A Dynamic Video Combiner for Multipoint Video Conferencing using Wavelet Transform

Kai-Tat Fung, Wan-Chi Siu and Ngai-Fong Law

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

ABSTRACT
A new architecture of video combiner for multipoint video conferencing is proposed in this paper. The proposed video combiner is wavelet-based which extracts the motion activities information from the video bitstreams produced by a wavelet-based video coder. Using the progressive properties of wavelet transform, the encoded bitstream become scalable. Hence, the video quality of inactive sub-sequences can be easily adjusted in the video combiner by discarding the fine detail information bitstreams. In other words, more bits can be reallocated to the active sub-sequences to achieve a good visual quality with smooth motion. In addition, the video coder is region-based so that different wavelet kernels can be used for the foreground and the background. This setting can on one hand reduce the computational complexity significantly. On the other hand, by considering the unequal importance of various regions, a high video quality in foreground can always be guaranteed and an acceptable quality in background can be maintained even under low bit rate environments. Since the video combiner only needs to rearrange the video quality level according to their motion activities, no re-encoding process is required. Therefore, a significant computational complexity saving can be achieved as compared to the conventional video combiner using transcoding approach. The new video combiner is then used to realize a multipoint video conferencing and some results are presented to show the improvement in performance due to our proposed architecture.

1. INTRODUCTION
With the advance of video compression, networking technologies and international standards, video conferencing is widely used in our daily life[1-5]. In fact, more and more video conferencing products are appearing on the market.

Figure 1 shows a typical scenario in which a video conferencing is held over a practical wide-area network such as the Public Switch Telephone Network (PSTN) or the Integrated Service Digital Network (ISDN) in which the channel bandwidth is constant and symmetrical. Assume that encoded video bitstream of each conference participant is R kb/s in a quarter common intermediate format (QCIF: 176 × 144 pixels). The MCU receives and decodes the multiple video bitstreams from all the conference participants. The decoded videos are combined in a common intermediate format (CIF: 352 × 288 pixels) through the video combiner. The combined video is re-encoded at R kb/s in order to fulfill the requirement of the channel bandwidth. Therefore video transcoding must be performed at the video combiner. Transcoding is regarded as a process of converting a previously compressed video bitstream into a lower bit rate compressed video bitstream.

Figure 1. An example of multipoint video conferencing.

Transcoding is a practical approach for video combining in multipoint video conferencing over a symmetrical wide-area network. 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. More importantly, additional degradation is introduced since the video quality of the transcoding approach suffers from its intrinsic double-encoding process. In order to provide a satisfactory visual quality of the combined and transcoded video, the re-distribution of bits in the re-encoding process to different parts of the combined video is a crucial step.

Some of the widely used transcoders for video conferencing application adopt requantization of DCT coefficients or frame skipping [2-4] to reduce the bandwidth of the inactive speakers. In this paper, we make use of the progressive properties of wavelet transform and the region of interest (ROI) features to build a video conferencing system.

Using successive quantization of the wavelet coefficients, scalable video in both ROI and background
can easily be obtained. Considering the unequal importance of various regions, a high decomposition level of wavelet transform and fine quantization level is not required in background region. Then the video combiner calculates the motion activity of each conference participant and adjusts the video quality by discarding the fine detail information bitstreams in the inactive conference participants or background according to the bandwidth requirement. Since no re-encoding process is required in the bit reallocation process, computational complexity can be greatly reduced in the video combiner. More importantly, video quality degradation can be avoided. The organization of this paper is described as follows:

The basic architecture and the features of our proposed video conferencing system will be given in section 2. In section 3, the progressive transmission in the video conferencing system is discussed. In section 4, the adaptive bit reallocation algorithm in the video combiner will be described. The experimental results are compared with the conventional video coder in section 5 while in section 6 we draw some conclusions.

2. THE PROPOSED VIDEO CODER AND VIDEO COMBINER

Figure 2 shows the system architecture of the wavelet-based video coder in multipoint video conferencing system[5]. Since the video conferencing system is wavelet-based, blocking artifacts are avoided[6-7]. The wavelet-based video conferencing system has four major features:

3. progressive transmission updating, and
4. adaptive bit reallocation algorithm in the video combiner

The purpose of region selection is to identify the region of interest in an image, e.g., the speaker’s face in video conferencing. This region is updated automatically by tracking the object’s motion using histogram information[5]. The wavelet-based coder is then applied separately to the foreground and the background. Because the size of the region of interest is small, the computation time can be reduced significantly. Adaptive bit allocation is then performed. It makes sure that the video quality of the foreground is always better than that of the background. This is particularly important for unstable networks or low bit rate applications. The progressive transmission updating is employed so that the video quality and the coding efficiency can be improved even under low bitrate environment. The video combiner extracts motion activities from the incoming video bitstreams and performs bit reallocation such that the output bitrate is the same as the incoming bitrate. In the bit reallocation process, low complexity can be achieved since the video is scalable and no re-encoding process is required. The video combiner only needs to calculate the motion activities and the targeted number of bits for each conference participants. According to the targeted number of bits for each conference participants, the video combiner adjusts the video quality level by discarding part of the high quality level video bitstreams, if necessary. In the following discussion, we will focus on scalable video combiner using successive quantization techniques which is our major contribution in this paper.

3. PROGRESSIVE TRANSMISSION IN VIDEO CONFERENCING SYSTEM

In the zero-tree coding, a successive quantization scheme is proposed. Initially, a coarse quantization level is applied to the wavelet coefficients in both ROI and background. In this quantization process, a base level video quality can be obtained as shown in Figure 3a. This coarse quantization level favours for low bitrate video conferencing applications. In order to enhance the video quality, a second level quantization level is applied to the residue of the quantized wavelet coefficients. Combining these two levels of quantized coefficients, an enhanced level video quality can be obtained as shown in Figure 3b. A final quantization level is applied to the residue of the quantized wavelet coefficients if the bandwidth is available. Combining these three levels of quantized wavelet coefficients, a high video quality can be achieved as shown in Figure 3c. The encoded wavelet bitstreams show no redundant information. Figure 2 presents the architecture of our proposed video coder using wavelet transform. Since the image at time t and t-1 are highly correlated, it is beneficial to make use of this correlation. Firstly, the difference between the video
signal at time t and the encoded image at time t-1 are fed into the encoder as an input signal. Then the input signal passes into the region of interest video encoder and the background encoder. According to the region of interest selected by the user, a simple tracking will be applied using histogram or chrominance information. After the encoding process, both video streams will be combined and transmitted to the client side for decoding. Note that the encoded image at time t-1 can be reconstructed in the front encoder. This arrangement is used to improve the video coding efficiency, video quality and facilitate the progressive transmission process.

Since the difference between the current frame and the previous reconstructed frame is small in both background and slow motion regions, less bits are required to encode these regions in a practical situation. In practice, even 30% of the targeted bits are spent in the background video coding, the video quality still can be improved progressively. Besides, as compared with the traditional video transcoder, no re-encoding process is required. Hence, high computational complexity and quality degradation can be avoided in the transcoding process.

![Figure 3: Different levels of video qualities. (3a) first level, (3b) second level and (3c) third level.](image)

4. ADAPTIVE BIT REALLOCATION ALGORITHM IN THE VIDEO COMBINER

According to the motion activities and the bandwidth constraints of each conference participants, the video combiner adjusts the scalable video quality level[6-7] for each conference participants. Firstly, our proposed video combiner extracts the motion activities information from the incoming bit streams as shown in Figure 4.

![Figure 4: Our proposed video combiner.](image)

In most multipoint video conferencing application, usually only one or two conference participants are active at a given time, while other participants have no or little motion. To make best use of the available bit rates, a rate control scheme is proposed to calculate the targeted bits based on the motion inside the region of interest. Since the speaker could have little motion when active, the motion activities is defined in terms of the magnitude of the motion of the region of interest and the energy of the audio signal as shown in following equations:

\[
\text{cost}_{\text{video}} = |\text{motion}_{\text{of the region of interest}}| \quad (1)
\]

and

\[
\text{cost}_{\text{audio}} = \lambda \cdot \text{Audio energy level} > \text{audio background energy threshold} \quad (2)
\]

\[
\text{cost} = \text{cost}_{\text{video}} + \text{cost}_{\text{audio}} \quad (3)
\]
where \( \text{cost}_i \) represents the importance of user \( i \) and \( \lambda \) is a constant. The video combiner evaluates the percentages of bits for every conference participants by normalizing the speaker activities as described in equation (4).

\[
\% \text{ of } \text{Bits assigned to } CP_i = \frac{\text{cost}_i}{\sum_i \text{cost}_i} \times 100\%
\] (4)

After the percentages of bits for every speaker are calculated, the video combiner adjusts the video quality level according to the bandwidth constraints for each conference participant. Consequently, more bits will be allocated to those sub-sequences with higher motion activities. Since the incoming video is scalable, our proposed video combiner can achieve a lower complexity by discarding the fine level or the second level quantized coefficients according to the bandwidth requirement and the users' activities.

5. EXPERIMENTAL RESULTS

The proposed video conferencing system is tested for the 64kbit/sec case. Figure 5 and Figure 6 show a comparison between our proposed system and the conventional DCT-based system. The overall performance is shown in Table 1. Although the background has a lower PSNR as compared to the conventional approach[4], the foreground has a much higher PSNR. Moreover, the subjective performance is much better than the conventional approach and the blocking artifacts are avoided. In fact, the subjective superiority is even more profound at low bit rates. In this case, the blocking artifacts associated with the DCT-based encoders are severe. However, by using our proposed video conferencing system, the quality in the foreground can still be maintained.

Table 1. Comparison of average PSNR's performances with the conventional video transcoder.

<table>
<thead>
<tr>
<th>Our proposed video conferencing system</th>
<th>Conventional video conferencing system</th>
</tr>
</thead>
<tbody>
<tr>
<td>Foreground Foreground</td>
<td>Background Background</td>
</tr>
<tr>
<td>CP1</td>
<td>36.4</td>
</tr>
<tr>
<td>CP2</td>
<td>36.8</td>
</tr>
<tr>
<td>CP3</td>
<td>37.8</td>
</tr>
<tr>
<td>CP4</td>
<td>35.6</td>
</tr>
</tbody>
</table>

6. CONCLUSIONS

A region-based video combiner is proposed. The proposed architecture consists of region of interest selection, wavelet-based encoders for the foreground and the background, motion-based region updating steps and adaptive bit re-allocation strategy in the video combiner. Based on the motion activities of the users, our proposed video combiner adjusts the speaker video quality without performing re-encoding process. Experimental results confirm that our proposed system produces a good video quality even under low bit rates for real-time video conferencing applications.

7. ACKNOWLEDGEMENTS

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

REFERENCES


