ISCAS 2000 - IEEE International Symposium on Circuits and Systems, May 28-31, 2000, Geneva, Switzerland

DYNAMIC RATE CONTROL IN MULTIPOINT VIDEO TRANSCODING

Chia-Wen Lin
Computer and Communications Research Labs
Industrial Technology Research Institute
Hsinchu, Taiwan 310, ROC

Te-Jen Liou, and Yung-Chang Chen
Department of Electrical Engineering
National Tsing Hua University
Hsinchu, Taiwan 300, ROC

ABSTRACT

This paper presents a dynamic rate control method for video transcoding to enhance the visual quality of the participants and regions of interest in multipoint video conferencing. This method firstly identifies the active conferees from the multiple incoming video streams by calculating the temporal and the spatial activities of the conferee sub-windows. The sub-windows of inactive participants are dropped and the saved bits are reallocated to the active sub-windows by using a rate-distortion optimized bit allocation approach. The simulation results show that the visual quality of the active sub-windows is significantly improved with the cost of degrading the temporal resolution of the inactive sub-windows which is relatively invisible to human perception. In addition, we also propose a dynamic distortion weighting adjustment based on H.263 TMN-8 framework to improve the quality of the regions of interest such as the face regions, since the face regions are usually the main focuses in video conferencing. The quality of face regions can be effectively enhanced at most frames in our simulation results. The computational complexity of the proposed algorithm is pretty low thus makes it well suited for real-time applications.

1. INTRODUCTION

With the rapid advance of video technologies, digital video applications become more and more popular in our daily life. In recent years several international standards such as H.261, H.263, MPEG-1, and MPEG-4 have been established to support various video services and applications. The ITU-T H.26x standards have been successfully adopted in commercial two-way video telephony applications. Video conferencing is an efficient way for business persons, engineers, scientists, etc. to exchange their information at remote locations. In a multipoint video conference, multiple participants are connected to the central server, referred to as Multipoint Control Unit (MCU), which coordinates and distributes video and audio streams among multiple participants in a video conferencing according to the channel bandwidth requirement of each participant. Video transcoder [1] is included in an MCU to combine the incoming video streams into a single stream and sent the re-encoded bit-stream back to each participant over the same channel with the required bit rates and formats. Bit-rate conversion from high bit-rate to low bit-rate in video transcoding will introduce video quality degradation. The visual quality and the available channel bandwidth need to be traded off in video transcoding to find a feasible solution.

The problem of how to efficiently redistribute the limited bit-rates in different parts of a video in video transcoding is quite practical in providing satisfactory viewing experience. Region-based rate control strategies involving segmentation of the regions/areas of interest and dynamic bit allocation to the regions of interest based on some quality criteria are now attracting many researchers' attention and considered as an efficient way to resolve the problem mentioned above. In video telephony applications, often the talker's head-and-shoulder image is viewed on the display. Thus the face area is often the region of interest drawing the viewers' attention most of the time. It thus deserves to allocate more bits to the face regions to obtain sharper face quality by sacrificing the quality of the other regions of less interest to some acceptable extent from the viewpoint of human visual system (HVS). Several recent studies [3] suggested to allocate more bit-rates to the regions of interest (e.g., face regions in video telephony applications, and motion regions) so as to produce better visual quality.

Moreover, in a multipoint video conference, most of the time only one or two conferees are active at one time. The active conferees need higher bit-rates to produce good quality video while the inactive conferees only requires less bit-rates to produce acceptable quality video [4]. Simply uniformly distributing the bit-rates to the conferees will result in non-uniform video quality. To make best use of the bit-rates resources, a joint rate control scheme which can take into account each sub-window's activity is preferred [4,5], which has also been studied for the statistical multiplexing (StatMux) problem. Sun *et al.* [4] proposed to measure the activity of each sub-window by calculating the sum of the magnitudes of its corresponding motion vectors, and allocate the bit-rates to each sub-window according to its activity. Thus more bits will be allocated to those sub-windows with higher activities. Wu *et al.* [5] suggested to allocate the bit-rates to each sub-window according to its spatial-temporal activity which takes into account the motion, variance of the residual signal, and the number of the encoded macroblocks. A similar idea appeared in MPEG-4 multiple video objects coding [6].

The flow chart of the proposed transcoding algorithm is depicted in Fig.1. The proposed algorithm is based on the H.263 TMN-8 framework. In our approach, the bit allocation is divided into three layers: frame, sub-window, and macroblock (MB). After frame layer bit allocation, we propose a dynamic sub-window skipping scheme which classifies the sub-windows into active and inactive classes by estimating the corresponding spatial-temporal activities. The inactive sub-windows are then dropped so that the saved bits can be used to enhance the visual quality of the active ones. The rate control is subsequently down to the MB layer, and those regions of interest (e.g., face MB's and MB's with high motion) are further enhanced by the proposed MB-layer dynamic distortion weighting adjustment scheme.

0-7803-5482-6/99/$10.00 ©2000 IEEE

2. PROPOSED RATE CONTROL FOR MULTIPOINT TRANSCODING

A. H.263 TMN-8 Rate and Distortion Models

H.263 TMN-8 test model suggested the rate and the distortion models which relate bit rate and distortion in terms of quantization parameters and the signal variances to estimate the produced bits and the distortion of an MB as follows [2]:

$$B_{m,n} = A(K_m \frac{\sigma_{m,n}^2}{Q_{m,n}^2} + C_m), \tag{1}$$

$$D_{m,n} = \alpha_{m,n}^2 \frac{Q_{m,n}^2}{12} \tag{2}$$

where $B_{m,n}$, $D_{m,n}$, and $Q_{m,n}$ represent the number of bits, the estimated distortion, and the quantization step-size of the *n*th macroblock of the *m*th sub-window in a frame respectively. K_m is a model parameter for the *m*th sub-window. *A* is the number of pixels in a macroblock, $\sigma_{m,n}^2$ represents the macroblock variance of the motion-compensated residual signal, C_m is the average rate to encode the motion vectors and the bit-stream header for the *m*th sub-window. $\alpha_{m,n}$ is the distortion weighting factor, and is set to:

$$\alpha_{m,n} = \begin{cases} 2\frac{B_m}{AN}(1-\sigma_{m,n}) + \sigma_{m,n}, & \frac{B_m}{AN} < 0.5 \\ 1 & \textit{otherwise} \end{cases} \tag{3}$$

where B_m is the target allocation for the *m*th subwindow, *N* is the number of macroblocks in a sub-window, thus B_m */AN* is the average bit-rate in bits/pixel for the *m*th sub-window. The optimal quantization step-size can be decided by finding the solution of

$$Q_{m,1}^*,\ldots,Q_{m,N}^* = \arg\min_{Q_{m,1},\ldots,Q_{m,N}} \frac{1}{N}\sum_{n=1}^{N} \alpha_{m,n}^2 \frac{Q_{m,n}^2}{12}$$

$$\textit{subject to} \quad \sum_{n=1}^{N} B_{m,n} \le B_m^{\text{target}}$$

The Lagrange multiplier can be used to solve this constrained problem. The optimized quantization step-size was obtained by the following equation [2]:

$$Q_{m,n}^* = \sqrt{\frac{AK_m}{(B_m^{\text{target}} - ANC_m)} \frac{\sigma_{m,n}}{\alpha_{m,n}} \sum_{n=1}^{N} \alpha_{m,n}\sigma_{m,n}} \qquad i = 1,\ldots\ldots,N \tag{4}$$

B. Bit Reallocation Using Dynamic Sub-Window Skipping (DSWS)

In multipoint video conferencing, the talker is often the most important person. Since the talker usually has larger motion than others, the sub-window containing the talker usually attracts our attention [4]. We thus propose to skip the static sub-window and use the saved bits to enhance the quality of non-skipped sub-windows. The incoming motion vectors and the associated motion-compensated distortion are used to calculate the temporal and spatial activity measures respectively for dynamic sub-window skipping. The decision rule is described as follows:

```
if $(S_m^{MV} < TH_{\text{MV1}}) \,\&\&\, (\frac{SAD_m - SAD_m^{\text{prev}}}{SAD_m^{\text{prev}}} < TH_{\text{SAD1}})$
then
        Skip the mth sub-window
else
        Encode the mth sub-window
```

where the sum of the magnitude of the motion vectors of the MB layer and sub-window-layer are defined as:

$$S_{m,n}^{MV} = \left|MV_{m,n}^x\right| + \left|MV_{m,n}^y\right|. \tag{5}$$

$$S_m^{MV} = \sum_{n=1}^{N} S_{m,n}^{MV} \tag{6}$$

The sums of absolute difference (SAD) of the macroblock level and the sub-window level are defined respectively as follows:

$$SAD_{m,n} = \sum_{x,y\in \text{MB}_{m,n}} \left| f_m(x,y) - f_m^{\text{prev}}(x + MV_{m,n}^x, y + MV_{m,n}^y) \right|, \tag{7}$$

$$SAD_m = \sum_{n=1}^{N} SAD_{m,n}. \tag{8}$$

where $(MV_{m,n}^x, MV_{m,n}^y)$ is the motion vector associated with the *n*th macroblock of the *m*th sub-window with respect to its corresponding latest encoded sub-window (i.e., $f_m^{\text{prev}}(\cdot)$). The sum in (5-6) indicates the MB and sub-window motion activity. A sub-window is classified as active if the sum is larger than a predetermined threshold, otherwise it is classified as static. A static sub-window is considered to be skipped and, once it's skipped, the corresponding latest un-skipped sub-window is repeated to approximate the skipped sub-windows. Human visual perception is relatively insensitive to the little difference between the skipped sub-window and its reconstructed one from sub-window repetition if the sub-window is static. The thresholds, TH_{MV1} and TH_{SAD1}, are set as the border for classification, the larger the thresholds are set, the more the sub-windows will be skipped, and the more the saved bits will be used in other sub-windows but jerky motions will become more serious. The SAD value of each sub-window is used to constrain the frame skipping. If the current frame is static but the accumulated residual is larger than a threshold, we enforce the frame to encode. This measure can prevent error accumulation by using only the motion activity measure. Since the SAD value is a byproduct of the motion compensation operation, it does not need extra computation. So the extra computational cost required for determining sub-window skipping is pretty low. Moreover the computation saving is also achieved since no transcoding operation is required for the skipped sub-windows.

After determining which sub-windows should be skipped, we propose to use the following bit allocation scheme for the un-skipped sub-windows :

$$B_m^* = \tilde{S}_m^{\text{S-T}} \left[B_{\text{target}} - (M - M_i)B_{\text{skip}} - AN\sum_{m=1}^{M_i} C_m \right] + ANC_m, \quad m = 1,\ldots,M_i \tag{9}$$

where the joint spatial-temporal activity measure is

$$\tilde{S}_m^{\text{S-T}} = \alpha_\sigma \frac{S_m^\sigma}{\sum_{m=1}^{M_i} S_m^\sigma} + \alpha_{\text{MV}} \frac{S_m^{\text{MV}}}{\sum_{m=1}^{M_i} S_m^{\text{MV}}} + \alpha_{\text{MB}} \frac{S_m^{\text{MB}}}{\sum_{m=1}^{M_i} S_m^{\text{MB}}} \tag{10}$$

S_m^{MV} is the sum of the magnitudes of the motion vectors in the *m*th sub-window as defined in (6) and S_m^{MB} is the number of the coded macroblocks (i.e. the COD bit is "0") in the *m*th sub-window of the incoming video. And

$$S_m^\sigma = \sum_{n=1}^{N} \alpha_{m,n}\sigma_{m,n} \tag{11}$$

The weighting factors $(\alpha_\sigma, \alpha_{\text{MV}}, \alpha_{\text{MB}}) \in [0,1]$ and

$$\alpha_{\sigma} + \alpha_{\text{MV}} + \alpha_{\text{MB}} = 1 \qquad (12)$$

M_i is the number of the non-skipped sub-windows in a frame.

C. Face-Assisted MB-Level Bit Reallocation

In (2), the parameter $\alpha_{m,n}^2$ represents the distortion weight, and the larger its value is, the more important $MB_{m,n}$ will be, which provides an efficient means to dynamically control the distortion weighting. With this distortion model, the proposed dynamic distortion weighting adjustment (DDWA) of different regions of interest can be performed by assigning different weighting factors to the MB's belonging to different regions. We propose to classify the MB's into three regions: face region, active non-face region, and static non-face region. The face detection and tracking algorithm proposed in [7] is adopted to segment out the face regions. The rule used in our proposed algorithm to classify the active and static non-face MB's is as follows:

```
$if(\text{MB}_{m,n} \notin Face\_Region)$
    $if(SAD_{m,n} < TH_{\text{SAD2}})\,\&\,\&(S_{m,n}^{\text{MV}} < TH_{\text{MV2}}),$
            $\text{MB}_{m,n} \in Inactive\_Non\_Face\_Region$;
    $else$
            $\text{MB}_{m,n} \in Active\_Non\_Face\_Region$;
```

where $SAD_{m,n}$ and $S_{m,n}^{\text{MV}}$ are defined in (7) and (5).

Because the MB's of non-face regions may belong to background or human body, larger distortions in the non-face regions are tolerable to viewer's perception since the face region is usually the focus in video telephony applications. We propose to skip the static non-face MB's by setting the COD bit to 1 in H.263 syntax, and the saved bits are used to compensate the quality of the face and active non-face MB's using different weightings. The proposed MB-layer DDWA approach is summarized as follows:

```
for n = 1 to  N
{
    Set the initial $\alpha_{m,n}$ as in (3)
    if $(\text{MB}_{m,n} \in Face\_Region)$

        $\alpha_{m,n} = K_{\alpha} \cdot \sigma_{m,n} \cdot \dfrac{Face\_SAD_m / Face\_Area_m}{Total\_Face\_SAD / Total\_Face\_Area}$
    else
            if  $\text{MB}_{m,n} \in Inactive\_Non\_Face\_Region$
                $\alpha_{m,n} = 0$
            endif
    endif
    calculate $Q_{m,n}$ using (4)
}
```

where $Face\_SAD_m$ is the SAD of face regions in the mth sub-window, { $m=1,2,3,4$ }, and $Total\_Face\_SAD$ represents the sum of $Face\_SAD$ of the non-skipped sub-windows in a frame. $Face\_Area_m$ denotes the number of the MB's belonging to face region in sub-window m, $Total\_Face\_Area$ is the total number of the face MB's in a frame. K is the weight to determine the enhancement effect. With the above algorithm, the distortion weighting factors, $\alpha_{m,n}^2$, will be magnified by a ratio for the face MB's, thus finer quantization parameters will be used and more bits will be allocated to the face MB's. The quality on face region can thus be effectively enhanced. On the other hand, the quality of the static non-face region will be sacrificed, and only a little bit difference appears on the active non-face region. The improvement on the face region is significant as will be shown from the experimental results, while the degradation on the non-face region due to DDWA is relatively invisible to human perception in video telephony applications. Moreover, another advantage of the proposed DDWA method is that the computations (DCT, quantization, inverse quantization, and IDCT) required for the skipped MB's can be saved, thus the computational cost can be further reduced.

3. EXPERIMENTAL RESULTS AND CONCLUSION

In our experiments, four 400-frame QCIF image sequences were captured at 30 frames/sec from a 4-point video conference. We firstly encoded the QCIF image sequences at 128 Kbits/sec and 15 frames/sec as the input sequences of the video transcoder. Then the four QCIF sequences are combined into a CIF sequence and subsequently jointly transcoded to meet the predetermined target rate. The target rate for the combined CIF video is also set to 128 Kbits/sec, and the output frame rate is kept 15 frames/sec. The compression ratio of the transcoder is thus four in our experiments.

Table 1 compares the average PSNR performance for three transcoding methods: the direct transcoding using TMN8 rate control, the DSWS method, and the DSWS method followed by the R-D optimized bit reallocation based on joint spatial-temporal activity in (9) (DSWS+RDST). It shows that the proposed DSWS method brings the PSNR gain on the sub-windows with relative high activities, while the low-activity sub-windows are degraded. The PSNR of a skipped sub-window is computed from the incoming QCIF image and the latest previously reconstructed non-skipped sub-window. As shown in Table 1, the proposed DSWS scheme achieves 0.11dB average PSNR improvements on the non-skipped sub-windows. In sub-window 4, the performance is degraded by 0.38dB because it's with relatively low motion in several long intervals. In practices the degradation is relatively insensible to viewers' perception. The proposed DSWS+RDST method can further emphasize the sub-windows with higher joint spatial-temporal activities. Fig.2 shows the performance comparison of face regions with our proposed scheme and the direct cascaded transcoding. In the simulation results, the quality of face regions can really be improved in most frames. The largest PSNR improvement is 6.55 dB, while the maximal degradation is about 6.44 dB. The degradation is mainly introduced by sub-window repetition due to DSWS. The visual quality degradation of the skipped frame is, however, relatively insensitive in non-active sub-windows with low temporal activity. The proposed sub-window skipping method can save bits from the static frame, and low computation cost is required for determining sub-window skipping. This method can enhance the quality of the other non-skipped sub-windows without introducing significant quality degradation on the skipped sub-windows. A face-assisted MB-layer rate control scheme is proposed to further enhance the visual quality of face regions. This method is particularly useful in videoconferencing applications since the focuses in such applications are mainly on the face regions. Simulation results verify the effectiveness of the proposed method and the extra computational load is pretty low in our proposed scheme. The computation saving is further achieved by skipping the inactive sub-windows and MB's.

4. REFERENCES

[1] G. Keesman et al., "Transcoding of MPEG bitstream," *Signal Proc.. Image Commun.*, pp. 481-500, 1996.

[2] J. Ribas-Corbera and S. Lei, "Rate control in DCT video coding for low-delay communications," *IEEE Trans. Circuits Syst. Video Technol.,* vol 9, pp. 172-185, Feb. 1999.

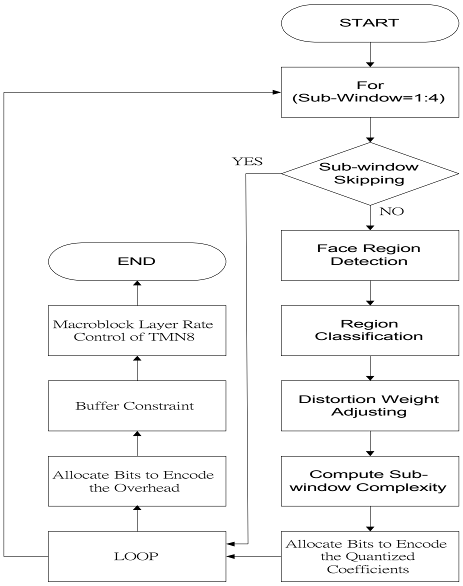
[3] S. Chai and K. N. Ngan, "Face segmentation using skin-color map in videophone applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol 9, pp. 551-564, Jun. 1999.

[4] M.-T. Sun, T.-D. Wu, and J.-N. Hwang, "Dynamic bit allocation in video combining for multipoint v conferencing," *IEEE Trans. on Circuit and Systems.,* vol. 45, No. 5, pp. 644-648, May. 1998

[5] T.-D. Wu, J.-N. Hwang, and C.-W. Lin, "Dynamic bit-rate conversion and bit re-allocation in multipoint video conference," submitted to *IEEE Trans. Circuits Syst. Video Technol.*,May, 1999

[6] A. Vetro, H. Sun, and Y. Wang, "MPEG-4 rate control for multiple video objects," *IEEE Trans. Circuits Syst. Video Technol.*, vol 9, pp. 186-199, Feb. 1999.

[7] C.-W. Lin, Y.-J. Chang, E. Fei, and Y.-C. Chen, "Efficient Video Coding with R-D Constrained Quadtree Segmentation," *PCS99,* pp. 57-60*,* Portland, Apr., 1999.

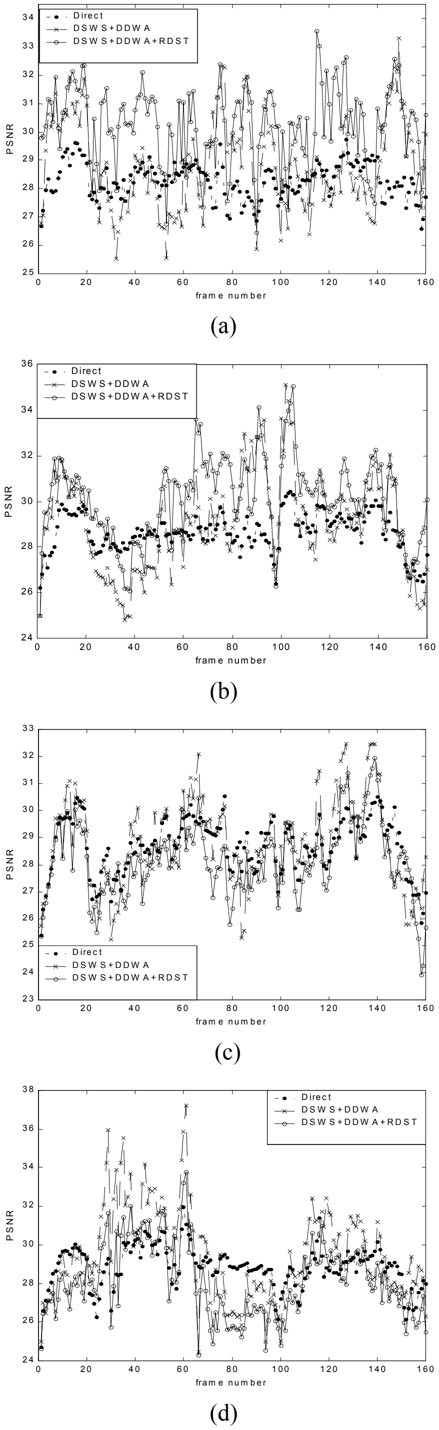


START
For (Sub-Window=1:4)
Sub-window Skipping
YES
NO
Face Region Detection
Region Classification
Distortion Weight Adjusting
Compute Sub-window Complexity
Allocate Bits to Encode the Quantized Coefficients
LOOP
Allocate Bits to Encode the Overhead
Buffer Constraint
Macroblock Layer Rate Control of TMN8
END

Fig. 1. The proposed dynamic rate control algorithm for multipoint video transcoding.

Table 1. Average PSNR Comparison with the proposed DSWS, DSWS+RDST, and direct transcoding schemes

	Average PSNR of all frames (dB)			Average PSNR of non-skipped frames	
	Direct	DSWS	DSWS+ RDST	DSWS	DSWS+ RDST
Sub-window 1	30.03	30.25	31.39	30.31	31.46
Sub-window 2	29.92	30.19	30.24	30.28	30.33
Sub-window 3	30.41	30.55	29.98	30.53	29.99
Sub-window 4	29.78	29.40	28.38	29.49	28.53
Average	30.04	30.10	30.00	30.15	30.08



(a)

(b)

(c)

(d)

Fig. 2. PSNR comparison of the proposed methods and direct cascaded transcoding on face regions: (a) letft top; (b) left bottom; (c) right top; (d) right bottom.