A REGION BASED MULTIPLE FRAME-RATE TRADEOFF OF VIDEO STREAMING

Wei Lai[+ *], *Xiao-Dong Gu*[*], *Ren-Hua Wang*[+], *Li-Rong Dai*[+], *Hong-Jiang Zhang*[*]

+ Department of Electronics Engineering & Information Science,
University of Science and Technology of China,
Hefei, Anhui, 230026, China
laiwei@ustc.edu, {rhw,lrdai}@ustc.edu.cn

* Microsoft Research Asia,
5F, Sigma Building, Zhichun Road,
Beijing, 100080, China
{t-xiaogu,hjzhang}@microsoft.com

ABSTRACT

Conventional frame rate up-conversion adopted in low bit-rate video coding/streaming cannot obtain good quality due to artifacts caused by composition problems. In this paper, a region based multiple frame-rate tradeoff scheme is proposed to reduce artifacts. The video frames are divided into regions of interest (ROIs) and backgrounds by the visual attention model. The motion intensity and complexity of ROI and background are computed by perceptive motion energy spectrum (PMES) to decide the ratio of the frame rate of ROI to that of background. The background is encoded at a low frame rate, and up-converted to the same frame rate as ROI at the decoder, with an integrated background motion composition model. The experimental results indicate that, compared to conventional frame rate up-conversion and normal codec scheme without frame rate up-conversion at the same bandwidth, the proposed scheme has higher PNSR performance and better subjective visual quality.

Key Words: Content-based Video Streaming, Visual Attention Model, Frame rate Up-conversion.

1. INTRODUCTION

In low bit-rate video coding/streaming, the videos may be encoded at a very low frame rate, or skip some frames at the encoder/sender, and recover these frames at the decoder/receiver by frame rate up-conversion. Frame rate up-conversion will cause artifacts by diverse composition problems [1]. Many interpolation algorithms have been proposed to reduce the artifacts. *Object based motion compensation interpolation (OB-MCI)*, *overlapped block motion compensation (OBMC)* and *multiple motion estimation* are proposed in [2], [3] and [4] respectively, to reduce the artifacts of object boundaries.

In this paper, we propose a coding scheme to solve the problem by another approach. The video frames are segmented into Regions of Interest (ROIs) and non-ROI regions (background), by the visual attention model. The ROIs are encoded at a high frame rate, while the background is encoded at low frame rate. Frame rate up-conversion is performed at the background at the decoder.

The ROIs are regions with high attention level provided by the visual attention model. The regions of complex motion and object boundary, which may cause composition problems, are included by ROIs. Encoding ROIs at a high frame rate can ensure smooth motion and accurate object boundary, which can give a good visual quality. The background is paid less attention, and contains no complex motion and evident edges, which tends to have a global motion. The frame rate of background can be decided by its motion mode and motion intensity, which is computed by perceptive motion energy spectrum (PMES) [6].

At the decoder/receiver, the frame rate of background is up-converted to the same frame rate of ROIs by an integrated background motion composition model developed in this paper, the integrated model can make ROIs and background have synchronous and smooth motion, and introduce least artifacts.

The encoded video with multiple frame rate scheme mentioned above (defined as the M-FR version), is compared to a video encoded at high frame rate (defined as the H-FR version), and a video encoded at low frame rate, and up-converted to high frame rate (defined as the L-FR version). These three versions of encoded video streaming have an approximate same bit-rate (bandwidth). Our experimental results intricate that the M-FR version gets a higher average PSNR than the H-FR version, and gets a higher PSNR at interpolated frames than the L-FR version. The evaluation results also show that our M-FR version has a better subjective visual quality that outperforms the H-FR version and L-FR version.

2. VISUAL ATTENTION MODEL AND ROI EXTRACTION

Attention is a neurobiological conception. It implies the concentration of mental powers upon an object by close or careful observing or listening, which is the ability or power to concentrate mentally [5].

In an image sequence, there are many visual features, including motion, color, texture, shape, text region, etc. Also, some recognizable objects, such as face, will more likely attract human attention. For static image, a saliency map is generated by fusing the three channel saliency maps computation: color contrasts, intensity contrasts, and orientation contrasts.

The motion attention model as dynamic, is built based on motion vector field (MVF). It is assumed that a MVF has three inductors: Intensity inductor, Spatial Coherence inductor, and Temporal Coherence inductor. The perceived motion energy spectrum (PMES) is proposed in [6] to extract the object motion from the global background motion effectively. The PMES is also used to decide the frame rate conversion ratio of the non-ROI region, which will be discussed in the next section.

With the visual attention model, a saliency map is computed from each frame of a video shot. The regions with high attention values in the corresponding saliency map are detected as ROIs. Since video is an image sequence having continuous motions, the ROIs of the frame sequence are performed a smoothing and filtering process, to provide good spatial and temporal consistency.

Generally, the objects with high and complex motion, complex texture, and clear edges are included in the ROIs, the remain non-ROI region is considered as background, which has only simple and global motion, and will be paid less attention by viewers. As an example, the detected ROI is shown in the rectangle in the right of figure 1.

3. CHOOSE FRAME RATE CONVERSION RATIO

It is mentioned that frame rate up-conversion will cause artifacts, especially when there are composition problems (complex motions, overlapping objects) [1]. So, in our multiple frame-rate scheme, the ROIs, which contain complex motions or overlapping boundaries, are encoded in full frame rate. Without frame rate up-conversion, no artifacts will be introduced into ROIs. These regions which catch more attention can maintain a high visual quality.

The background has no such factors as ROIs have that will cause artifacts. The slighter the motion is, and the more uniform the motion is, the better the frame rate up-conversion performance will be. The perceived motion energy spectrum (PMES) is used as a metric of motion strength and coherence, and then it is used to decide the frame rate up-conversion ratio of background.

The PMES at the macro block ($MB_{i,j}$) is computed as following [6]:

$$PMES_{i,j} = GMR_{i,j} \times \overline{MixEn_{i,j}} \tag{1}$$

The $MixEn_{i,j}$ reflects the strength of the motion, while $GMR_{i,j}$ stands for the entropy (consistency) of motion angles.

To decide the frame rate of background, the average PMES is computed as:

$$\overline{PMES}_{BG} = \sum_{(i,j)\in BG} PMES_{i,j} \Big/ \|BG\|$$
$$\overline{PMES}_{ROI} = \sum_{(i,j)\in R} PMES_{i,j} \Big/ \|R\| \tag{2}$$

BG denotes the background; R denotes all of the ROIs. ‖Region‖ is the number of MBs in the region.

The frame rate of ROIs is selected as full frame rate, denoted by f_{ROI}.

The frame rate of non-ROI region (background) f_{BG} is computed as following:

$$f_{BG} = k \cdot f_{ROI} \cdot \frac{\overline{PMES}_{BG}}{\overline{PMES}_{ROI}} \tag{3}$$

where k is an adjustment parameter.

4. BACKGROUND COMPOSITION MODEL

4.1. Estimated motion vectors

The estimated motion vector filed is defined as $\overrightarrow{MV_E}(x, y)$. It can be provided by the MPEG encoder. But unfortunately, the estimated motion vectors (MVs) cannot represent the real MVs. Some MVs may vary from the real MVs greatly. If use $\overrightarrow{MV_E}(x, y)$ to interpolate the background directly, it will cause many artifacts.

4.2. Boundary constraint continuous MVF model

To make a continuous distribution at ROI boundaries, a motion vector field $\overrightarrow{MV_B}(x, y)$ is constructed to satisfy the following conditions:

$$\overrightarrow{MV_B}(x_b, y_b) = \overrightarrow{MV_E}(x_R, y_R)$$
$$\left|\overrightarrow{MV_B}(x, y) - \overrightarrow{MV_B}(x', y')\right| < threshold \tag{4}$$

where (x_b, y_b) is a position in the background neighbor to the ROI boundary, (x_R, y_R) is the corresponding nearest position to (x_b, y_b) in ROI and at the boundary. (x', y') is the position neighbor to (x, y) in the background.

The MVF is constructed as the following steps:

(1) Clear the MVs of all positions in the background.

(2) For every positions (x_b, y_b) neighbor to the ROI boundary, $\overrightarrow{MV_B}(x_b, y_b) = \overrightarrow{MV_E}(x_R, y_R)$

(3) The positions in step 1 are considered as the new boundary, for every positions (x, y) neighbor to and outside the new boundary:

$$\overrightarrow{MV_B}(x,y) = \sum \overrightarrow{MV_B}(x_n, y_n) / N \quad (5)$$

where (x_n, y_n) are the N positions neighbor to (x,y) and at the new boundary.

(4) The positions in step 3 are considered as the new boundary and go back to step 3, until all the positions in the background have a motion vector.

4.3. Integrated background motion composition model

The integrated motion vector field is a weighted combination of the MVFs above:

$$\overrightarrow{MV_I}(x,y) = w_b(x,y)\overrightarrow{MV_B}(x,y) + (1 - w_b(x,y))\overrightarrow{MV_E}(x,y)$$
$$w_b(x,y) = e^{-\lambda D_b(x,y)} \quad (6)$$

where $D_b(x,y)$ is the distance from (x,y) to ROI boundary. λ is a parameter.

The backgrounds at missed frames are composed by motion compensation interpolation (MCI) [3] using $\overrightarrow{MV_I}(x,y)$.

Since the integrated background motion filed has good consistency, especially at the ROI boundary. The interpolated frames get few artifacts and can have a good visual quality.

5. EXPERIMENTAL RESULTS

The standard test sequences "*stefan*" and "*akyio*", which stand for sequence with big motion and light motion respectively, were chosen to compute the ratio of $PMES_{ROI}$ to $PMES_{BG}$. Only the first 150 frames of these two sequences are used. The results are in table 1.

Sequence	$PMES_{ROI}$	$PMES_{BG}$	Ratio
stefan	0.707	0.315	2.244
akyio	0.375	0.005	71.094

Table 1. PMES of ROI/background and the ratio

The results are reasonable. The ROI of "*stefan*" (the player) contains big and complex motions, so $PMES_{ROI}$ of "*stefan*" gets a high value. The background of "*stefan*" (the playground and audience) contains a big motion too, but the motion is a global pan motion, so $PMES_{BG}$ of "*stefan*" gets a lower value than $PMES_{ROI}$.

The PMES map of the 38th frame of the experimental sequence is shown in figure 1 (the rectangle is ROI).

As for "*akyio*", obviously $PMES_{BG}$ is very small, since the background (the region excluding the news reader) is almost still.

To show the performance of the proposed multiple frame-rate scheme, we use the 75 even frames of the first 150 frames of "*stefan*" to compose a 5 seconds, 15 fps sequence. 3 encode-decoded versions of it are generated:

(1) High frame rate (H-FR) version: the sequence is encoded and decoded at 15 fps.

(2) Low frame rate (L-FR) version: the sequence is encoded and decoded at 7.5 fps, and up convert to 15 fps.

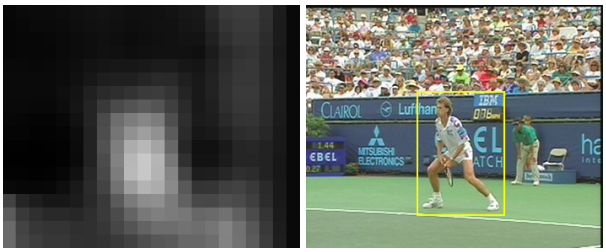

Figure 1. Left: PMES map, Right: original frame.

(3) Multiple frame-rate (M-FR) version: by the results of PMES ratio above, the frame rate conversion ratio is set to 3. The ROI of the sequence is encoded and decoded at 15fps, while the background of the sequence is encoded and decoded at 5 fps, and then up converted to 15 fps, by the integrated background composition model proposed in section 4.3.

The encoder and decoder are based on MPEG-4 FGS codec. The sizes of the bitstream of these three versions are almost equal to each other. That is to say, the comparison is under the same bandwidth.

The PSNRs of the first 30 frames of these three versions are shown in figure 2.

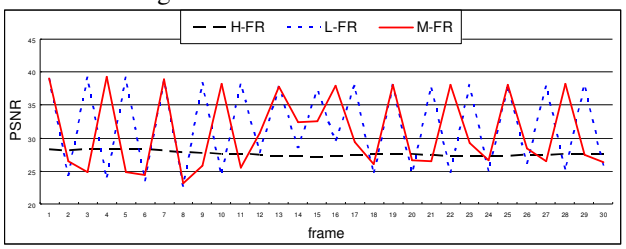

Figure 2. PNSR of the first 30 frames

The average PNSRs of these three versions are:

H-FR: 27.596 db
L-FR: 31.618 db
M-FR: 30.581 db

We can see that the average PNSR of M-FR is close to L-FR, but much higher than that of H-FR.

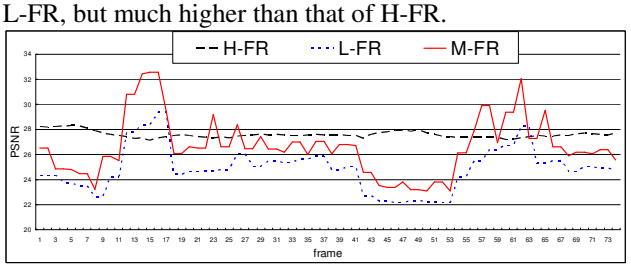

Figure 3. PSNR of interpolated frames

The PNSRs of interpolated frames of the M-FR version (frame $3*n+1$, $3*n+2$; n=0,1,…,24) and the L-FR version (frame $2*n+1$; n=0,1,…,37) are shown in figure 3, the PSNR of all frames of the H-FR version is also given out as a reference.

We can see that the PNSR at the interpolated frames of M-FR is higher than that of L-FR. It means that the frame rate up-conversion of background of M-FR gets more

precise result (less artifacts) than the frame rate up-conversion of the whole frame of L-FR.

The 23rd decoded frames of the three versions are shown in figure 4. This frame is an interpolated frame in both the L-FR version and M-FR version.

Figure 4. Top: H-FR, Middle: L-FR, Bottom: M-FR.

We can see that H-FR gets a poor quality of each frame. The even frames of L-FR have good qualities, but the odd (interpolated) frames get many artifacts due to the error motion compensation interpolation (the head and leg) of the region with complex motion (the player). The L-FR version gets a bad quality when it is played continuously.

The M-FR, which was generated by the proposed scheme, has no defects of the 2 versions above. It can provide the best visual quality.

6. SUMMARY

In this paper, a multiple frame rate tradeoff scheme of video coding/streaming is proposed to reduce the artifacts in conventional frame rate up-conversion. The video frames are divided into ROIs and background by the visual attention model. The ROI is encoded at full frame rate, while the background is encoded at a low frame rate and up-converted to the same frame rate as ROI at decoder with an integrated background motion composition model proposed in this paper. The frame rate up-conversion ratio is decided by the motion intensity and complexity, which is computed by PMES. Compared to the conventional frame rate up-conversion (L-FR version) and normal coding at full frame rate (H-FR version), the proposed scheme (M-FR version) has higher average PSNR than H-FR version, and has higher PSNR than L-FR version at the interpolated frames. The experimental results also show that the M-FR version can get the best subjective visual quality over the three versions.

7. REFERENCES

[1] Jeong-Woo Lee, Anthony Vetro, Yao Wang and Yo-Sung Ho, "Bit Allocation for MPEG-4 Video Coding with Spatio-Temporal Trade-offs", *IEEE Trans. on CSVT*, Vol. 13, pp. 488-502, 2003.6.

[2] Sung-Hee Lee, Ohjae Kwon, and Rae-Hong Park, "Weighted-Adaptive Motion-Compensated Frame Rate Up-Conversion", *IEEE Trans. on Consumer Electronics*, Vol. 49, 2003.8.

[3] Soo-Chul Han and John W. Woods, "Frame-rate Up-conversion Using Transmitted Motion and Segmentation Fields for Very Low Bit-rate Video Coding", *IEEE International Conference on Image Processing*, Vol. 1, pp. 747-750, 1997.10.

[4] Kunio Kawaguchi and Sanjit K. Mitra, "Frame Rate Up-Conversion Considering Multiple Motion", *IEEE Conference on Image Processing*, Vol. 1 pp. 727-730, 1997.10.

[5] Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang and Mingjing Li, "A User Attention Model for Video Summarization", *ACM Multimedia'02*, 2002.12.

[6] Yu-Fei Ma and Hong-Jiang Zhang, "A New Perceived Motion Based Shot Content Representation", *IEEE International Conference on Image Processing (ICIP 2001)*, 2001.10.