IMPLEMENTING DYNAMIC GOP IN VIDEO ENCODING

Xiaodong Gu, Hongjiang Zhang

Microsoft Research Asia, 49 Zhichun Road, Beijing 100080, China
Phone: 86-10-62617711-{3194, 5791}; Email: {i-xiaogu, hjzhang}@microsoft.com

ABSTRACT

Techniques for dynamic GOP in video encoding are proposed in this paper. Comparing to the average-sampling of current standard fixed GOP structure, with dynamic GOP the encoder has great freedom in I-, P- and B-frame selection. The dynamic GOP structure keeps temporal importance information of frames in the encoded video, which helps in adaptation to bandwidth decrease and temporal-SNR quality tradeoff decision. It also helps in storing more information with restricted resource, results in code efficiency increasing and better user perception. An effective temporal importance metric is also proposed in this paper. The results are validated by experiments.

1. INTRODUCTION

The efficient streamed delivery of compressed digital video has received considerable attention from both the research and application community. To stream video over the dynamic and unpredictable best-effort Internet, the sending rate of streamed video should be adapted to the varying network conditions. Dropping frames, if it can try to maximize the user satisfaction, is one of the major techniques for the rate adaptation to bandwidth decrease which has long been studied [1,2] because of its efficiency and simplicity. But since the frame-dropping algorithm should be applied for each client applying because of their varying network conditions, it will greatly increase total server calculation. And the frame dropping algorithm will be more and more complicated for several reasons. First, video SNR quality can also be certain extent sacrificed in network decreasing which lead to a tradeoff between SNR [7] and temporal quality. Second, though a lot of MCI (Motion Compensated Interpolation) methods have been proposed to compensate the dropped frame, we still do not want to drop whole frames in some applications.

In a traditional communication system, the encoder compresses the input video into a bit rate which is streamed to less than and closed to the channel capacity, and the client received these bits and reconstructed the video, as shown in Fig.1. When bandwidth decreased, the streaming process will try to sacrifice as little total video quality (including SNR quality and temporal quality) as possible in order to fit the decreasing bandwidth. When the server receives a large amount of client applying simultaneous, the server calculation will be greatly increased. So, can the streaming process be simplified?

In fact it can. If the common and similar features are abstracted in the encoding process, this will greatly decrease the streaming process calculation. We will study the problem in this paper.

Fig.2 shows a typical GOP structure where I-frame is coded independently, P-frame predicts nearest preceding I- or P-frame and B-frame predicts both nearest preceding and upcoming I- or P-frames. The GOP structured is fixed before a video is encoded and there are constant number of P-frames between two adjacent I-frames and constant number of B-frames between two adjacent I- or P-frames. GOP structure is corresponding to FGST (or FGS+) scalability structure as shown in Fig.3 [3]. In FGST structure, aside from the FGS layer in which each frame of base layer is enhanced in SNR quality, the base layer + FGST layer construct the GOP structure, here FGST layer enhance the temporal quality of base layer frames. The arrow tail in Fig.3 defines the reference (prediction) frames of P- and B-frames.

The research results of FGS+ show that until the SNR quality improves to an acceptable level, the users prefer that the
additional bandwidth be used to enhance the SNR quality. Thus the FGST layer has the lowest priority in bandwidth allocation and will be the first sacrifice at bandwidth decreasing. On the other hand, the base layer will be guaranteed to transmit regardless of bandwidth decreasing. Intuitively, we would like that the base layer (I- and P-frames) contains as much information as possible.

But in fixed GOP structure the base layer frames are constantly selected according to the GOP structure, we are not able to store more information in base layer with limited bit rate. In this paper, we will consider a dynamic GOP structure in video encoding which means number of B-frames between adjacent I- or P-frames and number of P-frames between adjacent I-frames are not fixed. It will lead to:

- With a dynamic GOP structure, the base layer can be combined of frames containing relatively more information. Thus in bandwidth decreasing, when FGST layer is fully or partially discarded the client can still receive as much information as possible.
- The dynamic GOP structure contains the drop-frame information. Just as do a frame-dropping in encoding process, the less important frames are dropped to the FGST layer and the number of FGST frames between two adjacent base layer frames indicates their temporal importance. This will in fact store as much temporal information as possible through GOP structure.
- By dynamic GOP structure we have fixed the temporal importance in coded video data, which will help in deciding tradeoff between SNR and temporal quality.
- Since in FGST structure, both the FGS and FGST frames are predicted from base layer frames and the base layer contains more information with dynamic GOP structure than with fixed GOP structure, surely the encoding efficiency will increase to some extent.

An important point is, it doesn’t take much encoding time to implement the dynamic GOP structure and it doesn’t break the indexing function of classic GOP structure.

This paper is organized as follows. Section 2 describes the video temporal importance and details of dynamic GOP structure are provided in section 3. Section 4 is experimental results and section 5 concludes the paper.

2. TEMPORAL IMPORTANCE OF FRAMES

There are a lot of research works about detecting temporal importance of video frames before. We will use the method similar to that introduced in [4] where an important metric called perceived motion energy (PME) is defined for frame at time \( t \) of a video.

\[
\text{PME}(t) = \text{Mag}(t) \cdot \alpha(t)
\]

In the definition, \( \text{Mag}(t) \) is the average magnitude of motion vectors for all macro blocks and \( \alpha(t) \) represents the percentage of the dominant motion direction, it is defined as

\[
\alpha(t) = \frac{\max(AH(t,k), k \in [1,n])}{\sum_{k=1}^{n} AH(t,k)}
\]

The angle in \( 2\pi \) is quantized into \( n \) angle ranges. Then number of angles in each range is accumulated over the whole motion vectors of macro blocks to form an angle histogram with \( n \) bins, denoted by \( AH(t,k) \).

In the PME definition, \( \alpha(t) \) is used to emphasize camera panning/tilting and it ideally reaches the task. But at the same time, it weakens another important global camera motion, camera zooming. As experiments shown, both frames with camera panning and with camera zooming are of great temporal importance. That is, without enough temporal rate support, the video will leave you great “motion judder” [4,8]. For this reason, we will modify the definition of \( \alpha(t) \) to emphasis all of these global camera motions, including camera panning and zooming.

Consider all the motion vectors of macro blocks of a frame with global camera motion, we can find either these motion vectors have a same motion direction (direction coherence: pan, tilt, etc.) or point to/ radiate from a centralized area (point coherence: zoom in, zoom out, etc.), as shown in Fig.4. For direction coherence, entropy can give a more effective emphasis than \( \alpha(t) \) does because it will not drop information of non-dominant motion direction:

\[
\beta(t) = 1 - \frac{\sum_{k=1}^{n} AH(t,k) \log AH(t,k)}{\log n}
\]

Both \( \alpha(t) \) and \( \beta(t) \) weaken point coherence, besides they cannot distinguish point coherence with random motions. So we managed to define \( \gamma(t) \) which indicates point coherence. Suppose a frame has \( \omega \) blocks, the motion vector of block \( B_{ij} \) (0≤\( i < u \), 0≤\( j < v \)) has an angle of \( \theta_{0} \) and the motion vectors of block \( B_{ij} \) and \( B_{ij'} \) intersect at point \( (x_{ij}, y_{ij}) \). Let \( P_{ij} = (x_{0}, y_{0}) \) be the center point of all intersection points with \( x_{0} \) and \( y_{0} \) be the average of \( x_{ij} \) and \( y_{ij} \).

\[
\gamma(t) = 1 - \frac{\sum_{k=1}^{n} AH(t,k) \log AH(t,k)}{\log n}
\]

Both \( \alpha(t) \) and \( \beta(t) \) weaken point coherence, besides they cannot distinguish point coherence with random motions. So we managed to define \( \gamma(t) \) which indicates point coherence. Suppose a frame has \( \omega \) blocks, the motion vector of block \( B_{ij} \) (0≤\( i < u \), 0≤\( j < v \)) has an angle of \( \theta_{0} \) and the motion vectors of block \( B_{ij} \) and \( B_{ij'} \) intersect at point \( (x_{ij}, y_{ij}) \). Let \( P_{ij} = (x_{0}, y_{0}) \) be the center point of all intersection points with \( x_{0} \) and \( y_{0} \) be the average of \( x_{ij} \) and \( y_{ij} \).
\[
\sigma(P_{\bar{ij} \bar{j}')} = \frac{\sum_{i,j,i',j'} | P_i P_{i'-j'} |}{uv(uv-1)}
\]

is the average distance of all intersection points to \( P_o \) with some simple modification:

- If \( \sigma(P_{ij}) > \sigma_0 \), where \( \sigma_0 \) is a constant for receivable mismatch, \( \sigma(P_{ij}) \) is set to 0.
- If two motion vectors are parallel or close to parallel, their intersection is set to a pixel between these two vectors with a considerable distance.
- To simplify the calculation, the block may not be a single macro block but a collection of adjacent macro blocks and the motion vector then be the summation.

Then \( \gamma(t) \) is defined as

\[
\gamma(t) = \frac{1}{1 + \sigma(P_{ij}^t)}
\]  \hspace{1cm} (2)

to emphasize the point coherence.

At last we modify the definition of PME to

\[
PME(t) = Mag(t) \cdot \max \{ \beta(t), \gamma(t) \} \]  \hspace{1cm} (3)

The definition reflects the below two perception rules: (1) More judder will be perceived if frames with fast motion are dropped; (2) The judder is most noticeable when the whole frame motions in a same direction or toward/from a same point, for example, global camera motions.

With definition (3), we apply model of [4] to define the priority rank of each frame in a video series.

### 3. DYNAMIC GOP STRUCTURE

We implement dynamic GOP in MPEG-4 FGST structure in which all B-frames are used to enhance temporal quality of I- and P-frames. Thus a high temporal ratio means more B-frames between adjacent I- or P-frames. For both code efficiency and user perception reasons, we’d like to allocate high temporal ratio to those important frame sections.

**Definition:** Suppose \( F_i \) is an I- or P-frame of a FGST-structured video, there are \( m \) B-frames between \( F_i \) and the next I- or P-frame. Then the **temporal ratio** of \( F_i \) is defined as:

\[
t(F_i) = \frac{1}{m}
\]

Fig.5 shows an example of priority rank of a video series. The leveled result ranks frame importance in which higher rank means more temporal importance. It is some different to that of key-frame which emphasizes a typical frame is prior to its neighbor frames. Fig.5 separates the video series into sub-sections. The sub-section with high priority rank is considered more important and will be given higher temporal ratio, which means higher sampling rate.

For convenience, frames with same priority ranks have a same temporal ratio and we will decide a weight function for temporal ratio of each priority rank. Currently, the weight function is a set of threshold determined by a large number of experiments.

Before detailed discussion, let’s introduce some related notations:

<table>
<thead>
<tr>
<th>MaxRank</th>
<th>Priority rank of frames is included in {0,1,2,...,MaxRank}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count[i]</td>
<td>Number of frames with priority rank ( i ) in the video</td>
</tr>
<tr>
<td>w[i]</td>
<td>Weight function for priority rank ( i )</td>
</tr>
<tr>
<td>( t[i] )</td>
<td>Temporal ratio of frames with priority rank ( i )</td>
</tr>
<tr>
<td>( t[F_i] )</td>
<td>Temporal ratio of frame ( F_i )</td>
</tr>
</tbody>
</table>

And the below three notations are defined in standard fixed GOP structure:

<table>
<thead>
<tr>
<th>SamplingRate</th>
<th>Temporal ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>PBetweenI</td>
<td>Number of P-frames between two adjacent I-frames</td>
</tr>
<tr>
<td>BBetweenP</td>
<td>Number of B-frames between two adjacent I- or P-frames</td>
</tr>
</tbody>
</table>

#### 3.1 B-frame in dynamic GOP structure

In deciding dynamic GOP structure we still want the encoded video has an average sampling rate of SamplingRate as for fixed GOP structure, which means

\[
\sum_{i=0}^{MaxRank} Count[i] \cdot t[i] = SamplingRate \cdot \sum_{i=0}^{MaxRank} Count[i]
\]

And according to the definition of weight function \( w[i] \),

\[
t[i] = w[i] \cdot w[j] \quad \text{for} \quad 0 \leq i, j \leq MaxRank
\]

Simplify above two equations, we get the temporal ratio for each priority rank \( i \)

\[
t[i] = \frac{\text{SamplingRate} \cdot \sum_{i=0}^{MaxRank} Count[i] \cdot w[i]}{\sum_{i=0}^{MaxRank} Count[i] \cdot w[i]} \]  \hspace{1cm} (4)

In a FGST structured digital video, frames can be identified into two types: Base-layer frames (include all I-frames and P-frames) and FGST-layer frames (all B-frames). Thus for each frame in the video, we need to decide if it is encoded as a base-layer frame or FGST-layer frame. Since each frame has been decided a priority rank, with the help of (4) we can get the temporal ratio of each frame of the video. The following algorithm tries to distinguish base-layer frames and FGST-layer frames:

**Algorithm B-Frames_Identify()**

```c
double dFlag=1.0;
for (int i=0; i<FrameNumber; i++){
    if (dFlag>=1.0){
        // do something
    }
```
In dynamic GOP structure, we have more freedom to decide whether a frame is encoded as an I-frame, P-frame or B-frame. Other optimization methods can be adopted in I-, P- and B-frame decision. For example, let I-frame be the first frame of a certain shot [5], etc.

Through section 3.1 and 3.2, we define the dynamic GOP structure of a video.

3. EVALUATIONS

In dynamic GOP structure, we will use \( len = (PBetweenI + 1) \cdot (BBetweenP + 1) \) as a threshold. (1) The first base-layer frame is I-frame; (2) If the number of frames between a base-layer frame and the nearest preceding I-frame is larger than or equal to \( len \), this base-layer frame is then encoded as an I-frame. Otherwise it is encoded as a P-frame.

In a dynamic GOP structure, more base-layer frames are chosen from Part III of the video sequence which is of more temporal importance and thus the judders are weakened and the encoded video gains better perception.

As a byproduct, the average SNR value has an increase ranged from 0.1-0.2 through adopting dynamic GOP to a great number of experimental videos compared to adopting standard fixed GOP structure.

4. CONCLUSION REMARKS

A measure to estimate the temporal importance of frame in a video sequence, PME, is detailed discussed in this paper. With the help of this metric, we propose dynamic GOP structure in digital video encoding. It provides better user perception in bandwidth decreasing, better code efficiency and this make the digital video contains temporal importance information itself which will be benefited in temporal-SNR quality tradeoff. More over, dynamic GOP structure maintains all strong points of standard fixed GOP structure.

Since the fixed structure broken, a great deal of beneficial operation can be adopted in dynamic GOP structure such as synchronizing shot and GOP by making inter-coded I-frames the first frame of a shot.

5. REFERENCES