because of high sensitivity to abrupt perception change and low correlation between the remaining frames. When a sequence is played with good temporal quality for a while the viewer starts to get used to this playing smoothness. When suddenly, several frames are dropped because of the
network condition, then viewers feel very uncomfortable because of the large difference in playing smoothness.

4.5.1. Temporal Quality Fluctuation

Based on the reasons above, a Temporal Quality Fluctuation (TQF) metric is designed to estimate the importance of each group-dropping. More detail is illustrated in Fig. 3. The quantity $s'_{m,n}$ is input into the temporal fluctuation estimator and the temporal fluctuation ($tf$) is determined by

$$tf_{m,n} = [s'_{m,n} - \frac{1}{I} \sum_{i=1}^{I} s'_{m,n}]^\beta,$$

where $I$ is the size of look-backward window, here it is set as 3, $\beta$ is an important factor used to differentiate similar group-droppings and it is set as 2. Afterward, because the range of fluctuation for each frame rate varies, $tf$ is normalized within 0 - 1 by

$$tf'_{m,n} = \frac{tf_{m,n}}{UB[R]},$$

where the upper bound $UB[R] = [d(R) - \frac{1}{I} \sum_{i=1}^{I} d(R)]^2$ and $d(R) = (30 - R)/29$. A non-linear TQF

$$TQF(tf'_{m,n}) = \eta \left[ 1 - \left( \frac{tf'_{m,n}}{\beta} \right)^\psi \right]$$

is applied on $tf'_{m,n}$, where

$$\psi = \begin{cases} 
4 & \text{if } 20 \leq R \leq 30 \\
8 & \text{if } 19 \leq R \leq 14 \\
9 & \text{if } 1 \leq R \leq 13 
\end{cases}$$

constant $\eta = 1.25$ is obtain experimentally, $\kappa$ is a balance factor that controls the dominance between the effect of temporal fluctuation and the length of consecutive frame loss; the value of $\kappa$ is provided at Table 1 and it is obtained experimentally. As shown in Table 1, $\kappa$ decreases as frame rate decreases, which means the effect from quality fluctuation is less observable at low frame rates because humans’ perceived temporal quality is mainly dominated by the amount of consecutive frame loss.

<table>
<thead>
<tr>
<th>Motion Activity</th>
<th>$R \in {30, 20}$</th>
<th>$R \in {20, 10}$</th>
<th>$R \in {10, 1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low motion</td>
<td>10</td>
<td>2.6</td>
<td>1.5</td>
</tr>
<tr>
<td>Medium motion</td>
<td>9.7</td>
<td>2.5</td>
<td>1.5</td>
</tr>
<tr>
<td>High motion</td>
<td>9.6</td>
<td>2.4</td>
<td>1.4</td>
</tr>
</tbody>
</table>

Table 1. Value of $\kappa$

After weighting by TQF and normalization, the final temporal quality score of the whole sequence is estimated by

$$Q = \delta_T + \beta_T \cdot \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N}(1 - TQF(tf'_{m,n})s'_{m,n}).$$

where $\delta_T$ and $\beta_T$ are the normalization factors and $\delta_T = 1$, $\beta_T = 4$, $Q \in \{5, 1\}$ and a higher value means better temporal quality, $M$ is the number of shots in one sequence and $N$ is the number of group-dropping for each scene. The temporal quality of each scene is calculated by averaging all the dropping severities weighted with the output of TQF. The overall temporal quality of the input sequence is calculated by simply averaging the quality score of all scenes. Since each scene is already weighted by motion mapping, the average of all scenes does not require further weighting.

5. SUBJECTIVE EXPERIMENT

The purpose of the subjective experiment is to measure the performance of the PTQM. Six standard sequences are involved in this experiment. The sequences can be classified into three groups based on their motion activity. Low motion sequences: Container, Mother and Daughter, medium motion sequences: Carphone, Highway and high motion sequences: Rugby, Football. The selected sequences have wide coverage of different motion activity and content. The original sequences are played at 30fps and the playing duration is 10 seconds long. The video sequences are sampled in 4:2:0 with QCIF size.

First, three of these sequences are frame rate down sampled to 23fps, 15fps and 10fps respectively. The different frame rates are generated by dropping frames artificially and the lost frames are replaced by duplicating the last frame before frame loss occurs. All the testing sequences are not compressed in order to isolate the temporal effect from spatial distortion. Since different combinations of group-dropping introduce considerably different subjective impact, three sub-cases with different combinations of group-dropping were generated for each frame rate of each scene. The first case is the combination of group-dropping with lowest fluctuation, which is consistent frame dropping. The second sub-case represents a combination with moderate fluctuation and the third case represents the highest fluctuation. A detailed profile of each sub-case is given in Table 2. In Table 2, the $a - b$ means $a$ group-droppings and each group has $b$ consecutive dropped frames. The notation $(a', b')$ means two group-dropping only and each group has $a'$ and $b'$ consecutive dropping frames respectively. The experiment was carried out using a standard personal computer with Samsung 17’ LCD display. The lighting condition of the viewing environment and the viewing distance are not fixed and allows viewers to adjust it based on their own comfort. Viewers are asked to score the testing video sequences on a five-level scale with the corresponding semantic meanings, which are (5) Excellent, (4) Good, (3) Fair, (2) Poor and (1) Bad, and the increment between each grade is 0.25 that provides more flexibility for viewers to score. A total 20 examiners participated in the experiment, which includes twelve non-expert and eight expert viewers. The adopted experimental methodology is similar to the Double-Stimulus Continuous Quality Evaluation (DSCQE) [9]. An education section is conducted prior to running the testing sequences by showing all the reference sequences played at full frame rate. Afterward, the testing and reference sequences are shown in random order across different sub-cases and content. The final MOS data is the difference of the MOS score between reference and testing sequences.

The performance of the objective metric is evaluated by calculating the correspondence between model-predicted quality score and subjective data. The correspondence is quantified by the Pearson, Spearman correlation and Root-Mean-Square-Error (RMSE),
by both consistent and inconsistent frame loss, the local temporal
performed. In order to estimate the temporal degradation caused
accurate motion mapping associated with scene cut detection is
important attribute when estimating the perceived temporal quality,
rithms at the frame level in the future. Since motion is an im-
can be employed to develop temporal quality enhancement algo-
this new basis, PTQM can measure local temporal quality and it
quality contrast is used to emulate the sensitivity to each group-
dropping. The final temporal quality is obtained by averaging all
weighted group-droppings through all scenes. Based on the com-
parison between PTQM’s output and MOS data and the associated
 correspondence measurement, the high accuracy of humans per-
ceived temporal quality prediction is proved.

In future work, PTQM will be deployed in a synthetic quality
evaluation system to provide video quality feedback from tempo-
rnal aspects. Also, in the temporal quality enhancement wise, the
PTQM metric can serve as guidance for a new frame-level rate
controller. It will lead the rate controller to reduce the subjective
impact and meet the bits budget by skipping more frames in low
motion segments and spending more bits in the medium or high
motion segments. Furthermore, the bits saving in low motion seg-
ments can be more aggressive by dropping frames consecutively,
as long as the quality stays within the tolerable range.

6. CONCLUSION

A novel and reliable temporal quality - PTQM is proposed. The
PTQM estimates temporal quality by observing the amount of frame
loss, object motion, and local temporal quality contrast. Instead of
frame rate, PTQM measures the temporal degradation based on
the number of consecutive frames lost between each frame. With
this new basis, PTQM can measure local temporal quality and it
can be employed to develop temporal quality enhancement algo-
ithms at the frame level in the future. Since motion is an im-
portant attribute when estimating the perceived temporal quality,
accurate motion mapping associated with scene cut detection is
performed. In order to estimate the temporal degradation caused
by both consistent and inconsistent frame loss, the local temporal

Table 2. Different combination of sub-cases of each frame rate

<table>
<thead>
<tr>
<th>Frame Rate</th>
<th>C_P</th>
<th>C_S</th>
<th>C_R</th>
</tr>
</thead>
<tbody>
<tr>
<td>23fps</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10fps</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15fps</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10fps</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3. Performance evaluation of each sequences at 23fps

<table>
<thead>
<tr>
<th>Frame Rate</th>
<th>C_P</th>
<th>C_S</th>
<th>C_R</th>
</tr>
</thead>
<tbody>
<tr>
<td>10fps</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15fps</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Performance evaluation of each sequences at 15fps

<table>
<thead>
<tr>
<th>Frame Rate</th>
<th>C_P</th>
<th>C_S</th>
<th>C_R</th>
</tr>
</thead>
<tbody>
<tr>
<td>23fps</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10fps</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5. Performance evaluation of each sequences at 10fps

<table>
<thead>
<tr>
<th>Frame Rate</th>
<th>C_P</th>
<th>C_S</th>
<th>C_R</th>
</tr>
</thead>
<tbody>
<tr>
<td>10fps</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15fps</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

7. REFERENCES

metric for video sequences with temporal scalability,” in
Proc. IEEE International Conference on Acoustics, Speech,

transcoder driven by the jerkiness,” in Proc. IEEE Interna-
tional Conference on Multimedia and Expo, 2005, pp. 1452–
1455.

Yang, “Measuring the negative impact of frame dropping on
perceptual visual quality,” in Proc. of the SPIE Human Vision
and Electronic Imaging X, March 2005, vol. 5666, pp. 554–
562.

“Objective quality evaluation of video services,” in Proc. of
the Third International Workshop on Video Processing and
Quality Metrics for Consumer Electronics, January 2006.

assessment of video fluidity impairments using a no-reference
metric,” in Proc. of the Third International Workshop on Video
Processing and Quality Metrics for Consumer Electronics,
January 2006.

C. Hocine, “Sporadic frame dropping impact on quality per-
ception,” in Proc. SPIE Electronic Imaging, Human Vision

quality assessment method for freeze distortion based on
freeze aggregation,” in Proc. SPIE Electronic Imaging, Image Quality and System Performance, 2006, p. 60590Y.
