QoE Enhancement of Audiovisual IP Communications by Joint Application of FMO and a QoE–Based Video Output Scheme

Shuji Tasaka, Akifumi Hirashima, and Toshiro Nunome
Department of Computer Science and Engineering,
Nagoya Institute of Technology, Nagoya 466–8555, Japan
Email: tasaka@nitech.ac.jp, tiery@inl.nitech.ac.jp, nunome@nitech.ac.jp

Abstract—This paper studies a method of enhancing QoE (Quality of Experience) in H.264–video and audio IP transmission by jointly applying FMO (Flexible Macroblock Ordering) and SCS (Switching between error Concealment and frame Skipping), which is a QoE–based video output scheme simply based on the receiver’s operation. We conducted objective and subjective experiment on the method for combinations of two slice group map types (interleaved and dispersed), four contents and three picture patterns (I, I+4P’s and I+14P’s); we measured QoS (Quality of Service) and then assessed QoE in terms of the psychological scale. As a result, we find that for the patterns I+4P’s and I+14P’s, combinations of SCS and the dispersed map type work best, whereas for the pattern I, the frame skipping only (i.e., SCS without FMO) achieves the highest QoE.

Index Terms—QoE, audiovisual IP communications, H.264, FMO, SCS

I. INTRODUCTION

Video–Audio transmission is a key function in multimedia application services over IP networks, which are becoming increasingly important in our daily life. Application services of this type include IPTV, streaming such as YouTube, videophone like Skype, and network gaming, which play important roles in consumer communications.

Since the IP networks inherently offer best–effort services, the users may suffer transmission impairment such as packet loss, error and delay jitter; as measures to cope with the impairment, a variety of methods have been proposed and in practical use.

FMO (Flexible Macroblock Ordering) of the H.264/AVC video coding [1] is one of the methods of this type; it modifies the way macroblocks are associated with slices, which can be decoded independently. FMO has been shown to be effective as an error resilience technique by many researchers (e.g, see [2]–[4]). Therefore, we can expect that FMO enhances QoS (Quality of Service) offered to the users.

The flexible ordering of slice transmission by FMO can improve the effectiveness of video error concealment, which conceals the visual effects of packet loss and error; it is performed by exploiting the spatial or temporal correlation with adjacent data [5], [6]. The spatial approach restores missing blocks by interpolating each missing block from its neighboring blocks. The temporal approach, which is usually applied to inter–coded frames, replaces a missing block with some appropriate block in a previously decoded frame. Either approach prevents the degradation of picture quality, i.e., the spatial quality of video, while keeping the output frame rate (i.e., temporal quality) approximately original.

It should be noted here that the error concealment may cause residual errors; such errors can propagate to the succeeding frames. In this case, the spatial quality of the output video stream deteriorates compared to the original stream; it degrades QoE (Quality of Experience), which is the subjective quality perceived by the end–user [7] and therefore the ultimate target of QoS control. This suggests that if we have so many lost/erroneous blocks that the error concealment cannot work sufficiently, skipping the frame can prevent the degradation of QoE; this type of conditional frame skipping is a special case of the (unconditional) frame skipping, which does not decode a frame unless all packets (i.e., slices) of the frame are correctly received. This (unconditional) frame skipping can be easily implemented in video decoding.

It is obvious that the (unconditional) frame skipping keeps the spatial quality of output pictures original, since only frames whose slices are all correctly received are actually output; however, once a frame is skipped, the skipping continues until an intra–coded frame is decoded. Thus, the frame skipping decreases the output frame rate; it leads to lower temporal quality.

Noting that QoE is affected by both spatial quality and temporal quality of video, we easily conceive an idea that an appropriate mixture of the error concealment and (conditional) frame skipping can provide the highest QoE; this is the basic idea of SCS (Switching between error Concealment and frame Skipping) [8]. SCS switches between the error concealment and the frame skipping according to the percentage of error–concealed video slices in a frame. The effectiveness of SCS in the case where FMO is not employed has been shown in [8].

This paper applies SCS to the transmission of audio and H.264–video with FMO over IP networks; then, by measuring QoS and QoE, it demonstrates that joint application can achieve higher QoE than the case of using FMO only.

The remainder of the paper is organized as follows. Section II specifies slice group map types of FMO employed in this paper and gives an overview of SCS. Section III describes an experimental network, contents to be assessed, and a method of assessing QoE. Section IV presents experimental results and examines how effective the joint application is. Section V concludes the paper.
II. JOINT APPLICATION OF FMO AND SCS

In order to enhance QoE, we employ both FMO and SCS in this paper. FMO involves both encoder and decoder, while SCS is purely a receiver’s operation. FMO associates each macroblock to one of slice groups, each of which consists of one or more slices, by the macroblock to slice group map. Each slice is a sequence of macroblocks which are processed in the order of a raster scan; it can be decoded independently of the other slices. ITU-T Rec. H.264 [9] presents seven slice group map types; in this paper, we employ the interleaved and the dispersed. The two slice group map types used in this paper are illustrated in Fig. 1, where the numbers represent the slice group identification (ID) numbers. In this paper, we create an IP packet using a slice and transmit the packets in the order of the slice group ID number.

SCS switches from the error concealment to the frame skipping once the percentage of error–concealed slices in a frame exceeds a threshold value $T_h$; the frame skipping continues until an intra–coded frame is decoded. Note that the case of $T_h = 100\%$ is equivalent to the pure error concealment technique with FMO. On the other hand, $T_h = 0\%$ corresponds to the (unconditional) frame skipping; a picture with one or more lost slices is skipped without error concealment.

SCS exploits a QoE tradeoff relationship between spatial and temporal quality for video packet loss. Although SCS is a video output scheme, it can reflect the effect of audio on the overall QoE; this is because we can adjust the weight of video information in the output audiovisual stream through the threshold value, depending on whether the content is audio–dominant like music video or video–dominant like sport. Thus, SCS takes into consideration the cross-modal interaction between video and audio.

How to select the threshold appropriately is a key issue in SCS; see [8] and [10] for the selection methods.

III. EXPERIMENTAL METHODOLOGY

By experiment, we examine how effective the joint application of FMO and SCS is. For that purpose, we built an experimental IP network over which audio and H.264–video streams are transmitted. We then conducted subjective experiment of assessing QoE of the output audio–video streams by using many subjects (i.e., assessors).

(1) Experimental network

As shown in Fig. 2, a media sender (MS) transmits a pair of audio and video streams to a media recipient (MR); at the same time, as the interference to the audio–video streams, a Web server (WS) sends Web traffic generated by WebStone 2.5 [11] to a Web client (WC). WebStone generates Web client processes on the WC; those client processes retrieve specified files from the WS continuously. Table I shows the set of files to be retrieved in our experiment.

<table>
<thead>
<tr>
<th>file name</th>
<th>size [kbyte]</th>
<th>probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>file500.html</td>
<td>0.5</td>
<td>0.350</td>
</tr>
<tr>
<td>file5k.html</td>
<td>5.0</td>
<td>0.500</td>
</tr>
<tr>
<td>file50k.html</td>
<td>50.0</td>
<td>0.140</td>
</tr>
<tr>
<td>file500k.html</td>
<td>500.0</td>
<td>0.009</td>
</tr>
<tr>
<td>file5m.html</td>
<td>5000.0</td>
<td>0.001</td>
</tr>
</tbody>
</table>

In the experiment, the number of the Web client processes were set to 20, 30, 40, 50, 75 and 100. As the number of the processes increases, the amount of the Web traffic becomes larger, and therefore packet loss occurs more frequently.

The MS transmits an audio stream and the corresponding video stream as two separate transport streams with RTP/UDP. The information unit for transfer between the application layers is referred to as the MU (Media Unit). A video MU is defined as a video frame, and an audio MU as a constant number of audio samples. The MR exerts playout buffering control of 1 second to absorb delay jitters of received MU’s. For video encoding and decoding, we utilize the H.264/MPEG–4 AVC reference software JM13.2 [12].

Table II gives specifications of audio and video used in the experiment. A picture (i.e., a video MU) is divided into 15 slices each of which corresponds to a UDP datagram.

As the error concealment technique in this paper, we employ the one implemented in JM13.2. A missing block in an I frame is interpolated from its four neighboring blocks. For P–frames, we use a temporal approach of Frame Copy; it simply replaces the missing block with the spatially corresponding one of the previously output frame.
In the experiment, we set the threshold value \( T_h \) of SCS to 100%, 40%, 20% and 0% as in [8].

### TABLE II
<table>
<thead>
<tr>
<th>specifications of audio and video</th>
</tr>
</thead>
<tbody>
<tr>
<td>audio coding scheme</td>
</tr>
<tr>
<td>audio MU size [byte]</td>
</tr>
<tr>
<td>audio average bit rate [kb/s]</td>
</tr>
<tr>
<td>video coding scheme</td>
</tr>
<tr>
<td>image size [pixel]</td>
</tr>
<tr>
<td>number of slices in a picture</td>
</tr>
<tr>
<td>number of slice groups</td>
</tr>
<tr>
<td>slice group map type</td>
</tr>
<tr>
<td>video average MU rate [MU/s]</td>
</tr>
<tr>
<td>picture pattern (GOP)</td>
</tr>
<tr>
<td>recording time [s]</td>
</tr>
</tbody>
</table>

(2) Contents

Since the content type affects QoE, we have prepared two types, *sport* and *music video*, for each of which we have made two contents: sport 1 (aerobics), sport 2 (a racing car), music video 1 (a sitting man playing the ukulele), and music video 2 (a female singer playing the piano and singing while dancing). Sport has been selected as a video–dominant content type, where video plays a more important role than audio, while music video is considered audio–dominant.

The application of FMO incurs some overhead whose amount depends on the employed slice group map type; in general, the dispersed produces more overhead than the interleaved. Therefore, in the experiment, we adjusted the quantization parameters of the dispersed and the interleaved so that the two map types result in approximately the same average encoding bit rate of video.

Table III shows the video average bit rates used in the experiment; it also presents the value of \( T(I)(T_i p e r p e c t u a l ~ I n f o r m a t i o n) \) for each content. The TI measure is defined in ITU–T Rec. P.911 [13]; it indicates the amount of temporal changes of a video sequence. A higher value implies higher motion. The TI values in this table have been calculated by eliminating the effect of scene changes. Note that the TI value is not used in the operation of SCS; it is shown just for information of the degree of video motion. From Table III, we find that the second content of the same type (say sport 2) has higher motion video than the first one (say sport 1).

(3) QoE Assessment

We have recorded the audio–video streams output at MR as *stimuli* for the subjective experiment, where assessors performed the *Absolute Category Rating* with the following five–level quality scale: “excellent” assigned score 5, “good” 4, “fair” 3, “poor” 2 and “bad” 1. Instead of taking an average of the scores for a stimulus over all assessors (i.e., calculating *MOS* (*Mean Opinion Score*)), we applied the *law of categorical judgment* to the scores in order to obtain the *psychological scale*, which can represent the human subjectivity more accurately than MOS. See [14] for further details.

In the experiment, we have also measured objective QoS parameters such as the MU loss ratio.

In this way, we had 576 stimuli because of four contents, two slice group map types, four \( T_h \) values, three picture patterns and six kinds of the number of Web clients. By adding 56 dummy data to the 576 stimuli, we prepared 632 stimuli to be assessed. We presented the 632 stimuli to 39 assessors by putting them in a random order for each assessor, using a PC with headphones and a 17–inch LCD display. The assessors are 15 Japanese male students and 24 female students in their twenties. It took about three hours including break time for an assessor to assess all the stimuli.

### IV. EXPERIMENTAL RESULTS

This section calculates the psychological scale from the result of the subjective experiment. We then examine the effectiveness of the joint application of FMO and SCS.

We applied the law of categorical judgment to all the results of the 576 stimuli together. We then carried out *Mosteller’s test* [15] for a test of the goodness of fit. By removing 46 stimuli which give large errors of Mosteller’s test, we found that a significance level of 0.05 cannot reject the hypothesis that the observed value equals the calculated one. Thus, the 530 (= 576 – 46) stimuli produced the psychological scale.

We can select an arbitrary origin in the psychological scale since it is an *interval scale* [16]; we set the minimum value of the psychological scales for the 530 stimuli to the origin. Under this condition, we also calculated the lower boundaries of the categories as shown in the figures below.

Figures 3 through 10 plot the psychological scale versus the number of Web client processes for the four \( T_h \) values. Straight broken lines parallel to the abscissa in the figures represent the lower boundaries of the categories. Figures 3 and 4 present the result of sport 1 with the picture pattern I for the interleaved type and that for the dispersed type, respectively. Figs. 5 and 6 show cases of I+4P’s in sport 2 for the interleaved and the dispersed, respectively, while Figs. 7 and 8 correspond to I+4P’s. Figures 9 and 10 display the results of music video 1 with I+4P’s. Note that the results removed by the Mosteller’s test are not shown in the figures.

In addition, as an example of the measured objective QoS parameters, we plot the video MU loss ratio in Fig. 11, which corresponds to sport 1 of the dispersed type for the pattern I.

In the figures, we immediately see that the psychological scale value tends to decrease as the number of Web client processes increases, because the increase in the amount of
Fig. 3. Psychological scale versus number of Web client processes (sport 1, interleaved, Picture pattern: I).

Fig. 4. Psychological scale versus number of Web client processes (sport 1, dispersed, Picture pattern: I).

Fig. 5. Psychological scale versus number of Web client processes (sport 2, interleaved, Picture pattern: IPPPP).  

Fig. 6. Psychological scale versus number of Web client processes (sport 2, dispersed, Picture pattern: IPPPP).  

Fig. 7. Psychological scale versus number of Web client processes (sport 2, interleaved, Picture pattern: I+14P’s).

Fig. 8. Psychological scale versus number of Web client processes (sport 2, dispersed, Picture pattern: I+14P’s).  

Fig. 9. Psychological scale versus number of Web client processes (music video 1, interleaved, Picture pattern: I+14P’s).  

Fig. 10. Psychological scale versus number of Web client processes (music video 1, dispersed, Picture pattern: I+14P’s).
interference traffic degrades the output quality of the audio–video streams mainly owing to packet loss. When the number of Web client processes is 20, however, we have only observed packet loss as seen in Fig. 11; the difference in the psychological scale value among the four Th values is due to the fluctuation of the measurement. Therefore, in the following discussion, we will focus on the results for 30 and more Web client processes and examine each of the three picture patterns in turn.

1) Picture pattern I: From Figs. 3 and 4, we see that $T_h = 0\%$ clearly achieves the highest QoE. This is because skipping a frame with the picture pattern I incurs no skipping of the succeeding frames and therefore keeps the degradation of the temporal quality (i.e., decrease in the frame rate) minimum, while the spatial quality (i.e., the picture quality) remains original. Thus, only the frame skipping is sufficient for the pattern I; FMO is not necessary. It should be noted that $T_h = 100\%$ corresponds to the application of FMO only.

2) Picture pattern I+4P's: A comparison of Figs. 5 and 6 indicates that the selection of $T_h = 40\%$ is appropriate and that the dispersed map type provides higher QoE than the interleaved map type. Note that skipping a frame with this picture pattern can incur skipping the succeeding frames; this implies that $T_h = 40\%$ is a suitable compromise between the temporal quality degradation and the spatial one. In the error concealment, the dispersed map type can utilize more information than the interleaved one; the former type can have macroblocks with different slice group ID numbers in the eight directions (i.e., right, left, above, upper right, upper left, below, lower right and lower left), compared to the two directions (above and below) of the latter type, as we can see in Fig. 1. This provides the dispersed type with more powerful error concealment ability.

3) Picture pattern I+14P's: From Figs. 7 and 8, we see that the conventional application of the dispersed map type only (namely, $T_h = 100\%$) is a good choice for video–dominant contents with high motion video. This result is understandable by noting that if a frame with the picture pattern I+14P's is skipped, many succeeding frames are also skipped, and the frame rate decreases largely; consequently, the pure error concealment with the dispersed map type is a good strategy for the picture pattern I+14P's.

In audio–dominant contents with low motion video, which correspond to Figs. 9 and 10, however, we make a different observation; the difference in QoE between the dispersed and the interleaved is small. This is because the audio plays a more important role than the video in this type of contents, and therefore video error concealment does not dominate QoE enhancement.

V. Conclusions

We saw that a purely receiver’s operation SCS can enhance QoE, compared to using FMO only. We also observed that the effectiveness depends on the picture pattern and the content type. For the picture pattern I, SCS without FMO (the frame skipping only) achieves the highest QoE, whereas combinations of SCS and the dispersed map type work best for I+4P's and I+14P's. Note that we can easily implement SCS at the receiver without changing the other parts of the system; this is suitable for consumer communications environments.

As future work, we will examine other kinds of content types and picture patterns to support the effectiveness of the joint application. We also need to perform statistical analysis of a larger set of experiments.

Acknowledgment

This work was supported by the Grant-In-Aid for Scientific Research of Japan Society for the Promotion of Science under Grant 21360183.

References