A DYNAMIC FRAME-SKIPPING VIDEO COMBINER FOR MULTIPOINT VIDEO CONFERENCING

Kai-Tat Fung, Wan-Chi Siu and Yui-Lam Chan

Centre for Multimedia Signal Processing
Department of Electronic and Information Engineering
The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

ABSTRACT

This paper proposes a new architecture of video combiner for multipoint video conferencing. The proposed video combiner first identifies the active and talking conference participants from the multiple incoming video bitstreams by computing their corresponding motion activities and level of audio signals. By reducing the frame rates of inactive sub-sequences, more bits are reallocated to the active sub-sequences such that higher frame rates are obtained to produce a better visual perception as well as a smoother motion. We also propose a low-complexity and high-performance frame-skipping transcoder which is achieved by a direct summation of the DCT coefficients for non-moving macroblocks and an adaptive motion vector refinement for motion compensated macroblocks. The new transcoder is then used to realize the multipoint video conferencing and some results are presented to show the improvement in performance due to our proposed architecture.

1. Introduction

For multipoint video conferencing over a wide-area network, the conferencing participants are connected to a multipoint control unit (MCU) [1-2] which coordinates and distributes audio, video and data streams among multiple participants according to the requirement of the channel bandwidth. An audio mixer in the MCU accepts audio data in a variety of formats, with different data rates. The audio mixer must decode and mix these different audio bitstreams from all the conference participants and the mixed audio signal is encoded again for distribution to the conference participants. Similarly, a video combiner is also included in the MCU to combine the multiple coded video bitstreams from the conference participants into a coded video bitstream which conforms to the video coding standard such as H.263[3], and send it back to the conference participants for decoding and presentation.

Recently, multipoint video conferencing using the transcoding approach[4-5] has become a popular and an important means of communication. However, the computational complexity is inevitably increased since the individual video bitstream needs to be decoded and the combined video signal needs to be encoded. In addition, the video quality of the transcoding approach suffers from its intrinsic double-encoding process which introduces additional degradation. In order to provide a satisfactory visual quality of combined and transcoded video, the re-distribution of the limited bits in the reencoding process to different parts of the combined video is a critical step.

In this paper, we propose a new architecture for a video combiner which allows the sub-sequences to be transcoded in different frame rates according to their motion activities and audio levels. The proposed video combiner has two new main features:

1. a dynamic sub-frame skipping (DSFS) scheme for a dynamic selection of the most representative sub-frames according to the motion activities and the audio levels of the conference participants;
2. a high-quality and low-complexity frame-skipping transcoder which provides a satisfactory visual quality and reduces the computational burden of the MCU by employing a direct summation of the DCT coefficients and fast motion re-estimation in the transcoder.

2. The Proposed Video Combiner

Figure 1 shows the proposed system architecture for a video combiner in multipoint video conferencing. For simplicity and without loss of generality, four QCIF video bitstreams are received by the video combiner from the conference participants. Each QCIF bitstream is first parsed with a variable-length decoder (VLD) to extract the header information, coding mode, motion vector and quantized DCT coefficients for each macroblock. The frame-skipping transcoder processes and updates the quantized DCT coefficients of the current frame in the DCT-domain buffer for each QCIF sub-sequence. The transcoded and quantized DCT coefficients in the DCT-domain buffers in QCIF format are subsequently combined into a single buffer in CIF format through a multiplexer. At the beginning of the formation of a combining sequence, a decision is made as to which DCT-domain buffers (QCIF format) should be included in the new buffer (CIF format) by the dynamic sub-frame skipping (DSFS) controller.

A dynamic sub-frame skipping (DSFS) is proposed which can dynamically distribute the encoded sub-frames to each sub-sequence by considering the motion activities and audio levels. Thus, it is necessary to regulate the frame rate of each sub-sequence according to the motion activity in the current frame (MA_t) of the sub-sequence and its corresponding audio level. To obtain a quantitative measure for MA_t, we use the accumulated magnitudes of all of the motion vectors estimated for the macroblocks in the current frame, i.e.,

$$MA_t = \sum_{i=1}^{M} \left|(u_t^n)_i\right| + \left|(v_t^n)_i\right| \qquad (1)$$

where M is the total number of macroblocks in the current sub-frame, and $(u_t^n)_i$ and $(v_t^n)_i$ are the horizontal and vertical components of the motion vector of the ith macroblock which uses the previous non-skipped sub-frame as a reference.

If the value of MA_t after a non-skipped sub-frame exceeds the predefined threshold, T_{MA}, the incoming sub-frame should be kept. By adaptively adjusting the frame rate of each sub-sequence according to the MA_t, the proposed architecture can allocate more

0-7803-7448-7/02/$17.00 ©2002 IEEE

sub-frames for a sub-sequence with high motion activity and fewer sub-frames for a sub-sequence with low motion activity.

It is interesting to note that the different sub-sequences have different values of T_{MA} which are set according to the audio level of the sub-sequence. The larger the T_{MA} is set in the sub-sequence, the more the sub-frames will be skipped. Consequently, more saved bits will be used in other active and talking sub-sequences. The threshold setting process consists of testing the linear samples of the audio signal of a sub-sequence to detect whether this channel is active or silent. In the threshold setting process, the active talker can be identified and a lower T_{MA} will be used. The DSFS scheme is summarized in Figure 2.

If the current sub-frame of a sub-sequence is selected, it is picked out by the DCT-domain multiplexer and the DCT coefficients in the corresponding DCT-domain buffer are copied in the video combining processor's buffer with the size of CIF. The quantized DCT coefficients of each conference participant only need to be mapped according to Figure 1. Hence, the DCT-domain buffer of conference participant 1, *CP1*, is mapped to the first quadrant of the combined picture, the DCT-domain buffer of conference participant 2, *CP2*, is mapped to the second quadrant, etc. The data in the DCT-domain buffer will be encoded in the compressed bitstream by the stream processor. If the current sub-frame of a sub-sequence is skipped, it simply sets all the COD (coded macroblock indication) bits in the H263 syntax[3] of the macroblocks which belong to the skipped sub-frames to "1" to avoid sending the associated DCT coefficients, motion vectors, and macroblock overhead bits. Only 99 COD bits are required to represent a skipped QCIF sub-frame, thus the overhead is relatively negligible.

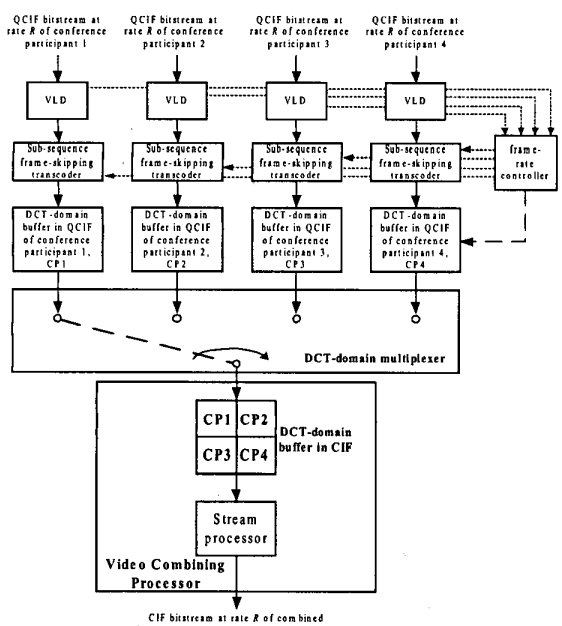


Figure 1. Architecture of the proposed video combiner.

T_{MA} setting process

if (energy level of sub-sequence's audio < T_{audio})
then
$T_{MA} = T_l$;
else
$T_{MA} = 0.5 \times T_l$;
where T_{audio} and T_l are the predefined threshold.

Dynamic sub-frame skipping

if ($MA_t < T_{MA}$)
then
skip the current sub-frame;
else
keep the current sub-frame;

Figure 2. Pseudocode of the DSFS scheme.

3. The High Performance Frame-Skipping Transcoder

According to the skipping decision provided by the DSFS controller, the frame-skipping transcoders in Figure 1 must perform sub-frame skipping on different sub-sequences. Each transcoder also has the responsibility for updating its corresponding DCT-domain buffer (QCIF format) in the proper manner such that these buffers can be multiplexed by the video combining processor as mentioned in section 2. However, the skipped sub-frame must be decompressed completely, and should act as the reference sub-frame to the non-skipped sub-frame for reconstruction. The newly quantized DCT coefficients of prediction error and the motion vectors need to be re-computed for the non-skipped sub-frame with reference to the previous non-skipped sub-frame, which introduces a re-encoding process; this can create an undesirable complexity in real time application as well as introduce re-encoding error. In this section, we propose a new sub-sequence frame-skipping transcoder for improving picture quality and reducing complexity. In [4], we have proposed a direct summation of the DCT coefficients in the frame-skipping transcoding to avoid the re-encoding errors and to reduce the complexity for macroblocks coded without motion compensation. In this paper, we present a new frame-skipping transcoder which is an extension of the work of [4]. The new transcoder has the following features:

1. a direct summation of the DCT coefficients for re-computing the newly quantized DCT coefficients of prediction error in the macroblocks without motion compensation;
2. an adaptive motion vectors composition for the new motion vectors.

3.1 Macroblocks without motion compensation

In Figure5, a situation in which one frame is dropped is illustrated. We assume that MB_t represents the current macroblock and MB_{t-1} represents the best matching macroblock to MB_t. Since MB_t is coded without motion compensation, the spatial position of MB_{t-1} is the same as that of MB_t, and MB_{t-2} represents the best matching macroblock to MB_{t-1}. Since R_{t-1} is dropped, for MB_t, we need to compute a motion vector, (u_t^s, v_t^s), and the prediction error in quantized DCT-domain, $Q[DCT(e_t^s)]$, by using R_{t-2} as a reference. Since the motion vector in MB_t is zero, it simply re-computes such a motion vector as, $(u_t^s, v_t^s) = (u_t, v_t)$. Since re-

encoding can lead to an additional error, it could be avoided if $Q[DCT(e_t^s)]$ can be computed in the DCT-domain. For the sake of simplicity, we define the incoming residual signal with the quantization error to the transcoder as $\hat{e}_t = e_t + \Delta_t$. From Figure5, the pixels of MB_t can be reconstructed by performing inverse quantization and inverse DCT of $Q[DCT(\hat{e}_t)]$ and summing this residual signal to pixels in MB_{t-1} which can be similarly reconstructed by performing inverse quantization and inverse DCT of $Q[DCT(\hat{e}_{t-1})]$ and summing this residual signal to pixels in the corresponding MB_{t-2}. However, by considering the linearity of inverse DCT and inverse quantization, we obtain,

$$Q[DCT(e_t^s)] = Q[DCT(\hat{e}_t)] + Q[DCT(\hat{e}_{t-1})] \quad (2)$$

Equation (2) implies that the newly quantized DCT coefficient $Q[DCT(e_t^s)]$ can be computed in the DCT-domain by summing directly the quantized DCT coefficients between the data in the DCT-domain buffer and the incoming quantized DCT coefficients, whilst the updated DCT coefficients are stored in the DCT-domain buffer. Since it is not necessary to perform motion compensation, DCT, quantization, inverse DCT and inverse quantization, the complexity is reduced. Furthermore, since re-quantization is not necessary for this type of macroblock, the quality degradation of the transcoder introduced by $\Delta_t^s(i,j)$ is also avoided.

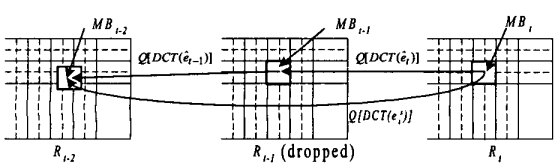


Figure 3. Residual signal re-computation of frame-skipping for macroblocks without motion compensation.

3.2 Motion-compensated macroblocks

For motion-compensated macroblocks, the motion vectors in the dropped sub-frames are usually not available in the incoming bitstreams, as depicted in Figure 4, thus new motion vectors for the transcoded bitstream need to be re-computed. Motion vector re-estimation is undesirable due to the intensive computation. Instead, motion vector re-estimation using the available motion vectors is a better approach [6-7]. It is possible to use bilinear interpolation from the motion vectors mv_{t-1,MB_1}, mv_{t-1,MB_2}, mv_{t-1,MB_3} and mv_{t-1,MB_4} which are the four neighboring macroblocks, MB_{t-1}^1, MB_{t-1}^2, MB_{t-1}^3 and MB_{t-1}^4, of MB_{t-1} to come up with an approximation of mv_{t-1} [8]. However, bilinear interpolation of motion vectors leads to inaccuracy of the resultant motion vector because the area covered by the four macroblocks may be too divergent and too large to be described by a single motion vector [6-7]. Thus, the dominant vector selection approach is used [6-7] to select one dominant motion vector from four neighboring macroblocks. A dominant motion vector is defined as the motion vector carried by a dominant macroblock. The dominant macroblock is the macroblock that has the largest overlapped segment with MB_{t-1}. For example, the new motion vectors in Figure 4 can be composed by

$$mv_t^n = mv_t + mv_{t,MB_1} \quad (3)$$

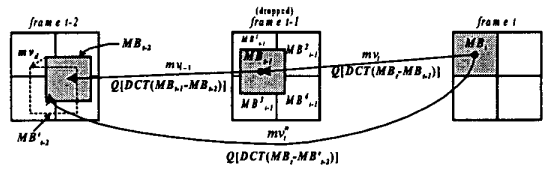


Figure 4. Motion-compensated macroblocks.

In [7], the motion vector refinement scheme uses the motion vector computed in equation (3) as the base motion vector and then performs a motion estimation in a very small search range around the base motion vector. The performance of this scheme is further improved in this paper. If there is a strong dominant macroblock which overlaps the reference macroblock in a significantly large area, this motion vector is reliable for composing the new motion vector. In other words, refinement of the motion vector is not required. The adaptive motion vector refinement scheme is summarized as follows:

1. The base motion vector, mv_t^n, is composed according to equation (3).
2. If the largest overlapping area is greater than a T_{oa}, which is a predefined threshold (e.g. say 80% of the macroblock area), then select the base motion vector as the final motion vector, otherwise go to step (3).
3. The delta motion vector, mv_d, is estimated within a small search area (±3 pixels) around the base motion vector (motion vector refinement scheme), as depicted in Figure 4, and the new motion vector is computed by

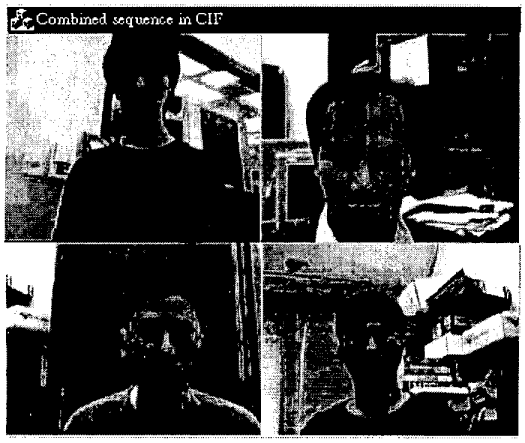
$$mv_t^n = mv_t + mv_{t,MB_1} + mv_d \quad (4)$$

After obtaining the new motion vector, the newly quantized DCT coefficients, $Q[DCT(MB_t - MB'_{t-2})]$, are required to be computed. Again, since MB_{t-1} is not on a macroblock boundary, direct summation cannot be employed. In other words, re-encoding of the motion-compensated macroblocks is inevitable. Another advantage of the proposed transcoder is that when multiple frames are dropped, it can be processed in the forward order, thus eliminating the multiple DCT-domain buffers that are needed to store the incoming quantized DCT coefficients of all dropped frames.

4. Experimental Results

A series of experiments was conducted to evaluate the overall efficiency of the proposed video combiner for multipoint video conferencing. For our experiments, we recorded a four-point video conferencing session. Each conferee's sequence was encoded into a QCIF format at 128kb/s according to the H.263 standard [3]. At the front encoder, the first frame was coded as intraframe (I-frame), and the remaining frames were encoded as interframes (P-frames). These picture-coding modes were preserved during the transcoding. The four sequences were then transcoded and combined using the CIF format. The PSNR of a skipped sub-frame is calculated from the incoming sub-frames and

the latest previously non-skipped sub-frame, since the sub-frame repetitions will occur for the skipped sub-frames at the video decoders. It can be seen that our proposed video combiner offers a significant gain for both active and non-active periods as shown in Table 1 and Table 2, which represents the average PSNR's of the sub-sequence of all frames and non-skipped frames, respectively. Also, direct summation and adaptive motion refinement in the proposed transcoder can reduce the computational burden of re-encoding, which gives rise to a significant speed-up of about 9.2 times as compared with Sun *et al's* video combiner. The video quality of the active participant (at the upper right corner) produced by the proposed video combiner is better than that of the Sun *et al's* video combiner as shown in Figure 5.



(a)

(b)

Figure 5. Frame 184 of the combined video sequence using (a) Sun *et al.*'s video combiner [5] (b) our video combiner. The active conference participant is at the upper right corner.

Table 1. Average PSNR's of the sub-sequence of all frames.

	Frame 1 – 100 (most active: CP1)		Frame 101 – 200 (most active: CP2)		Frame 201 – 300 (most active: CP3)		Frame 301 – 400 (most active: CP4)	
	A	B	A	B	A	B	A	B
CP1	**32.62**	**35.58**	34.68	36.54	34.48	36.57	34.39	36.54
CP2	33.95	36.57	**32.80**	**35.89**	34.45	36.18	34.38	36.32
CP3	33.81	36.32	33.99	36.49	**31.95**	**35.13**	34.32	36.38
CP4	33.86	36.42	33.91	36.69	33.77	36.66	**32.21**	**35.68**

A – Sun *et al.*'s video combiner [5]. B – Proposed video combiner.

Table 2. Average PSNR's of the sub-sequence of non-skipped frames.

	Frame 1 – 100 (most active: CP1)		Frame 101 – 200 (most active: CP2)		Frame 201 – 300 (most active: CP3)		Frame 301 – 400 (most active: CP4)	
	A	B	A	B	A	B	A	B
CP1	**32.62**	**35.77**	34.68	36.66	34.48	36.72	34.39	36.68
CP2	33.95	36.71	**32.80**	**36.06**	34.45	36.30	34.38	36.48
CP3	33.81	36.48	33.99	36.65	**31.95**	**35.34**	34.32	36.55
CP4	33.86	36.59	33.91	36.83	33.77	36.81	**32.21**	**35.89**

A – Sun *et al.*'s video combiner [5]. B – Proposed video combiner.

5. Conclusions

In this paper, we have implemented a new video combiner for multipoint video conferencing by using the dynamic sub-frame skipping (DSFS) scheme and the high-performance frame-skipping transcoder. We also propose a frame-skipping transcoder for minimizing re-encoding error and reducing computational burden by employing the direct summation of the DCT coefficients and the adaptive motion refinement. This transcoder can be processed in the forward order when multiple frames are dropped. Thus, only one DCT-domain buffer is needed to store the updated DCT coefficients of all dropped frames. The proposed video combiner can provide better performance than the conventional video combiner in terms of better quality and reduced complexity.

6. Acknowledgments

This work is supported by the Centre for Multimedia Signal Processing, Hong Kong Polytechnic University. K.T. Fung acknowledges the research studentships provided by the University.

7. References

[1] ITU-T Study Group XV, Recommendation H.231, "Multipoint control units for audiovisual systems using digital channels up to 2 Mb/s," May 1992.

[2] ITU-T Study Group XV, Recommendation H.243, "Procedures for establishing communication between three or more audiovisual terminals using digital channels up to 2 Mb/s," May 1992.

[3] ITU-T Study Group XV, Recommendation H.263, "video coding for low bitrate communication," May 1997.

[4] Kai-Tat Fung, Yui-Lam Chan and Wan-Chi Siu, "Low-Complexity and High Quality Frame-Skipping Transcoder," in Proc. IEEE Int. Symposium on Circuits and Systems'2001, Sydney, Australia, pp. 29-32, May 6-9, 2001.

[5] Ming-Ting Sun, Tzong-Der Wu, and Jenq-Neng Hwang, "Dynamic bit allocation in video combining for multipoint conferencing," IEEE Trans. on Circuits and Systems – II: Analog and Digital Signal Processing, vol. 45, no. 5, May 1998.

[6] Jeongnam Youn, Ming-Ting Sun and Chia-Wen Lin, "Motion vector refinement for high-performance transcoding," IEEE Transactions on Multimedia, vol. 1, pp. 30-40, March 1999.

[7] Jeongnam Youn, Ming-Ting Sun and Chia-Wen Lin, "Motion estimation for high performance transcoding," IEEE Transactions on Consumer Electronics, vol. 44, pp. 649-658, August 1998.

[8] Jenq-Neng Hwang, Tzong-Der Wu and Chia-Wen Lin, "Dynamic frame-skipping in video transcoding," 1998 IEEE Second Workshop on Multimedia Signal Processing, pp. 616 –621, 1998.