

Motion-Compensated Layered Video Coding for Playback Scalability

Jae-Yong Lee, *Member, IEEE*, Hyo-Sub Oh, *Student Member, IEEE*, and Sung-Jea Ko, *Senior Member, IEEE*

Abstract—In this paper, we propose a multilayered video coding scheme based on motion estimation which enables a decoder to dynamically change its temporal and spatial resolution during playback. In the proposed scheme, a new motion-prediction structure with a temporal hierarchy of frames is adopted to afford temporal resolution scalability and the wavelet decomposition with a new intra-update algorithm is used to offer spatial scalability. Experimental results show that the proposed scheme exhibits a higher compression ratio than conditional replenishment schemes because it further reduces the temporal redundancy using motion estimation. Also the proposed intra-update technique enables adaptation to dynamic change of spatial resolution with the effective follow-up of prediction in the decoder while preventing the large peaks in bit rate. Therefore, the proposed scheme is expected to be effectively used in heterogeneous environments such as the Internet, ATM, and wireless networks where dynamic scalability and interoperability are required.

Index Terms—Adaptation, dispersive intra-subband update, intra-update, layered video, motion estimation, scalability.

I. INTRODUCTION

FUTURE digital video services demand interoperability among various applications. To fulfill different requirements by using one common bit stream in a wide range of video services in heterogeneous environments, video coding techniques are needed which can simultaneously support a variety of bit rates tailored to individual services. One approach toward this purpose is to represent a video information in a scalable manner which enables a receiver to select a part of the generated bit stream and decode it with the given resources.

There are three major kinds of scalability in case of video: spatial resolution, temporal resolution, and data-rate (SNR) scalabilities [1]. Spatial resolution scalability is functionality to decode images at different sizes, e.g., from size of thumbnail for fast browsing to that of HDTV. Temporal scalability means that the refresh rate of the frames can be adjusted and data-rate scalability implies that any target data rate can be achieved from a single compressed bitstream according to the user requirements based on the available network bandwidth or system capability.

A method to attain scalability is to split and distribute the video information over a number of *layers*. In layered coding scheme, the most fundamental information to reconstruct video frames constructs a layer referred to as *base layer*. The other layers, called *enhancement layers*, are for additive information which, when combined with the base or lower enhancement layers, produce more refined information. The more layers the decoder receives, the better quality it will achieve.

Motion estimation (ME) and motion compensation (MC) schemes which have been effectively utilized for conventional video coding are known difficult to be incorporated in scalable video coding techniques based on the layered representation [2]–[5], since the reference frame (decoded previous frame) at the decoder may not be the one that has been used for the prediction at the encoder when some enhancement layers are dropped due to either network congestion or end-system capacity limit. If the decoder uses a different prediction to the one assumed by the encoder, a form of mismatch referred to as *drift* may occur.

Temporally subband coding techniques which are free from prediction loop could be used to prevent this problem. 2-D scalable video coding techniques based on subband decomposition can be extended to 3-D scalable coding [4]–[8] which not only provides spatial resolution scalability, but also affords temporal scalability in nature. For rate scalability, Shapiro's embedded zerotree wavelet (EZW) [9] and many improved variants or alternatives such as set partitioning in hierarchical trees (SPIHT) [10] can be extended to three dimensional case [8], [11]. But temporally subband coding has a drawback that it blurs moving objects in temporally low resolution pictures, which is caused by the averaging effect of the low-pass filter performed in temporal decomposition. Motion-compensated 3-D subband video coding techniques have been proposed to alleviate this problem and also to improve compression efficiency by concentrating signal energy in the temporal low subband [12], [13]. Still, because these techniques operate on several consecutive frames, they require a large frame memory in temporal subband decomposition. As the target number of temporal layers increases, the number of required frames is exponentially increasing and consequently an unavoidable excessive coding/decoding delay is introduced. A promising approach is, therefore, to extend the classical hybrid coding principle to fulfill the demands of scalabilities.

For SNR scalability, motion prediction has been utilized only for the base layer information, which is assumed to be always delivered to the receiver [14], [15]. Recently, MPEG-4 standard adopted an SNR scalable coding tool, the fine-granular-scalability (FGS) framework, which is also based on this concept [16], [17]. The FGS scheme is expected to be effectively used

Manuscript received November 12, 1999; revised August 10, 2000. This paper was recommended by Associate Editor A. Luthra.

J.-Y. Lee is with the R&D Center, Serome Technology Inc., Seoul 137-064, Korea (e-mail: jlee@ieee.org).

H.-S. Oh is with the Next Generation Lab, LG Electronics Inc., Anyang 431-749, Korea (e-mail: fred999@ieee.org).

S.-J. Ko is with the Department of Electronics Engineering, Korea University, Seoul 136-701, Korea (e-mail: sjko@ieee.org).

Publisher Item Identifier S 1051-8215(01)03825-3.

for video streaming service both in unicast and multicast environments.

A sophisticated technique which provides temporal and spatial scalabilities as well as SNR scalability based on conditional replenishment (CR) was proposed for layered video coding and transmission over best-effort networks [18]. In CR approaches, only those blocks of a picture that have changed significantly between frames beyond a certain threshold are selected, coded, and transmitted [19]. This scheme, however, cannot achieve such a high compression performance as motion compensated prediction can.

In this paper, we propose a layered video coding scheme which affords temporal and spatial resolution scalabilities. A new motion-prediction structure with a temporal hierarchy of frames is adopted to afford temporal resolution scalability. It can have a higher compression ratio than replenishment schemes since motion estimation further reduces the temporal redundancy. For spatial scalability, the proposed scheme uses wavelet decomposition with multiresolution motion estimation (MRME), where motion vectors of a spatial layer are refined based on those of lower layers [20]. Since only small differential motion vectors in each pyramid level are stored or transmitted with prediction error information separately for the corresponding layer, we can achieve a layer-wise prediction. The prediction error signal is encoded by the SPIHT algorithm which is an embedded coding technique to enable precise target rate control. In addition, a new effective intra update scheme which we will refer to as *dispersive intra-subband update* (DISU) is proposed to admit dynamic change of spatial resolution at a decoder and also to enable recovery from errors. The proposed video coding algorithm provides a multilayered bitstream for heterogeneous environments and the spatial resolution and frame rate can be independently adjusted according to various user requirements.

This paper is organized as follows. In Section II, we propose a novel multilayer video coding technique to provide temporal and spatial scalabilities. Section III describes the intra update method DISU, which is an efficient resynchronization method to recover from reference mismatches in encoders and decoders. Experimental results are given and discussed in Section IV. Finally, Section V summarizes our work and concludes the paper.

II. MOTION-COMPENSATED LAYERED VIDEO CODING

Unfortunately, international video coding standards such as MPEG-1 and H.261 currently used in wide areas do not provide layered representations. Although MPEG-2 and H.263 do support scalability and layered representation in their extensions, they are not able to produce an arbitrary number of layers [21]. For example, the MPEG-2 scheme may allow only two discrete levels of scalability for each scalable scheme. Even in the case of combining different scalability tools into a hybrid coding scheme, its syntax supports only up to three different layers. In this section, we propose a multilayered video coding method based on motion estimation and compensation.

A. Motion Estimation over Temporal Layering

To provide temporal scalability, the subsampled frames can be arranged so that any set of cumulative layers produces frames

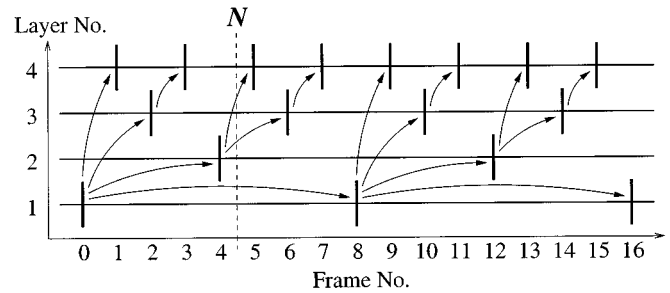


Fig. 1. Motion estimation over temporal layering.

spaced evenly over time [18]. If the total number of temporal layers is equal to $N_T = M + 1$ and we let $L_M(n)$ denote the temporal layer number where the frame number $n \geq 0$ is delivered, then

$$L_M(n) = M - r(n \bmod 2^M + 2^M) + 1 \quad (1)$$

with

$$r(m) = \min\{k > 0 : \lfloor \frac{m}{2^k} \rfloor 2^k \neq m\} - 1 \quad (2)$$

where $r(m)$ is the bit position (numbered from 0) of the rightmost nonzero bit in the binary representation of m . Temporal layering with four layers ($M = 3$) is illustrated in Fig. 1.

In this temporal layering scheme, however, motion vectors cannot be estimated directly from the previous frame as in conventional predictive coding techniques because the previous frame might be on any higher layer that is not utilized by a decoder. Therefore, we propose a new prediction method over this temporal layering. As the reference frame for the frame n , we select the closest one in time among frames on layer $L_M(n)$ or on the lower layers as depicted in Fig. 1. That is, the reference frame number for the frame n is given by

$$R_M(n) = n - 2^{M-w(n)} \quad (3)$$

where

$$w(n) = \min\{k \geq 0 : n \bmod 2^{M-k} = 0\}. \quad (4)$$

By performing motion estimation and compensation across different layers, as shown in Fig. 1, different temporal resolution can be simultaneously achieved without affecting the performance no matter how many layers a decoder exploits.

Distribution of frames over the multiple layers in temporal hierarchy leads to storage of multiple references. At the time N in Fig. 1, for example, we can see that two frames should be stored as references; frame 4 for prediction of frames 5 and 6, and frame 0 for prediction of frame 8. The required number of frame memories in the buffer for motion estimation is $\lceil (M + 1)/2 \rceil$, where $\lceil \cdot \rceil$ denotes rounding off to the nearest integer. Note that, because prediction from the reference frame may result in a large motion due to the varying temporal interval between the predicted frame and its reference, the search range should be adjusted proportionally to the interval.

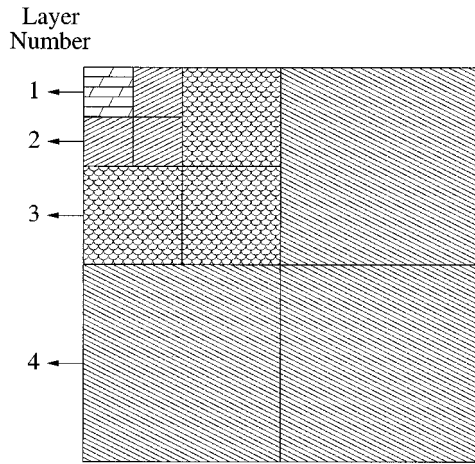


Fig. 2. Spatial layering with four-level wavelet pyramid.

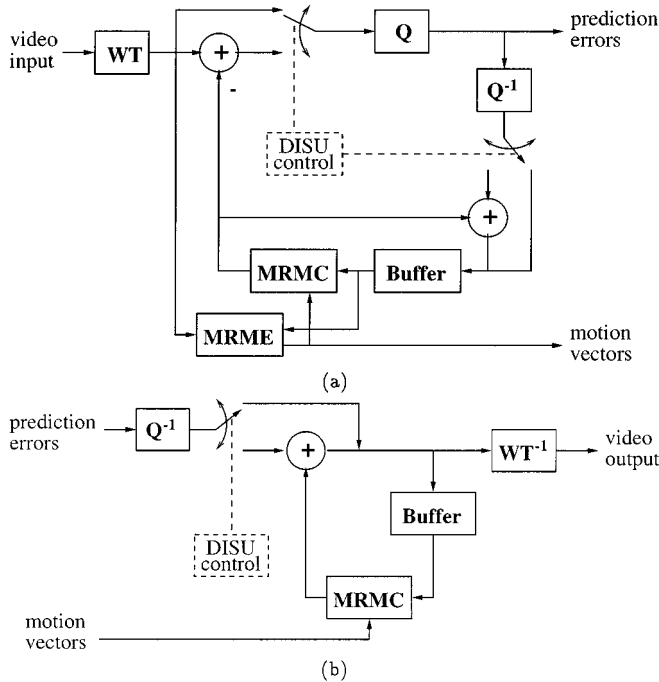


Fig. 3. Proposed layered video coder: (a) encoder and (b) decoder.

B. Spatial Layering with Wavelet Decomposition

By combining the image pyramid decomposition technique which can generate multiresolution images with the aforementioned encoding scheme, spatial scalability also can be provided. Multiple spatial layers generated from a pyramid structure characterize the signal in different spatial resolutions. In this paper, the wavelet image decomposition scheme is adopted, but other decomposition techniques can be easily employed in a similar way. Fig. 2 illustrates a four-level wavelet pyramid. Each level of the image pyramid corresponds to a spatial layer with different resolutions.

The proposed encoder and decoder employing MRME are depicted in Fig. 3. A video frame is first decomposed into a wavelet pyramid, and motion estimation and compensation is performed in wavelet domain. Motion activities at different levels of the wavelet pyramid are different but highly correlated

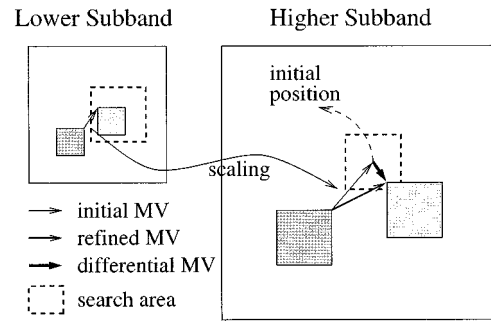


Fig. 4. Multiresolution motion estimation. Only two associated subbands are shown.

since they actually characterize the same motion structure at different scales and different frequency ranges. Therefore, the proposed coding algorithm incorporates MRME where motion vectors for higher spatial resolutions are refined using the motion information obtained at lower resolutions, as illustrated in Fig. 4 [20]. The search range in the higher subband can be kept small because the motion vectors found in the lower subband give good initial positions where the search process can start. The MRME scheme is expected to reduce the searching time significantly and to provide a smooth motion vector field. Motion-compensated error frames are quantized and may be further compressed exploiting an entropy coder. In our work, SPIHT is incorporated since it is superior in performance to other still image compression techniques [17], [22]. Also, thanks to the embedded characteristic of SPIHT, it is easy to control the overall rate of the generated bitstream. If we assume that the target data rate is R and the bits required to encode the motion vector fields for a frame have data rate $R_{MV} < R$ which is known to the encoder, the precise target rate can be achieved by encoding each prediction error image at the rate $R - R_{MV}$.

Existing multiresolution video coding techniques with MRME assume that the spatial resolution is first determined at the decoder and fixed throughout the video playback. Consequently, it is impossible to keep track of the right reference during video playback with dynamic change of the spatial resolution. However, as interactivity comes up as an important functionality of future video services, dynamic change of spatial size of video playback may be requested by users. For that purpose, switches are inserted in our encoder and decoder architectures (see Fig. 3) for bypassing the prediction loop and generating intra-coded signal. By periodically switching to this *intra* mode with an appropriate interval, both recovery from errors and tracking of correct prediction state can be achieved. The details of this intra-update algorithm will be described in the next section.

III. DISPERSIVE INTRA-SUBBAND UPDATE

In order to perform dynamic adaptation in decoding according to available resources, such as available network bandwidth and end-system capacity, a decoder may add or drop spatial layers. When spatial enhancement layers are appended to increase the quality of video playback, no reference information is available on the appended layer, and the video

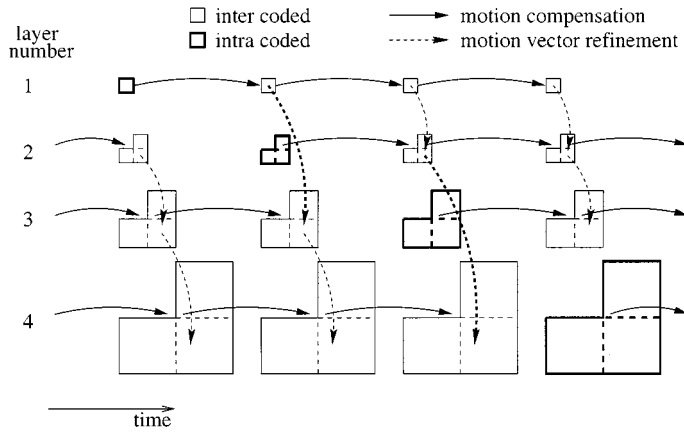


Fig. 5. One cycle of dispersive intra-subband update with 3-level wavelet decomposition (only the base temporal layer is shown).

quality cannot increase as high as expected. Moreover, when a decoder receives information corrupted by channel error, a general predictive coding algorithm may suffer from loss propagation. In the proposed temporal layering scheme with motion compensation, however, the loss propagation is confined within a prediction hierarchy which is rooted on the base layer. A prediction hierarchy is composed of 2^M frames and the effects of all possible losses are propagated only within this hierarchy and never affect the following new prediction hierarchy. In Fig. 1, for example, an error which occurs at frame 6 on the third temporal layer can affect only the next frame, frame 7, on the fourth temporal layer without degrading the other subsequent frames. This temporal layering should be used in conjunction with a priority-drop policy so that all unavoidable loss in congestion may occur only on the highest temporal layer that a decoder receives and thus the error propagation can be managed within as small duration as possible.

To further reduce this effect on the base temporal layer, as well as to catch up with the potentially best quality under given circumstances, intraframes like I-frames of the MPEG scheme may be inserted periodically. The drawback in this case, however, is that the intraframes may cause very large peaks in bit rate. In transmission, these peaks may lead to congestion in network and cause oscillations in the level of layers the decoder receives. To solve this problem, the DISU scheme is proposed, in which only one level of subbands in the wavelet pyramid instead of a whole frame is intra-coded on the base temporal layer at a time and then the other levels in turn as illustrated in Fig. 5. The horizontal arrows in Fig. 5 represent motion compensation. Since motion estimation is not performed in the intra-coded subbands, the refinement information of motion vectors for the higher spatial layer in MRME cannot be generated in these subbands. The refinement of motion vectors on the higher spatial layer is therefore achieved by utilizing the estimated motion vectors on the next lower spatial layer as the bold vertical arrows imply in Fig. 5. When a decoder newly adds a spatial-enhancement layer, the quality of double-sized video is somewhat lower than the expected level for the first some frames since only the differential information without the initial reference is available on the new spatial layer. As it is intra-updated and the reference information finally becomes complete, the quality reaches

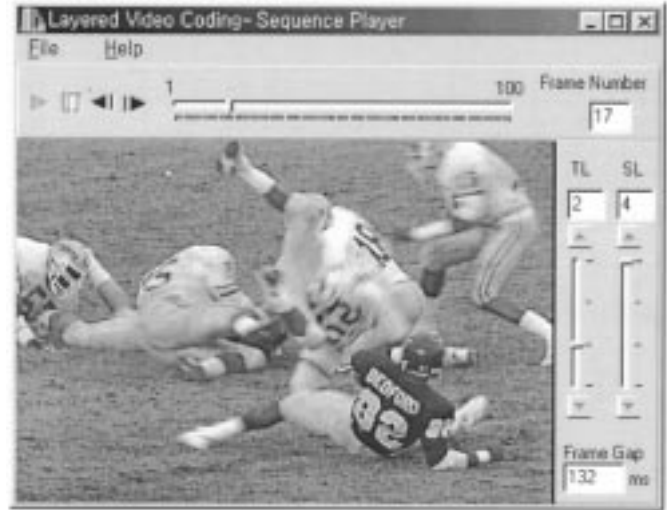


Fig. 6. Implemented video sequence player (*lvplayer*).

to the expected level. The time demanded for this follow-up is generally not so long and can be tolerable in real video communication. The quantitative investigation will follow.

There are two possible methods of DISU, depending on the frequency of update on each spatial layer, as follows.

- 1) *Simple Update*: Each spatial layer is intra-updated sequentially on the base temporal layer as shown in Fig. 5. After all the spatial layers are updated, a new cycle of intra-update is repeated. When the number of temporal layers and spatial layers are N_T and N_S respectively, the update period on all spatial layers is given by

$$P = 2^{N_T-1} \times N_S. \quad (5)$$

In case of four temporal layers ($N_T = 4$) and four spatial layers ($N_S = 4$), for example, the intra-update period becomes $2^3 \times 4 = 32$ frames. This implies that the full recovery of video quality takes about one second in the worst case.

- 2) *Hierarchical Update*: In general, higher spatial layers tend to be dropped more frequently than lower ones. Hence, we can consider a method to perform more frequent intra-update on the higher spatial layers than lower layers. If we apply the similar hierarchy of temporal layering in (1) to spatial layers on the base temporal layer, the update period of the i -th spatial layer is given by

$$P_i = \begin{cases} 2^{N_T-1} \times 2^{N_S-1}, & \text{for } i = 1 \\ 2^{N_T-1} \times 2^{N_S-i+1}, & \text{for } i \geq 2. \end{cases} \quad (6)$$

When $N_T = 4$ and $N_S = 4$, the update periods of the spatial layers are $P_1 = P_2 = 64$, $P_3 = 32$, and $P_4 = 16$. Even though the average time elapsed until degradation of visual quality disappears after an error occurs is same as in the simple update method, the actual elapsed time for the update varies according to the given circumstances because the probability of utilization of each spatial layer is not identical (the probability decreases as the number of layer increases). For example, in the case of utilizing all of

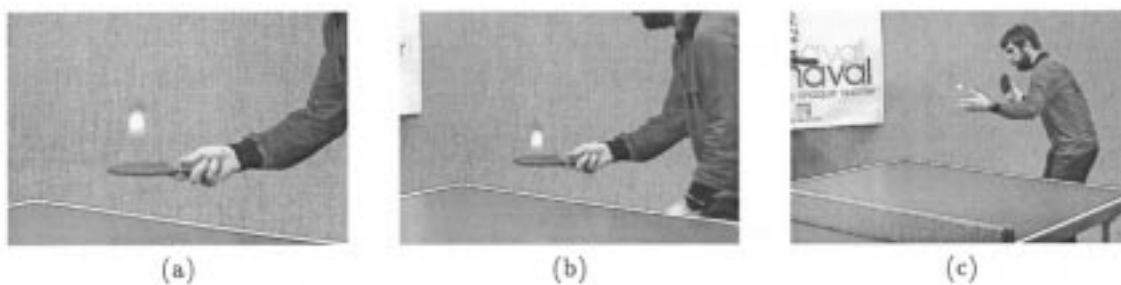


Fig. 7. Table Tennis sequence: (a) 10th; (b) 40th; and (c) 70th frame.



Fig. 8. Football sequence: (a) 10th; (b) 50th; and (c) 110th frame.



Fig. 9. Flower Garden sequence: (a) 1st; (b) 30; and (c) 70th frame.

the four spatial layers while one or two of those layers are added/dropped, the hierarchical update will work better than the simple update. Meanwhile, when only a small set of lower spatial layers are used in decoding, the hierarchical update is not expected to perform well since the period of intra update becomes too long.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. Implementations

For the experiment, we implemented the encoder and decoder depicted in Fig. 3. Decomposition was conducted three times using discrete wavelet transform (DWT) with the Daubechies 4-tap FIR filter bank [23], and hence four spatial layers were generated. After the decomposition, each subband was divided into blocks and MRME was carried out. We employed the block-matching algorithm for simplicity, although it does not necessarily result in the best rate-distortion tradeoffs [24]. The block size for motion estimation is 2×2 in the LL band which corresponds to the first spatial layer. Motion vectors estimated in the LL subband are refined in the higher subbands. The search range of ± 2 , both in the horizontal and vertical directions at all subbands, was utilized for the

TABLE I
AVERAGE NUMBER OF BYTES ASSIGNED TO MOTION VECTORS (FOOTBALL)

Temporal Layer	Spatial Layer				Total Bytes
	First	Second	Third	Fourth	
First	425.2	1380.1	1390.5	1397.6	4595.4
Second	349.3	1177.2	1188.1	1180.9	3897.5
Third	254.7	949.3	969.4	972.4	3148.2
Fourth	160.2	711.5	753.3	763.2	2390.3

brute force full-search method using mean absolute difference (MAD) criteria.

The motion vectors attained from the subbands were differentially coded, and then the QM coder which is an adaptive binary arithmetic coder utilized in JBIG standard was applied to these codes for entropy coding [25]. The residual errors of wavelet coefficients after motion compensation were then quantized and compressed using SPIHT. Because SPIHT exploits cross-band correlation in the wavelet domain, it is not easy to separate the generated bitstream into spatial layers. For our experiment, SPIHT was specially implemented by rearranging generated bits into different layers which correspond to different resolutions (for details, see Appendix).

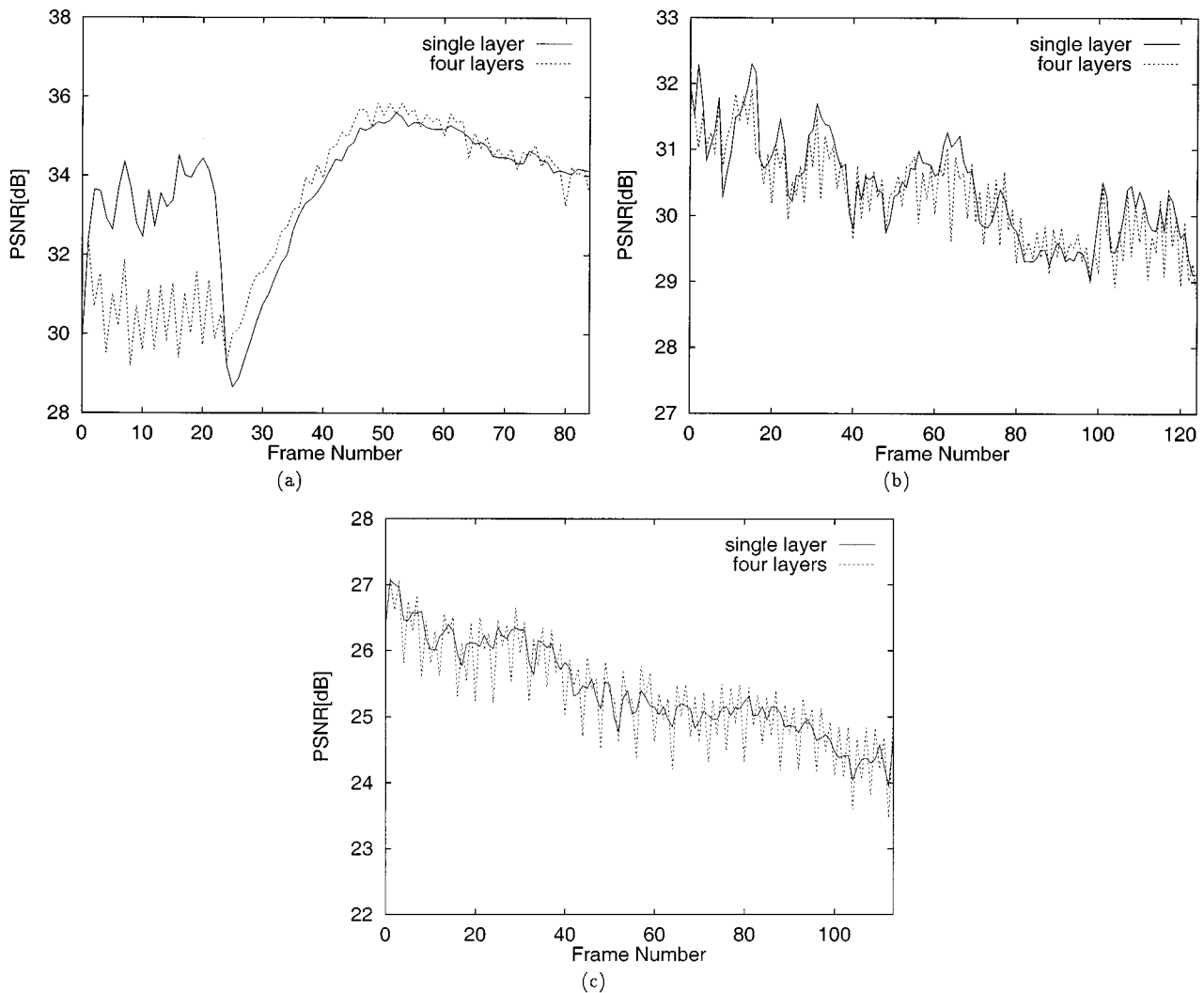


Fig. 10. Bit rate of each spatial layer: (a) Table Tennis; (b) Football; and (c) Flower Garden.

A video sequence player *lvplayer*, running on Windows with a user-friendly interface, was implemented for the observation of subjective visual quality in real video playout, especially with layer dropping. The *lvplayer* is shown in Fig. 6. One can select the number of temporal and spatial layers involved in decoding with this software. The refresh rate of video frames can also be controlled by adjusting the frame gap so that each frame can be carefully examined in playout.

B. Test Sequences

The proposed coding scheme was applied to 352×240 (SIF) grayscale version of three test sequences with 30 frames/s: *Table Tennis*, *Football*, and *Flower Garden*, which are composed of 85, 125, and 115 frames, respectively. Some selected frames from these sequences are given in Figs. 7–9. The first part of the Table Tennis test sequence shows only one hand holding a racket and bouncing a ball with slight motion as shown in Fig. 7(a), but the camera starts to zoom out around frame number 25 until a scene with a player and a part of the table appears. On the other hand, the Football sequence involves many football players with a lot of motion in the

TABLE II
AVERAGE PSNR OF VIDEO RECONSTRUCTED WITH A PART OF SPATIAL LAYERS (IN DECIBELS)

Test Sequence	Number of Used Spatial Layers			
	one	two	three	four (all)
Table Tennis	21.73	23.06	25.73	33.61
Football	19.75	21.70	24.94	30.72
Flower Garden	16.43	17.54	19.89	25.87

TABLE III
AVERAGE PERCENTAGE OF BITRATE ASSIGNED TO EACH SPATIAL LAYER

Test Sequence	Assigned Rate to Each Spatial Layer (%)			
	First	Second	Third	Fourth
Table Tennis	15.41	16.18	36.23	32.18
Football	14.18	17.91	38.31	29.60
Flower Garden	10.73	15.45	38.88	34.94

whole range of display as shown in Fig. 8. The Flower Garden sequence shown in Fig. 9 is a slow “drive-by” of a house and garden generated by camera panning.

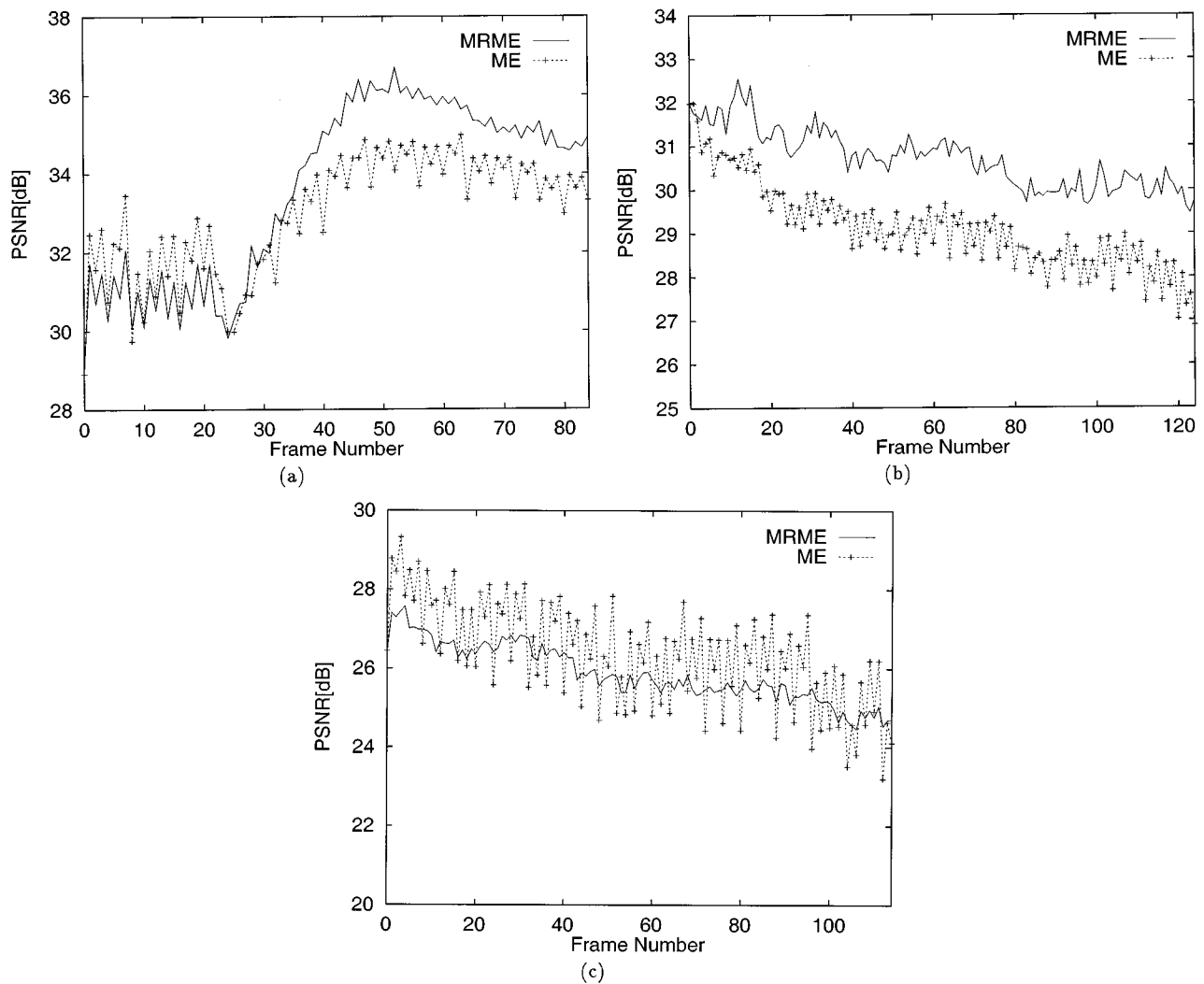


Fig. 11. ME versus MRME: (a) Table Tennis; (b) Football; and (c) Flower garden.

C. Performance Evaluation

First, the effects of the proposed temporal layering with motion compensation were investigated. Table I shows the average rate required to encode motion vector fields for the Football sequence on each spatial and temporal layer. As the distance between the current frame and the reference frame becomes larger, i.e., the temporal layer number decreases (see Fig. 1), motion vector values increase and consequently require more bits, as shown in Table I. Accordingly, the bit budget for prediction error decreases when the total bit rate is maintained constant and the decoded image quality becomes worse than in the case with a single temporal layer. This quality degradation is significant in those regions that call for a lot of bit budget, for example, due to a rough texture. This can be clearly seen from the big performance difference at the first part of the Table Tennis sequence [see Fig. 10(a)], which involves a static rough background over a wide range. Fig. 10 shows the results associated with the proposed video coder with a single temporal layer and also with four layers. The total bit rate was targeted at 3.0 Mbits/s for this experiment. The PSNR plots for the four-layer case exhibits periodically spiky behavior due to the periodic variations of the intervals between the predicted frames and their references.

As a decoder appends one more spatial layer, the spatial resolution doubles. Even if the decoder utilizes only a subset of spatial layers, it may reconstruct a video in full resolution by substituting 0's for wavelet coefficients in the other subbands. In this manner, we can attain the SNR scalability effect. When video frames are decoded in full resolution (352×240), the average reconstructed visual quality in respect of PSNR is shown in Table II. Table III shows the average percentage of the bit rate of each spatial layer.

In conventional ME methods, the motion-compensated error image is converted into its DCT or other transformed version. On the other hand, motion is estimated in wavelet transformed domain in MRME. One question arises if zerotree coding such as SPIHT is well combined with the MRME scheme, since separate motion estimation performed on each subband might affect the cross-band correlation of the prediction error image and degrade the coding efficiency of zero-tree coding. To see the effects of MRME on SPIHT coding efficiency, we fixed the bit rate of prediction-error frames independent of the number of bits assigned for motion vectors. Fig. 11 shows PSNR versus frame number when each prediction error frame is encoded at 1.0 bpp. It is seen in Fig. 11(a) and (b) that MRME even performs better than ordinary ME with SPIHT. This

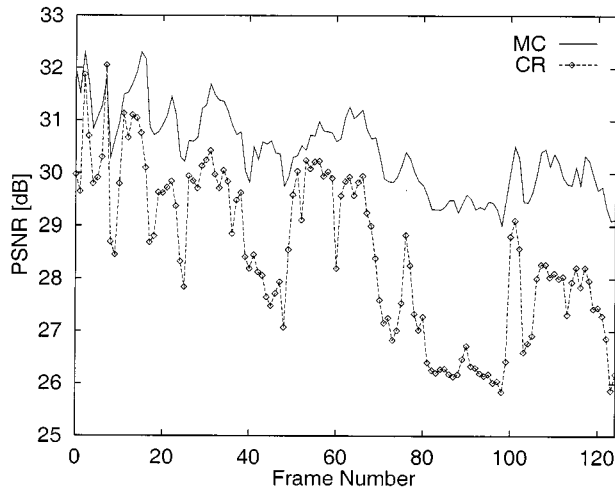


Fig. 12. Comparison of compression efficiency of a CR scheme and the proposed technique (Football).

indicates that the cross-band correlation of the residual image is maintained after MRME since motion activities at different subbands actually characterize the same motion structure. Also, due to the smoother hierarchical motion vector fields at different frequency ranges, more uniform and superior visual quality can be achieved, as seen in Fig. 11.

A video coder using the block-based CR technique was implemented using SPIHT and its compression performance was compared to that of the proposed coding scheme. The bit rate was again fixed at 3.0 Mbits/s for both of the schemes. For the CR scheme, the mean squared difference between the reference block and the current block is computed and compared with an empirical threshold value to decide whether or not to encode the current 16×16 block of a video frame. Fig. 12 gives the comparison of the CR coding and the proposed MC-based coding. It is seen that the MC-based coding technique outperforms the CR technique by 0.29–3.42 dB for the Football sequence. All these experimental results indicate that the proposed MC-based video coding scheme can produce a multilayer bitstream with much higher compression ratio than the CR scheme while maintaining a comparable performance with the conventional motion-estimation technique.

Finally, the proposed DISU scheme was evaluated. As can be observed in Fig. 13, periodic I-frame insertion for resynchronization of the decoder yields deep notches in visual quality when encoded in constant bit rate, i.e., the same number of bits are assigned to each frame. Conversely, it would produce large peaks in bit rate if a constant visual quality were to be maintained. Note that the proposed DISU algorithm effectively moderates the burstiness in decoded video quality (or bit rate). Although the average time for resynchronization increases in this case, it can be managed to be within a duration short enough to not much affect the perceived quality in motion pictures.

APPENDIX SPATIAL LAYERING WITH SPIHT

The entire SPIHT encoding algorithm is given in Fig. 14. The reader is referred to [10] for definition of the notations used here.

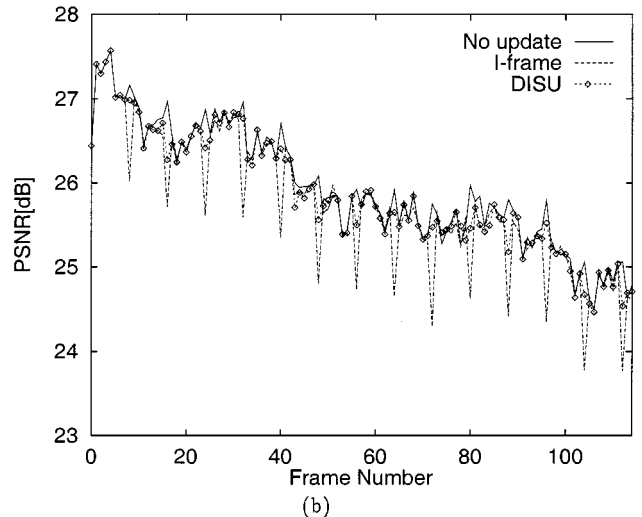
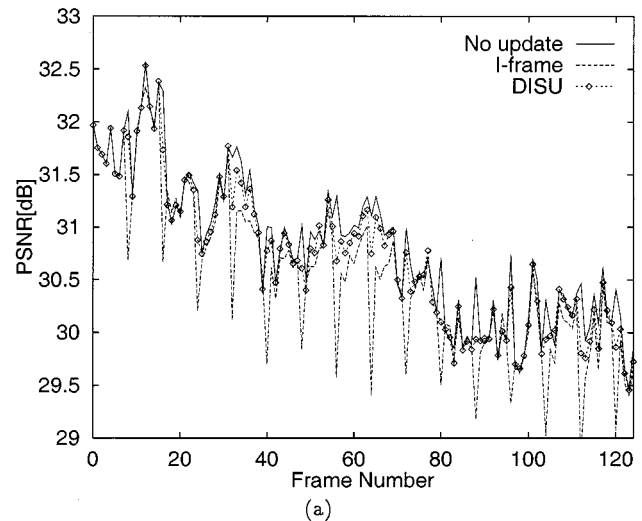


Fig. 13. Comparison of intra-update schemes. (a) Football. (b) Flower Garden.

The basic concept of spatial layering is to partition and reorganize all the information in the generated bit stream, which includes sorting bits as well as sign bits and refinement bits. This is done by sending out the generated bits at the boxed places in Fig. 14 into different spatial layers each of which corresponds to a different resolution. As a result, the bitstream becomes rearranged as illustrated in Fig. 15. In this manner, we can yield a layered representation of an image while not breaking the decoding order. In decoding, the list of insignificant sets (LIS), the list of insignificant pixels (LIP), and the list of significant pixels (LSP) are maintained only with the bits corresponding to target spatial resolution.

V. CONCLUSION

In this paper, we proposed a new multilayered video coding technique to provide temporal and spatial scalabilities with an arbitrary number of layers. This scheme incorporates multiresolution motion compensation into temporal layering. A video frame is represented in a pyramid structure using wavelet decomposition, and MRME is performed in the wavelet domain over a temporal hierarchy of frames. The prediction error is encoded by the SPIHT coder to achieve a precise target bit rate

Algorithm:

- 1) Initialization: output $n = \lfloor \log_2(\max_{(i,j)}\{|c_{i,j}|\}) \rfloor$; set the LSP as an empty list, and add the coordinates $(i, j) \in \mathcal{H}$ to the LIP, and only those with descendants also to the LIS, as type A entries.
- 2) Sorting pass:
 - 2.1) for each entry (i, j) in the LIP do:
 - 2.1.1) $\boxed{\text{output } S_n(i, j)}$;
 - 2.1.2) if $S_n(i, j) = 1$ then move (i, j) to the LSP and $\boxed{\text{output the sign of } c_{i,j}}$;
 - 2.2) for each entry (i, j) in the LIS do:
 - 2.2.1) if the entry is of type A then
 - $\boxed{\text{output } S_n(\mathcal{D}(i, j))}$;
 - if $S_n(\mathcal{D}(i, j)) = 1$ then
 - * for each $(k, l) \in \mathcal{O}(i, j)$ do:
 - $\boxed{\text{output } S_n(k, l)}$;
 - if $S_n(k, l) = 1$ then add (k, l) to the LSP and $\boxed{\text{output the sign of } c_{k,l}}$;
 - if $S_n(k, l) = 0$ then add (k, l) to the end of the LIP;
 - * if $\mathcal{L}(i, j) \neq \phi$ then move (i, j) to the end of the LIS, as an entry of type B, $\boxed{\text{output '1'}}$, and go to Step 2.2.2; otherwise, remove entry (i, j) from the LIS and $\boxed{\text{output '0'}}$;
 - 2.2.2) if the entry is of type B then
 - $\boxed{\text{output } S_n(\mathcal{L}(i, j))}$;
 - if $S_n(\mathcal{L}(i, j)) = 1$ then
 - * add each $(k, l) \in \mathcal{O}(i, j)$ to the end of the LIS as an entry of type A;
 - * remove (i, j) from the LIS.
 - 3) Refinement Pass: for each entry (i, j) in the LSP, except those included in the last sorting pass (i.e., with same n), $\boxed{\text{output } n\text{-th most significant bit of } |c_{i,j}|}$;
 - 4) Quantization-Step Update: decrement n by 1 and go to Step 2

FIG. 14. SPIHT algorithm marked for spatial layering.

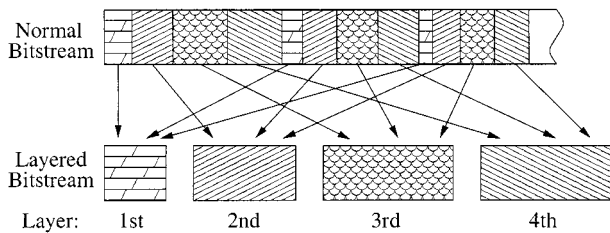


Fig. 15. Spatial layering by rearrangement of the bits generated by SPIHT.

from each frame and high compression. We also proposed the DISU scheme to admit dynamic change of spatial resolution at a decoder, and also to enable recovery from errors.

Experimental results showed that the proposed video coding scheme provides scalability in temporal and spatial resolution and outperforms the CR techniques, both in reconstruction quality and compression efficiency, while maintaining the overall quality comparable to a conventional single-layer motion predictive coding. Also, it was shown that the spatial resolution can be dynamically changed during playback by using the DISU scheme, which enables an effective follow-up of prediction in the decoder. Dynamic change and adaptation of processing load during playback are crucial features required today in many applications, e.g., streaming video on the Internet. Thus, the proposed scheme is expected to be successfully used for various multimedia applications in heterogeneous environments such as the Internet, ATM, and wireless networks where interoperability and scalability are needed.

REFERENCES

- [1] T. Sikora, "MPEG digital video-coding standards," *IEEE Signal Processing Mag.*, vol. 14, pp. 82–100, Sept. 1997.
- [2] J. Woods, *Subband Image Coding*. Norwell, MA: Kluwer, 1991.
- [3] K. Rose and S. Regunathan, "Toward optimal scalability in predictive video coding," in *Proc. IEEE Int. Conf. Image Processing '98*, vol. III, Chicago, IL, Oct. 1998, pp. 929–933.
- [4] K. Uz, M. Vetterli, and D. LeGall, "Interpolative multiresolution coding of advanced television with compatible subchannels," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 1, pp. 86–99, Mar. 1991.
- [5] D. Taubman and A. Zakhor, "Multirate 3-D subband coding of video," *IEEE Trans. Image Processing*, vol. 3, pp. 572–588, Sept. 1994.
- [6] W. Tan, E. Chang, and A. Zakhor, "Real time software implementation of scalable video codec," in *Proc. IEEE Int. Conf. Image Processing '96*, vol. I, Lausanne, Switzerland, Sep 1996, pp. 17–20.
- [7] U. Horn and B. Girod, "Scalable video transmission for the Internet," *Comput. Network ISDN Syst.*, vol. 29, no. 15, pp. 1833–1842, Nov 1997.
- [8] J. Tham, S. Ranganath, and A. Kassim, "Highly scalable wavelet-based video codec for very low bit-rate environment," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 12–27, Jan. 1998.
- [9] J. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. Signal Processing*, vol. 41, pp. 3445–3462, Dec. 1993.
- [10] A. Said and W. Pearlman, "A new, fast, and efficient image codec based on set partitioning in hierarchical trees," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, pp. 243–250, June 1996.
- [11] B.-J. Kim, Z. Xiong, and W. Pearlman, "Low bit-rate scalable video coding with 3-D set partitioning in hierarchical trees (3-D SPIHT)," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, pp. 1374–1387, Dec. 2000.
- [12] J. Ohm, "Three-dimensional subband coding with motion compensation," *IEEE Trans. Image Processing*, vol. 3, pp. 559–571, Sept. 1994.
- [13] S.-J. Choi and J. Woods, "Motion-compensated 3-D subband coding of video," *IEEE Trans. Image Processing*, vol. 8, pp. 155–167, Feb 1999.
- [14] M. Ghanbari, "Two-layer coding of video signals for VBR networks," *IEEE J. Select. Areas Commun.*, vol. 7, pp. 771–781, June 1989.
- [15] K. Shen and E. Delp, "Wavelet based rate scalable video compression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, pp. 109–122, Feb. 1999.

- [16] H. Radha and Y. Chen, "Fine-granular-scalable video for packet networks," in *Proc. Packet Video '99*, New York, Apr. 1999, pp. 69–72.
- [17] H. Radha, Y. Chen, K. Parthasarathy, and R. Cohen, "Scalable Internet video using MPEG-4," *Signal Processing: Image Commun.*, vol. 15, no. 1-2, pp. 95–126, Sep 1999.
- [18] S. McCanne, M. Vetterli, and V. Jacobson, "Low-complexity video coding for receiver-driven layered multicast," *IEEE J. Select. Areas Commun.*, vol. 15, pp. 983–1001, Aug. 1997.
- [19] F. Mounts, "A video encoding system with conditional picture-element replenishment," *Bell. Syst. Tech. J.*, vol. 48, no. 7, pp. 2545–2554, Sep 1969.
- [20] S. Zafar, Y. Zhang, and B. Jabbari, "Multiscale video representation using multiresolution motion compensation and wavelet decomposition," *IEEE J. Select. Areas Commun.*, vol. 11, pp. 24–35, Jan. 1993.
- [21] B. Haskell, P. Howard, Y. LeCun, A. Puri, J. Ostermann, M. Civanlar, L. Rabiner, L. Bottou, and P. Haffner, "Image and video coding—emerging standards and beyond," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8, pp. 814–837, Nov. 1998.
- [22] Z. Xiong, K. Ramchandran, M. Orchard, and Y.-Q. Zhang, "A comparative study of DCT and wavelet based coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, pp. 692–695, Aug. 1999.
- [23] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using wavelet transform," *IEEE Trans. Image Processing*, vol. 1, pp. 205–218, Apr. 1992.
- [24] F. Kossentini, W. Chung, and M. Smith, "Rate-distortion-constrained subband video coding," *IEEE Trans. Image Processing*, vol. 8, pp. 145–154, Feb 1999.
- [25] W. Pennebaker and J. Mitchell, *JPEG Still Image Data Compression Standard*. New York: Van Nostrand, 1992.



Jae-Yong Lee (S'98–M'00) received B.S., M.S., and Ph.D degrees, all in electronics engineering, from Korea University, Seoul, Korea, in 1993, 1995 and 1999, respectively.

He was a Research Assistant from 1996 to 1997 and a Researcher from 1997 to 1999 at the Research Institute of Information and Communication Technology, Korea University. During the summer of 1997, he was with the Institute of Industrial Process and System Communications, Technical University of Darmstadt, Darmstadt, Germany, supported by Deutscher Akademischer Austauschdienst (DAAD)

and Korea Science and Engineering Foundation (KOSEF). In September 1999, he joined Hyundai Electronics Industries Inc., Ltd., Seoul, Korea, and worked as a Member of Technical Staff. He has been with the R Center of Serome Technology Inc., Seoul, Korea, since September 2000. His current research interests are in the areas of video coding and transmission over wired and wireless networks.

Dr. Lee is a member of the IEEE Circuits and Systems Society and the IEEE Communications Society.



Hyo-Sub Oh (S'99) received the B.S. and M.S. degrees in electronic engineering from Korea University, Seoul, Korea, in 1999 and 2001, respectively.

In 1999, he joined the Research Institute for Information and Communication Technology, Korea University. He is presently an Assistant Research Engineer with Next Generation Terminal Laboratory, LG Electronics Inc, Anyang, Korea. His current interests include MPEG-4, multimedia messaging service, and visual communication over UMTS.



Sung-Jea Ko (M'88–SM'97) received the B.S. degree in electronic engineering from Korea University, Seoul, Korea, in 1980, and the M.S. and Ph.D. degrees in 1988 and 1986, respectively, both in electrical and computer engineering, from the State University of New York at Buffalo.

In 1992, he joined the Department of Electronic Engineering, Korea University, where he is currently a Professor. From 1988 to 1992, he was an Assistant Professor of Electrical and Computer Engineering at the University of Michigan-Dearborn. From 1981 to

1983, he was with Daewoo Telecom Corporation, where he was involved in research and development on data communication systems. His current research interests are in the areas of digital signal and image processing, and multimedia communications.

Dr. Ko was Associate Editor for the *Journal of Communications and Networks (JCN)* from 1998 to 2000, and is presently Chair of the IEEE Consumer Electronics Chapter in Korea. He received the Hae-Dong Best Paper Award from the Institute of Electronics Engineers of Korea in 1997, a Best Paper Award from the IEEE Asia Pacific Conference on Circuits and Systems in 1996, and the LG Academic Award for Outstanding Research in Information and Communication from LG Electronics Inc. in 1999. He is a Fellow of the IEEE.