Dynamic Region of Interest Transcoding for Multipoint Video Conferencing

Chia-Wen Lin, *Member, IEEE*, Yung-Chang Chen, *Senior Member, IEEE*, and Ming-Ting Sun, *Fellow, IEEE*

***Abstract*—This paper presents a region of interest transcoding scheme for multipoint video conferencing to enhance visual quality. In a multipoint video conference, usually there are only one or two active conferees at one time, which are the regions of interest to the other conferees involved. We propose a dynamic sub-window skipping scheme to firstly identify the active participants from the multiple incoming encoded video streams by calculating the motion activity of each sub-window and then dynamically reduce the frame rates of the motion inactive participants by skipping these less-important sub-windows. The bits saved from the skipping operation are reallocated to the active sub-windows to enhance the regions of interest. We also propose a low-complexity scheme to compose, as well as trace, the unavailable motion vectors with a good accuracy in the dropped inactive sub-windows after performing sub-window skipping. Simulation results show that the proposed methods not only significantly improve the visual quality of the active sub-windows without introducing serious visual quality degradation in the inactive ones, but also reduce the computational complexity and avoid whole-frame skipping. Moreover, the proposed algorithm is fully compatible with the H.263 video coding standard.**

***Index Terms*—Bit-rate control, multipoint control unit (MCU), video coding, video conference, video transcoding.**

I. Introduction

WITH THE rapid advance of video technologies, digital video applications have become increasingly popular in our daily life. In recent years, several international standards such as H.261 [1], H.263 [2], MPEG-1 [3], MPEG-2 [4], and MPEG-4 [5] have been established to support various video services and applications. In these standards, H.261 and H.263 have been successfully adopted in two-way video telephony applications. Video telephony is an efficient way for businesspersons, engineers, scientists, etc. to exchange information at remote locations. With the rapid growth of video telephony, the need of multipoint video conferencing is also growing. A multipoint videoconference involves three or more conference participants. In continuous presence video conferencing, each conferee can see others in the same window simultaneously [6]. Fig. 1 depicts an application scenario of multiple persons participating in a multipoint videoconference with a centralized server. In this scenario, multiple conferees are connected to the central server, referred to as the multipoint control unit (MCU), which coordinates and distributes video and audio streams among multiple participants in a video conference according to the channel bandwidth requirement of each conferee. A video transcoder [7]–[10] is included in the MCU to combine the multiple incoming encoded video streams from the various conferees into a single coded video stream and send the re-encoded bit stream back to each participant over the same channel with the required bit rate and format for decoding and presentation. In the case of a multipoint video conference over a public switch telephone network (PSTN) or integrated service digital network (ISDN), the channel bandwidth is constant and symmetrical. Assuming each conferee has a channel bandwidth of B kb/s, then MCU receives the conferees' videos at B kb/s each, decodes and combines the videos, and re-encodes the combined video at B kb/s so as to meet the channel bandwidth requirements for sending back the encoded video to the conferees. Therefore, it is required to perform bit-rate conversion/reduction at the video transcoder. Bit-rate conversion from high to low bit rate in video transcoding will, however, introduce video quality degradation. The visual quality, computational load, and used bit rates need to be traded off in video transcoding to achieve a good solution.

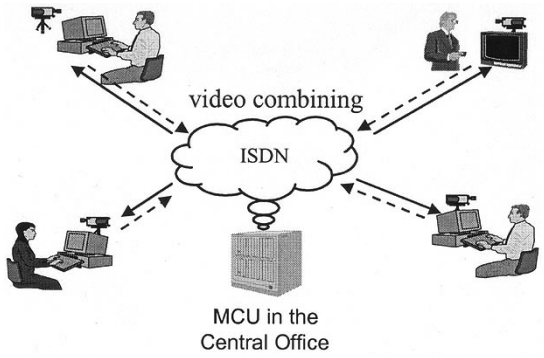


Fig. 1. Application example of multipoint video conferencing.

The problem of how to efficiently redistribute the limited bit rates to different parts of a video in video transcoding is critical in providing satisfactory visual quality. In a multipoint videoconference, usually 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 require lower bit rates to produce acceptable quality video [9]. Simply uniformly distributing the bit rates to the conferees will result in nonuniform video quality. To make the best use of the available bit rates, a joint rate-control scheme, which takes into account each conferee sub-window's activity, can be used [9], [10]. Sun *et al.* [9] proposed to measure the motion activity of each sub-window

Manuscript received April 4, 2001; revised December 1, 2002. This paper was recommended by Associate Editor H. Watanabe.

C.-W. Lin and Y.-C. Chen are with the Department of Computer Science and Information Engineering, National Chung Cheng University, Chiayi, Taiwan 621, R.O.C.

M.-T. Sun is with the Information Processing Laboratory, University of Washington, Seattle, WA 98195 USA.

Digital Object Identifier 10.1109/TCSVT.2003.816505

1051-8215/03$17.00 © 2003 IEEE

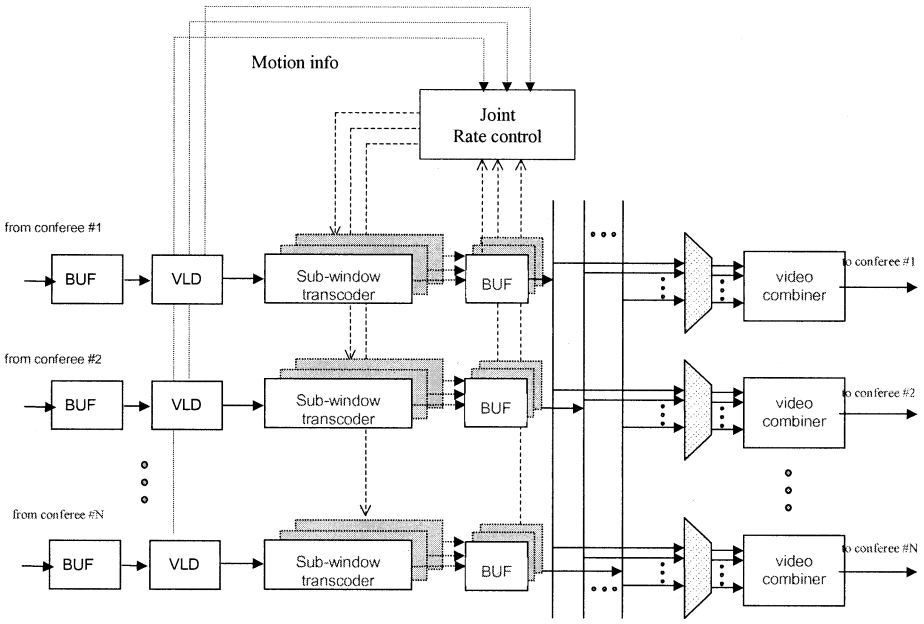


Fig. 2. Proposed multipoint video transcoding architecture.

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, thereby producing much more uniform quality video. Wu and Hwang [10] extended the work in [9] by allocating the bits to each sub-window according to its spatial-temporal activity, which takes into account the motion, variance of the residual signal, and number of encoded macroblocks. Similar work on joint rate control can also be found in the statistical multiplexing (StatMux) of multiple video programs [11]–[13] and MPEG-4 joint rate control of multiple video objects [14], [27]. However, the strong correlation among the conferees in a multipoint video conference usually does not exist in the general cases of the StatMux and MPEG-4 object rate control. In addition, we will show that dynamic temporal resolution control [15] for each sub-window, which we first proposed in [16], may achieve further coding gain and computation reduction in multipoint video transcoding.

In this paper, we present a dynamic sub-window skipping (DSWS) scheme, which provides the flexibility that sub-windows can be encoded in different temporal resolutions according to their motion activities. The proposed DSWS scheme classifies the sub-windows into active and inactive classes by calculating the associated motion activities. The inactive sub-windows can then be dropped without transcoding so that the saved bits can be used to enhance the visual quality of the active ones without introducing serious degradation on the inactive ones. In addition to the performance gain on active sub-windows, the DSWS scheme also presents two other advantages: achieving computation reduction and avoiding the whole-frame skipping. On the other hand, the sub-window skipping will also cause some problems. The first problem is that dropping sub-windows will cause visual motion jerkiness to some degree, which depends on the number of consecutive sub-windows dropped and the associated motion activities. As will be shown later, this problem can be alleviated by appropriately selecting the frames to be dropped. Second, it will cause a motion vector missing problem, which is similar to that mentioned in [15] and [18]. In order to resolve the problem, we present a motion-vector composing scheme, which can compose and trace the unavailable motion vectors in the dropped sub-windows with accuracy at a better than previous algorithms and very low computational and memory cost.

The remainder of this paper is organized as follows. Section II presents the proposed DSWS scheme. In Section III, a pre-filtered activity-based motion-vector composing scheme is proposed for composing the unavailable motion vectors in the dropped sub-windows after performing the DSWS scheme. Section IV reports the experimental results of the proposed algorithms and the comparison with the H.263 TMN8 [17] direct transcoding method. Finally, conclusions are drawn in Section V.

II. DSWS

Fig. 2 depicts the architecture for multipoint video transcoding discussed in this paper. In this architecture, the input data for the video transcoder consist of multiple H.263 encoded video streams from the client terminals through a heterogeneous network environment. The video streams could be transmitted through PSTN, ISDN, or local area network (LAN) with various bandwidth requirements. For simplicity, but without loss of generality, we assume each video stream is encoded in the Quarter Common Intermediate Format (QCIF; 176×144) format and each participant can see four participants in a Common Intermediate Format (CIF; 352×288) frame in a continuous presence fashion. In our experiments, we assume the video transmission is over ISDN, as shown in the scenario in Fig. 1. As shown in Fig. 2, the input video streams are first buffered at the input to regulate the difference

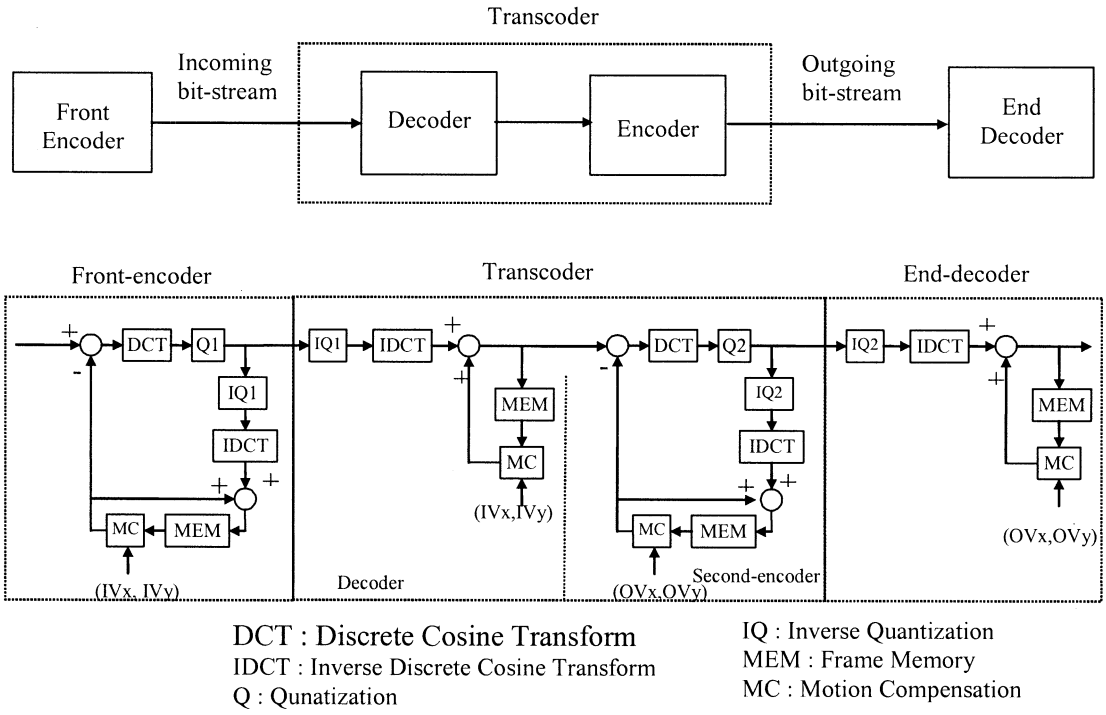


Fig. 3. Cascaded pixel domain transcoder architecture.

between the input data rate and the transcoding data rate for each video frame. Each video stream is decoded through a variable length decoder (VLD) and then transcoded into lower data rates. To meet various user bit-rate requirements, more than one transcoder may be required for the same video stream to generate multiple video streams with different bit rates. The transcoded bit streams are subsequently combined into CIF frames through a number of multiplexers and video combiners. The multiplexing unit for the video combiner is the group of blocks (GOBs), as specified in the H.263 standard [2].

Fig. 3 shows the cascaded pixel-domain transcoder used for transcoding each sub-window in our method. This transcoder provides the flexibility that the decoder and encoder loops can operate at different bit rates, frame rates, picture resolutions, coding modes, and even different standards, since they can be totally independent to each other, while this may not be achievable in the simplified transcoder in [7] and the compressed-domain transcoder in [8]. It can also be implemented to achieve a drift-free operation for conversing the bit rate and spatial/temporal resolution if the implementations of inverse discrete cosine transform (IDCT) in the front encoder and the end decoder are known. In this case, the decoder loop and the encoder loop can be implemented to produce exactly the same reconstructed pictures as those in the front encoder and end decoder, respectively. If the implementations of the IDCTs are not known, as long as they satisfy the IDCT standards specification [19] and the macroblocks are refreshed, as specified in the standards [1]–[5], the drift will not be a major issue. Since some information (e.g., coding modes and motion vectors) extracted from the incoming video bit stream after the decoding can be reused at the encoding, the overall complexity is not as high as the sum of a decoder and an encoder [18], [21].

As mentioned above, in a multipoint video conference, usually only one or two persons are motion active at one time. The active conferees (e.g., the conferees who are speaking) are often the center of focus. Therefore, allocating the active sub-windows with relatively higher bit rates can provide a much better visual experience to the viewers. The active conferees usually have larger motions than others, thus, they can be easily detected by using the incoming motion information.

In this paper, we observe that, in multipoint video conferencing, the temporal resolutions of the low-active sub-windows may be reduced without introducing significant visual degradation from the human visual system (HVS) point of view. The main reason is that, since the motions presented by the inactive sub-windows are relatively slow and the conferees usually concentrate their focuses on the active ones, the effect of the temporal resolution reduction by skipping inactive sub-windows can often be masked by the high motions in the active sub-windows and, thus, is not sensitive to viewers' perceptions. To make best use of this property, we propose to drop motion inactive sub-windows by using sub-window repetition to approximate those dropped sub-windows at the end decoder so that the saved bits can be used to enhance the quality of the remaining non-skipped active ones, which are usually the regions of interest. In addition, if a sub-window is decided to be skipped, much computation in transcoding this sub-window can be saved, thus, significant computation reduction can be achieved. Sub-window skipping can be implemented in the H.263 syntax by simply setting all the coded macroblock indication (COD) bits [2] of the macroblocks belonging to the skipped sub-windows to "1" to get rid of sending the associated DCT coefficients, motion vector information, and MB overhead bits. Only 99 bits (for 99 macroblock COD bits, respectively) are required to represent a skipped QCIF sub-window, thus, the overhead is relatively neg-

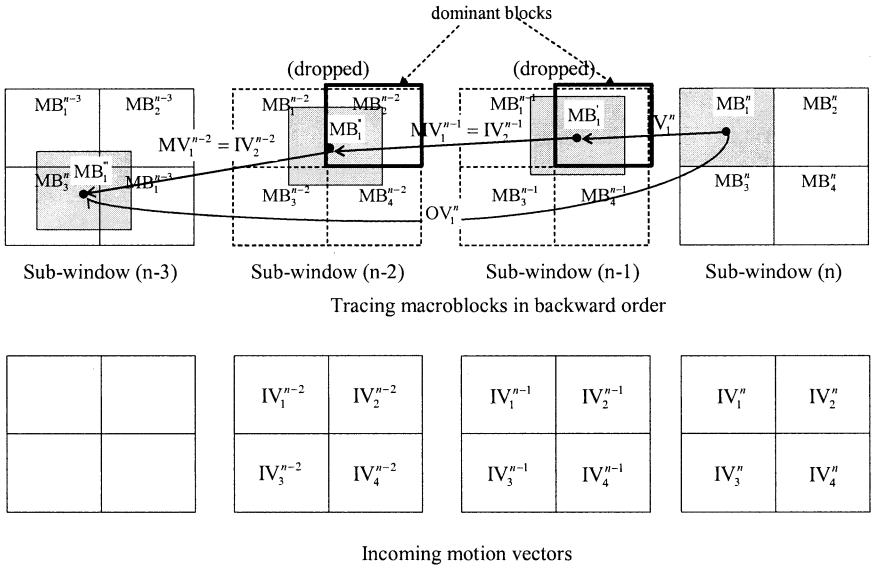


Fig. 4. Motion vector composition with FDVS [18].

ligible. In our proposed DSWS scheme, the motion information is used to calculate the motion activity of each sub-window for DSWS control. The DSWS scheme is summarized as follows:

```
if (S_m^MV < TH_MV)&&(MAAD_m < TH_MAAD)
then
 Skip the transcoding of the mth sub-
window
else
 Transcode the mth sub-window
```

where the mean accumulated magnitude of motion vectors of the mth sub-window is defined as

$$S_m^{MV} = \frac{1}{N}\sum_{n=1}^{N}\left(\left|MV_{m,n}^{x}\right| + \left|MV_{m,n}^{y}\right|\right) \tag{1}$$

N is the number of MBs in a sub-window, $(MV_{m,n}^{x}, MV_{m,n}^{y})$ is the motion vector associated with the nth MB of the mth sub-window with respect to its corresponding previously encoded sub-window (i.e., $f_m^{\text{prev}}(\cdot)$), and the mean of accumulated absolute difference (MAAD) of the sub-window is defined as

$$\begin{aligned}\text{MAAD}_m &= \frac{1}{N}\sum_{n=1}^{N}\sum_{x,y\in \text{MB}_{m,n}} \left|f_m(x,y) - f_m^{\text{prev}}(x+MV_{m,n}^{x}, y \right. \\ &\qquad \left. +MV_{m,n}^{y})\right|. \end{aligned} \tag{2}$$

The mean accumulated magnitude of motion vectors of a sub-window can be used as a good indication of its motion activity. A sub-window is classified as active if the sum is larger than a predetermined threshold TH_{MV}, otherwise it is classified as inactive. An inactive sub-window will be skipped if its associated MAAD value defined in (2) is below a threshold TH_{MAAD}. If an inactive sub-window is skipped, the corresponding latest nonskipped sub-window is repeated to approximate the skipped sub-windows. Human visual perception is relatively insensitive to the little differences between the skipped sub-windows and their reconstructed ones from sub-window repetition if the skipped sub-windows are inactive. The two thresholds TH_{MV} and TH_{MAAD} are used for the 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 MAAD value of each sub-window is used to constrain the sub-window skipping. If the current frame is inactive, but the MAAD is larger than a threshold, the proposed method enforces that the sub-window, which would otherwise be skipped, be encoded. This measure can prevent the error accumulation caused by slow, but steady motions by using only the motion activity measure, as in [15]. Note, if cascaded discrete cosine transform (DCT)-domain transcoders [26] are adopted, the DCT-domain counterpart of (2) (e.g., accumulated sum of absolute DCT coefficient differences) can be used instead. It should also be noted that since the incoming sub-windows may be dropped in consecutive frames with the DSWS method, the incoming motion vectors may not be valid since they may point to the dropped sub-windows that do not

TABLE I
PERFORMANCE COMPARISON OF DIFFERENT MOTION VECTOR ESTIMATION AND COMPOSITION METHODS. INCOMING BIT STREAMS OF 128 kb/s AND 30 f/s WERE TRANSCODED INTO 32 kb/s AND 7.5 f/s

Test sequence	MV composition method	Average PSNR (dB)
Foreman	Full-scale ME	27.39
	Interpolation	23.72
	FDVS	25.51
	PA-FDVS	25.67 (+1.6; -0.6)
Carphone	Full-scale ME	29.47
	Interpolation	27.07
	FDVS	28.16
	PA-FDVS	28.27 (+2.9; -0.8)
Football	Full-scale ME	35.18
	Interpolation	32.38
	FDVS	34.02
	PA-FDVS	34.47 (+1.7; -0.3)
Akiyo	Full-scale ME	42.95
	Interpolation	41.61
	FDVS	42.35
	PA-FDVS	42.38 (+2.2; -2.1)

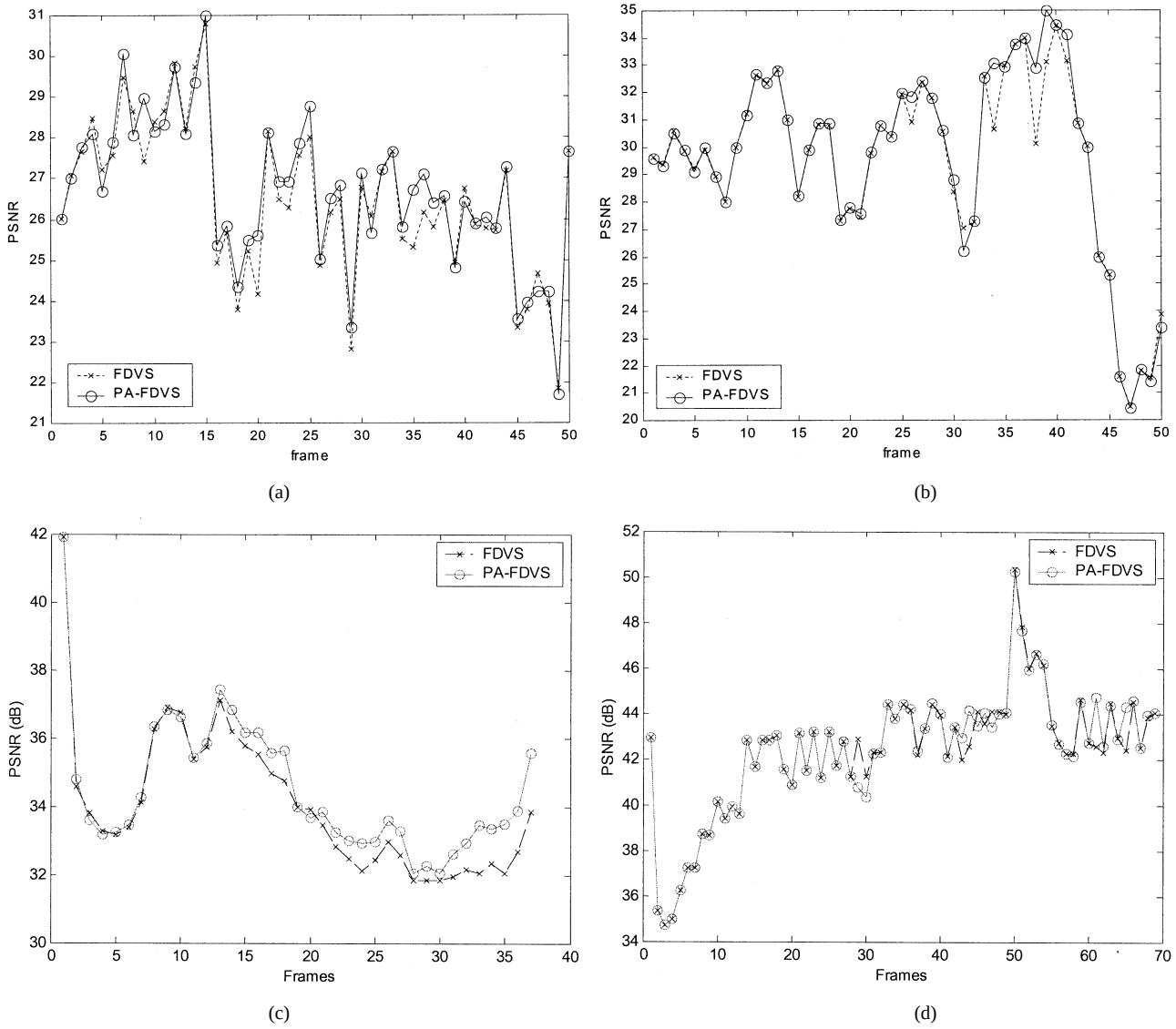


Fig. 5. Performance comparison of the FDVS and PA-FDVS schemes for the "Foreman" sequence. Incoming bit streams of 128 kb/s and 30 f/s are transcoded to the: (a) "Foreman" sequence, (b) "Carphone" sequence, (c) "Football" sequence, and (d) "Akiyo" sequence.

exist in the transcoded bit stream. To compose and trace the required, but unavailable motion vectors along the consecutively skipped sub-windows with respect to the corresponding latest encoded sub-windows, the motion vector compositing scheme proposed in Section III is used.

The proposed DSWS scheme presents several advantages. First, the quality of the active sub-windows can be effectively enhanced. The quality loss on the inactive sub-windows is relatively small and visually insensitive to the viewer's perception. Second, skipping a sub-window implies saving much computation in transcoding that sub-window (in our simulation, about 2/3 of the computation in transcoding that sub-window can be saved), thus achieving significant computation reduction. Finally, by skipping the motion inactive sub-windows, many whole-frame skippings due to insufficient bit allocation can be avoided so that the temporal resolution of the motion active sub-windows can be kept as high as possible. Moreover, the proposed method can be combined with the dynamic bit-allocation scheme presented in [10] and [16] to further improve the visual quality with almost no extra complexity.

III. Composing Motion Vectors in the Skipped Sub-Windows

After performing DSWS, sub-windows may be dropped in consecutive frames. However, the motion vectors in the dropped sub-windows are usually unavailable in the incoming bit stream. For example, in Fig. 4, a situation where one sub-window is dropped in two consecutive frames in transcoding is illustrated. In this example, the equivalent outgoing motion vector of the block MB_1^n should be $\mathrm{OV}_1^n = \mathrm{IV}_1^n + \mathrm{MV}_1^{n-1} + \mathrm{MV}_1^{n-2}$ instead of the incoming motion vector IV_1^n. However, MV_1^{n-1} and MV_1^{n-2} do not exist in the incoming bit stream since MB_1' and MB_1'' are not aligned with the block grid. Thus, the outgoing motion vector needs to be either reestimated using motion estimation schemes or composed using the incoming motion information of the macroblocks on the gird.

Motion vector reestimation is undesirable due to intensive computation. Instead, motion vector composing using incoming motion information is a better approach [15], [18]. Similar problem has also been discussed for video

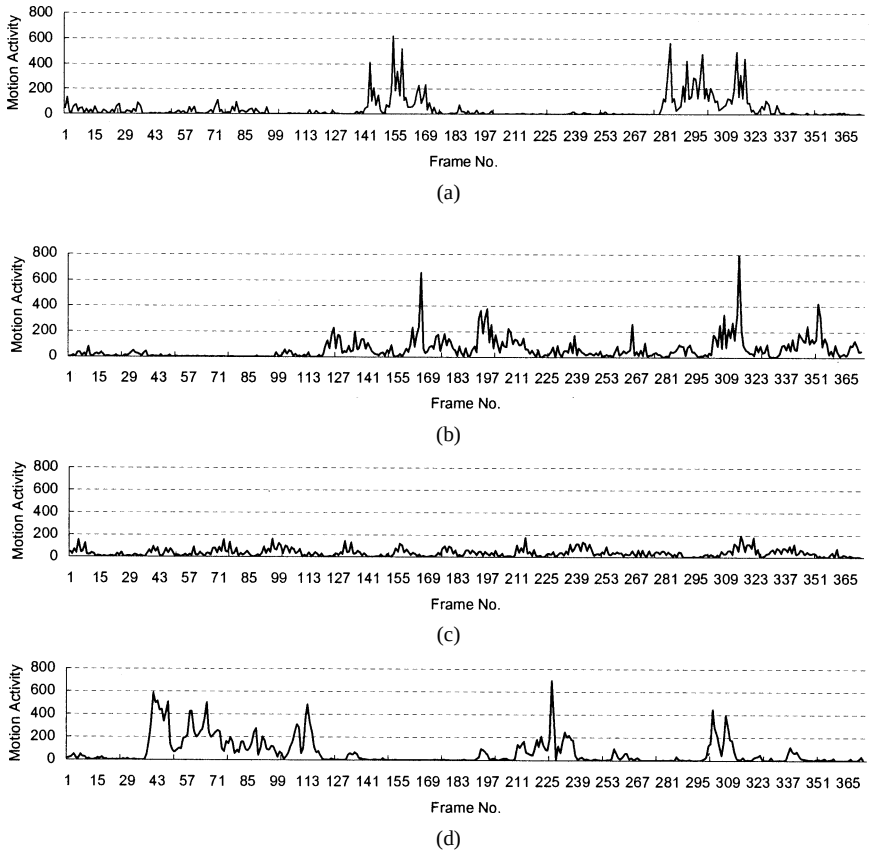


Fig. 6. Motion activity of each sub-window.

downscaling applications [19]–[21]. Composing the motion vector MV^{n-k} from the four neighboring motion vectors $\{\mathrm{IV}_1^{n-k}, \mathrm{IV}_2^{n-k}, \mathrm{IV}_3^{n-k}, \mathrm{IV}_4^{n-k}\}$ is to find a mapping function $\widehat{\mathrm{MV}}^{n-k} = f(\mathrm{IV}_1^{n-1}, \mathrm{IV}_2^{n-1}, \mathrm{IV}_3^{n-1}, \mathrm{IV}_4^{n-1})$, which can approximate MV^{n-k} with good accuracy, such as an interpolation function as follows [15], [20], [21]:

$$\widehat{\mathrm{MV}}^{n-k} = \frac{\sum_{i=1}^{4} \mathrm{IV}_i^{n-k} \mathrm{A}_i \mathrm{ACT}_i}{\sum_{i=1}^{4} \mathrm{A}_i \mathrm{ACT}_i}. \tag{3}$$

where A_i and ACT_i represents the corresponding overlapping area with and the activity of the ith neighboring residual block, respectively. In (3), the residual block activities ACT_i's can be computed in the pixel or DCT domains. Since the number of the DCT coefficients of a block is usually small, it would be much more efficient to compute ACT_i in the DCT domain. We propose to compute $\mathrm{ACT}i$ as follows:

$$\mathrm{ACT}_i = \sum_{j \notin \mathrm{DC}} |\mathrm{Coef}_{i,j}| \tag{4}$$

where $\mathrm{Coef}_{i,j}$ is the jth nonzero DCT AC coefficient of the ith neighboring block, which is decoded and de-quantized from the incoming bit stream.

As explained in [18], however, the interpolation scheme may not produce a good estimate of a motion vector due to the interpolation of diverse motion flows. In addition, it requires much extra memory to store all the motion vectors of the dropped sub-windows in the backward interpolation process. To improve the results, a forward dominant vector selection (FDVS) scheme was proposed in [18] for composing the unavailable motion vectors in the dropped frames with good accuracy and low computation/memory cost. The FDVS method selects one dominant motion vector that is carried by the neighboring block, which overlaps the target block the most, from the four neighboring motion vectors as the motion vector of the target block, as illustrated in Fig. 4.

The performance of the FDVS method can be further improved with a little extra computational complexity. If there is no strongly dominant block that overlaps the reference block with a significantly large area (e.g., the overlapping area is larger than a predefined threshold, say, 80% of the block area), selecting a weakly dominant vector that diverges largely from the other neighboring motion vectors may degrade the quality significantly since the motion vector chosen may be unreliable. To solve this problem, we propose to remove the unreliable motion vectors from the candidate list before selecting the dominant motion vector if no strongly dominant block is found. Furthermore, in the dominant vector selection, the "largest overlapping area" may not be the best criterion when the overlapping areas of some of the other neighboring anchor blocks are similar. In this case, we propose to select the neighboring block with the largest overlapping energy/activity as the dominant MB and use the activity measure as defined in (4).

The proposed pre-filtered activity-based forward dominant vector selection (PA-FDVS) scheme is summarized as follows.

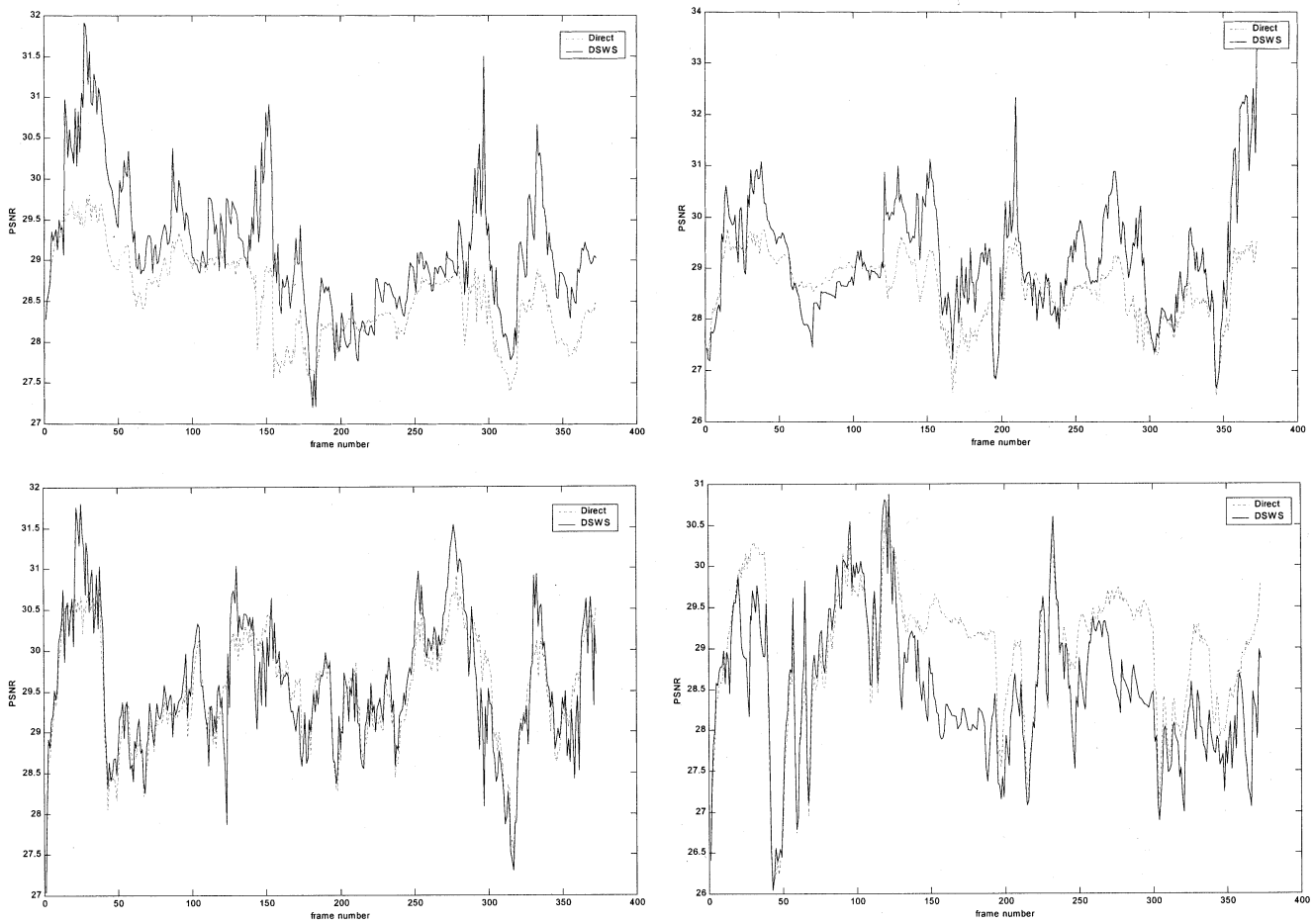


Fig. 7. PSNR comparison of the proposed method and TMN8.

Step 1) For each block, calculate the largest overlapping area, and if the largest overlapping area is greater than a predetermined threshold (e.g., 80% in our simulation), then select the motion vector of the neighboring block with the largest overlapping area as the dominant vector and process to the next block, otherwise go to step 2).

Step 2) Perform the following motion vector pre-filtering procedure:

Set the initial candidate list as the four neighboring vectors $\{\mathrm{IV}_1, \mathrm{IV}_2, \mathrm{IV}_3, \mathrm{IV}_4\}$
Calculate the mean and the standard deviation of the four neighboring motion vectors as follows:

$$\mathrm{IV}_{\mathrm{mean}} = \frac{1}{4}\sum_{i=1}^{4} \mathrm{IV}_i$$

$$\mathrm{IV}_{\mathrm{std}} = \sqrt{\frac{1}{4}\sum_{i=1}^{4} (\mathrm{IV}_i - \mathrm{IV}_{\mathrm{mean}})^2}$$

for $i = 1$ *to* 4
if $|\mathrm{IV}_i - \mathrm{IV}_{\mathrm{mean}}| > k_{\mathrm{std}} \cdot \mathrm{IV}_{\mathrm{std}}$
IV_i *is unreliable, remove it from the dominant vector candidate list*
else
IV_i *is reliable, keep it from the dominant vector candidate list*

Step 3) Calculate the area-activity products $\mathrm{A}_i \cdot \mathrm{ACT}_i$ for the blocks with the motion vector in the dominant vector candidate list, where A_i is the overlapping area with the ith neighboring residual block and ACT_i is the activity measure, as defined in (4). Then select the motion vector of the neighboring block with the largest area-activity product as the dominant vector.

IV. Experimental Results

In our experiments, four 200-frame standard QCIF test sequences "Foreman" and "Carphone," "Football" and "Akiyo" with a frame rate of 30 f/s are used to verify the performance of the proposed PA-FDVS scheme. The performance comparisons of the full-search motion estimation method and the motion-vector composition methods (interpolation, FDVS, and PA-FDVS methods) using the four test sequences are shown in Table I and Fig. 5. Table I shows the average peak signal-to-noise ratio (PSNR) comparisons for the two test

TABLE II
AVERAGE PSNR COMPARISON OF THE PROPOSED DSWS + TMN8 AND TMN8 TRANSCODING SCHEMES

	Skipped frame No.	Average PSNR of all frames (dB)			Average PSNR of non-skipped frames (dB)		
		TMN8	DSWS+TMN8		TMN8	DSWS+TMN8	
Sub-window 1	151	28.54	29.14	+0.60	28.55	29.31	+0.76
Sub- window 2	75	28.59	29.16	+0.57	28.52	29.25	+0.73
Sub- window 3	54	29.54	29.56	+0.02	29.48	29.59	+0.11
Sub- window 4	139	28.99	28.59	-0.40	28.73	28.68	-0.05
Average	104.75	28.91	29.11	+0.20	28.82	29.21	+0.39

sequences that were first encoded with 128 kb/s and 30 f/s and then transcoded with 56 kb/s and 7.5 f/s. The result in Table I indicates that PA-FDVS performs better than FDVS and significantly outperforms the interpolation scheme. The positive and negative numbers on right-hand side of the average PSNR value of PA-FDVS in Table I indicate the maximal coding gain and maximal degradation of PA-FDVS in comparison to FDVS. The frame-by-frame PSNR comparison of the PA-FDVS and FDVS schemes with the same test condition used in Table I are shown in Fig. 5. Although the average PSNR values of PA-FDVS and FDVS in Table I are close, Fig. 5 suggests that the PA-FDVS scheme achieves significant PSNR improvement (up to 1.6, 2.9, 1.7, and 2.2 dB for the four sequences, respectively) over the FDVS scheme on several frames with many diverse object motions.

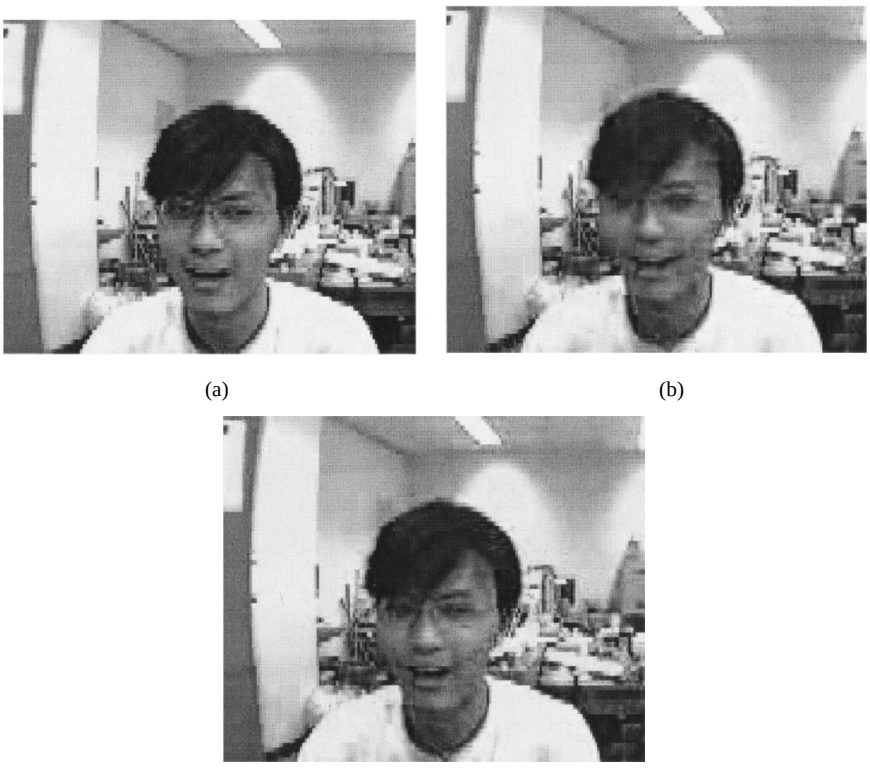
To verify the effectiveness of the proposed DSWS scheme, four 400-frame QCIF video sequences captured from a four-point video conference with a frame rate of 15 f/s are used for experiments. We firstly encoded the four QCIF video sequences with 128 kb/s and 15 f/s using the pubic-domain H.263 TMN8 software [22]. The four input bit streams are then jointly transcoded into a single CIF video with an output bit rate of 128 kb/s and an output frame rate of 15 f/s using the proposed DSWS scheme. Thus, the compression ratio performed by the transcoder is four in our experiments.

Fig. 6 depicts the motion activity of each sub-window. In the simulated video conference session, most of the time, only one or two sub-windows are motion active. Fig. 7 compares the frame-by-frame PSNR performance of the proposed and direct transcoding schemes. In the direct transcoding scheme, the bits are distributed into MBs using the ITU-T TMN8 rate control scheme [17], while in our proposed method, the DSWS algorithm is first used to determine the sub-windows to skip, and the TMN8 scheme is then adopted for MB bit allocation in the nonskipped sub-windows. Note that each skipped sub-window will only consumes 99 COD bits. The PSNR of a skipped sub-window is computed from the incoming QCIF sub-window image and the latest previously reconstructed nonskipped one since the sub-window repetitions will occur for the skipped sub-windows at the video decoders. The thresholds $\mathrm{TH}_{\mathrm{MV}}$ and $\mathrm{TH}_{\mathrm{MAAD}}$ are empirically set at 0.2 and 10, respectively. Fig. 7 illustrates that the proposed $\mathrm{DSWS} + \mathrm{TMN8}$ method achieves PSNR gain on the sub-windows with relatively high activities, while the low-activity sub-windows are degraded. Table II shows the comparison of the average PSNR of all the sub-windows using the two methods. As shown in Fig. 7 and Table II, the proposed DSWS scheme achieves 0.2- and 0.39-dB average PSNR improvements on the overall and nonskipped sub-windows, respectively. Figs. 8 and 9 compare the subjective quality of the whole frame and sub-window with maximal improvement using the proposed scheme. In Fig. 8, significant improvement on the visual quality of sub-window 1 (upper left) and sub-window 2 (upper right), the most active ones, can be observed while keeping comparable quality in other sub-windows. The improvement is more obvious in the face area of the maximally improved sub-window at frame 27. In sub-window 4, the average PSNR performance is degraded by 0.4 dB because of many long intervals with relatively low motion activity. The degradation is caused by the temporal resolution reduction by repeating the previously decoded sub-window for the period of sub-window skipping; the temporal resolution reduction is visually insignificant since the motions of these sub-windows are very small, as illustrated. Fig. 10 compares the subjective quality of the maximally

(a)

(b)

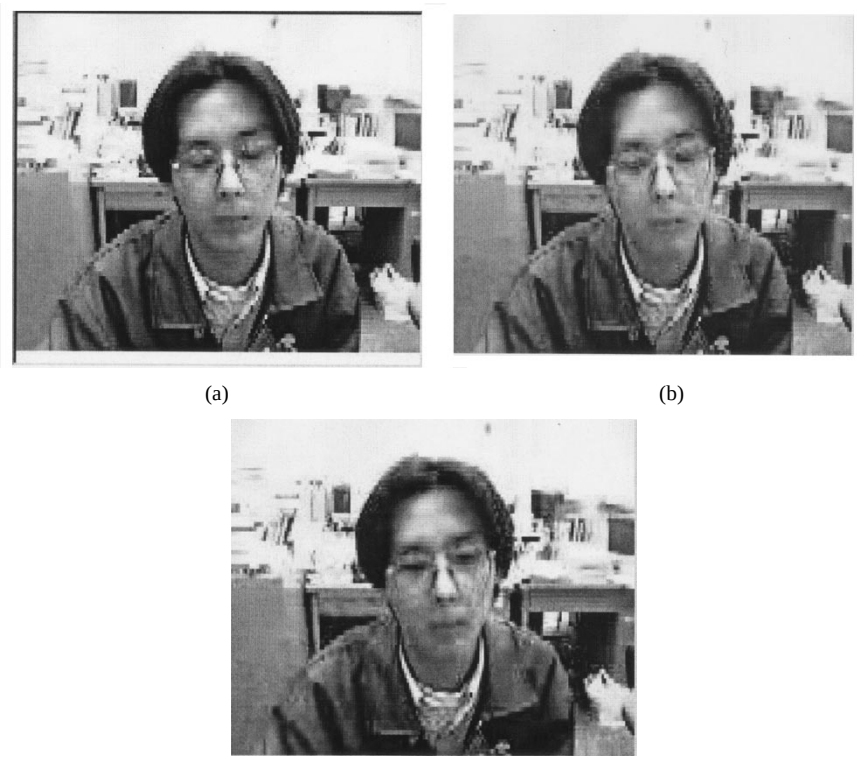
Fig. 8. Maximal improvement on whole frame (at frame #269). (a) Transcoding with TMN8 (29.25 dB). (b) Transcoding with $\mathrm{DSWS} + \mathrm{TMN8}$ (29.95 dB).



(a) (b)

(c)

Fig. 9. Maximal improvement of sub-window 1 (at frame #27). (a) Incoming video. (b) Transcoding with TMN8 (29.45 dB). (c) Transcoding with DSWS + TMN8 (31.90 dB).



(a) (b)

(c)

Fig. 10. Maximal degradation of sub-window 4 (at frame #170). (a) Incoming video. (b) Transcoding with TMN8 (29.4 dB). (c) Transcoding with DSWS + TMN8 (28.1 dB).

degraded sub-window. We can observe that, although with a 1.3-dB PSNR drop, the degradation still does not look serious. In our experiment, 418 out of 1600 sub-windows are skipped, thus achieving about 17% computation reduction since the

computation required for decoding a sub-window, skipping decision and motion vector composing, is only about 1/3 of the computation for transcoding a sub-window. The amount of computation reduction depends on the two threshold values $\mathrm{TH_{MV}}$ and $\mathrm{TH_{MAAD}}$. The higher the threshold values, the lower the computation demand; however, the lower the video quality. It is thus possible to achieve better tradeoffs between computational cost and video quality by adjusting the threshold values adaptively.

V. Conclusions

In this paper, we have proposed a DSWS scheme for multipoint video conferencing. The proposed scheme can enhance the visual quality of the active sub-windows by saving bits from skipping the inactive ones without introducing significant quality degradation. We also presented an efficient motion-vector composition scheme to compose and trace the motion vectors in the skipped sub-windows. Note, this motion-vector composition method can also be used in other transcoding applications involving frame-rate reduction.

The proposed method is particularly useful in multipoint video-conferencing applications since the focuses in such applications are mainly on the active conferees. Simulation results verify the effectiveness of the proposed method. In addition to the significant computation reduction due to sub-window skipping, the proposed method can also achieve both objective and subjective visual quality improvement. Furthermore, the proposed algorithm is fully compatible with the H.263 standard and, thus, can be integrated into current commercial products. The proposed method can also be further extended to enhance the quality of specific regions/objects in region-/object-based coding standards such as MPEG-4 [6].

Acknowledgment

The authors would like to thank the anonymous reviewers for their helpful comments that helped to improve the quality of this paper.

References

[1] *Video Codec for Audiovisual Services at $p \times 64$ kbits/s*, ITU-T Draft Recommendation H.261, Mar. 1993.
[2] *Video Codec for Low Bit-Rate Communication*, ITU-T Recommendation H.263, May 1997.
[3] *Coding of Moving Pictures and Associated Audio for Digital Storage Media at Up to About 1.5 Mbits/s*, ISO/IEC 11 172, Oct. 1993.
[4] *Coding of Moving Pictures and Associated Audio*, ISO/IEC 13 818, Nov. 1995.
[5] *Coding of Moving Pictures and Associated Audio MPEG98/W2194 (MPEG-4)*, ISO/IEC JTC1/SC29/WG11, Mar. 1998.
[6] M.-T. Sun and I.-M. Pao, "Multipoint video conferencing," in *Visual Commun. and Image Processing*. New York: Marcel Dekker, 1997.
[7] G. Keesman *et al.*, "Transcoding of MPEG bitstream," *Signal Process. Image Commun.*, pp. 481–500, 1996.
[8] P. A. A. Assuncao and M. Ghanbari, "A frequency domain video transcoder for dynamic bit rate reduction of MPEG-2 bit streams," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, pp. 953–567, Dec. 1998.
[9] M.-T. Sun, T.-D. Wu, and J.-N. Hwang, "Dynamic bit allocation in video combing for multipoint video conferencing," *IEEE Trans. Circuits Syst.*, vol. 45, pp. 644–648, May 1998.
[10] T.-D. Wu and J.-N. Hwang, "Dynamic bit rate conversion in multipoint video transcoding," in *Proc. IEEE Int. Conf. Image Processing*, Kobe, Japan, Oct. 1999, pp. 817–821.
[11] G. Keesman, "Multi-program video compression using joint bit-rate control," *Philips J. Res.*, vol. 50, pp. 21–45, 1996.
[12] L. Wang and A. Vincent, "Joint rate control for multi-program video coding," *IEEE Trans. Consumer Electron.*, vol. 42, pp. 300–305, Aug. 1996.
[13] ——, "Bit allocation and constraints for joint coding of multiple video programs," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, pp. 949–959, Sept. 1999.
[14] 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.
[15] J.-N. Hwang, T.-D. Wu, and C.-W. Lin, "Dynamic frame skipping in video transcoding," in *Proc. IEEE Multimedia Signal Processing Workshop*, Redondo Beach, CA, Dec. 1998, pp. 616–621.
[16] C.-W. Lin, T.-J. Liao, and Y.-C. Chen, "Dynamic rate control in multipoint video transcoding," in *Proc. IEEE Int. Symp. Circuits Syst.*, Geneva, Switzerland, May 2000, pp. II-17–II-20.
[17] *Video Codec Test Model, TMN8*, ITU-T/SG16, June 1997.
[18] J. Youn, M.-T. Sun, and C.-W. Lin, "Adaptive motion vector refinement for high performance transcoding," *IEEE Trans. Multimedia*, vol. 1, pp. 30–40, Mar. 1999.
[19] *Standard Specifications for the Implementations of 8×8 Inverse Discrete Cosine Transform*, IEEE Standard 1180-1990, 1990.
[20] H. Sun, W. Kwok, and J. W. Zdepski, "Architectures for MPEG compressed bitstream scaling," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, pp. 191–199, Apr. 1996.
[21] J. Youn, J. Xin, and M.-T. Sun, "Fast video transcoding architectures for networked multimedia," in *Proc. IEEE Int. Circuits Syst. Symp.*, Geneva, Switzerland, May 2000, pp. 25–28.
[22] N. Bjorkand and C. Christopoulos, "Transcoder architecture for video coding," *IEEE Trans. Consumer Electron.*, vol. 44, pp. 88–98, Feb. 1998.
[23] B. Shen, I. K. Sethi, and V. Bhaskaran, "Adaptive motion-vector resampling for compressed video downscaling," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, pp. 929–936, Sept. 1999.
[24] M. R. Hashemi, L. Winger, and S. Panchanathan, "Compressed domain vector resampling for downscaling of MPEG video," in *Proc. IEEE Int. Conf. Image Processing*, Kobe, Japan, Oct. 1999, pp. 276–279.
[25] "H.263 + encoder/decoder," Image Processing Lab., Univ. British Columbia, Victoria, BC, Canada, TMN (H.263) codec, Feb. 1998.
[26] W. Zhu, K. Yang, and M. Beacken, "CIF-to-QCIF video bitstream down-conversion in the DCT domain," *Bell Lab. Tech. J.*, vol. 3, no. 3, pp. 21–29, July–Sept. 1998.
[27] A. Vetro, H. Sun, and Y. Wang, "Object-based transcoding for adaptable video content delivery," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, pp. 387–402, Mar. 2001.

Chia-Wen Lin (S'94–M'00) received the M.S. and Ph.D. degrees in electrical engineering from the National Tsing Hua University, Hsinchu, Taiwan, R.O.C., in 1992 and 2000, respectively.

In August 2000, he joined the Department of Computer Science and Information Engineering, National Chung Cheng University, Taiwan, R.O.C., where he is currently an Assistant Professor. Prior to joining the National Chung Cheng University, he was a Section Manager with the Customer Premise Equipment (CPE) and Access Technologies Department, Computer and Communications Research Laboratories, Industrial Technology Research Institute (CCL/ITRI), Taiwan, R.O.C. From April 2000 to August 2000, he was a Visiting Research Scholar with the Information Processing Laboratory, Department of Electrical Engineering, University of Washington. From July 2002 to August 2002, he was a Visiting Professor with Microsoft Research Asia, Beijing, China. He has authored or coauthored over 40 technical papers. He holds eight patents with more pending. His research interests include video coding and networked multimedia technologies.

Dr. Lin was the recipient of the 2000 Research Achievement Award presented by the ITRI. He was also the recipient of the 2000 and 2001 Best Ph.D. Thesis Awards presented by the Acer Foundation and the Ministry of Education, R.O.C., respectively.

Yung-Chang Chen (M'85–SM'90) received the B.S. and M.S. degrees in electrical engineering from the National Taiwan University, Taipei, Taiwan, R.O.C., in 1968 and 1970, respectively, and the Ph.D. (Dr.-Ing.) degree from the Technische Universität Berlin, Berlin, Germany, in 1978.

In 1978, he joined the Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan, R.O.C. From 1980 to 1983, he was Chair of the Department of Electrical Engineering, National Central University, Chungli, Taiwan, R.O.C. From 1992 to 1994, he was Chair of the Department of Electrical Engineering, National Tsing Hua University. He is currently Dean of the College of Engineering and a Professor with the Department of Computer Science and Information Engineering, National Chung Cheng University, Chiayi, Taiwan, R.O.C. He is also a Professor with the Department of Electrical Engineering, National Tsing Hua University. His current research interests include multimedia signal processing, digital video processing, medical imaging, computer vision, and pattern recognition.

Dr. Chen serves as chair of the IEEE Consumer Electronics Society, Taipei Chapter.

Ming-Ting Sun (S'79–M'81–SM'89–F'96) received the B.S. degree from the National Taiwan University, Taipei, Taiwan, R.O.C., in 1976, the M.S. degree from the University of Texas at Arlington, in 1981, and the Ph.D. degree from the University of California at Los Angeles (UCLA), in 1985, all in electrical engineering.

In August 1996, he joined the University of Washington, Seattle, where he is currently a Professor. Prior to joining the University of Washington, he was the Director of the Video Signal Processing Research Group at Bellcore. He has authored or coauthored over 140 technical papers, including ten book chapters in the area of video technology. He holds eight patents.

Dr. Sun was a conference general co-chair of SPIE Visual Communications and Image Processing 2000 (VCIP2000). He was the Editor-in-Chief of the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT) (1995–1997) and the (1999–2001). From 1988 to 1991, he was the chairman of the IEEE Circuits and Systems (CAS) Standards Committee. He established an IEEE IDCT standard. He was the recipient of the 1987 Award of Excellence presented by Bellcore for his work on the digital subscriber line. He was a corecipient of the 1993 IEEE TCSVT Best Paper Award. He was also the recipient of a 2000 Golden Jubilee Medal presented by the IEEE CAS Society.