Journal of Zhejiang University SCIENCE A
ISSN 1009-3095 (Print); ISSN 1862-1775 (Online)
www.zju.edu.cn/jzus; www.springerlink.com
E-mail: jzus@zju.edu.cn

Multiple description scalable video coding based on 3D lifted wavelet transform*

JIANG Gang-yi[1,2], YU Mei[†1,2], YU Zhou[1], YE Xi-en[1,2], ZHANG Wen-qin[1], KIM Yong-deak[3]

([1]*Faculty of Information Science and Engineering, Ningbo University, Ningbo 315211, China*)
([2]*State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China*)
([3]*Division of Electronics Engineering, Ajou University, Suwon 442-749, Korea*)
[†]E-mail: yumei2@126.com

Received Dec. 8, 2005; revision accepted Feb. 14, 2006

Abstract: In this work, a new method to deal with the unconnected pixels in motion compensated temporal filtering (MCTF) is presented, which is designed to improve the performance of 3D lifted wavelet coding. Furthermore, multiple description scalable coding (MDSC) is investigated, and novel MDSC schemes based on 3D wavelet coding are proposed, using the lifting implementation of temporal filtering. The proposed MDSC schemes can avoid the mismatch problem in multiple description video coding, and have high scalability and robustness of video transmission. Experimental results showed that the proposed schemes are feasible and adequately effective.

Key words: Multiple description scalable coding (MDSC), Motion compensated temporal filtering (MCTF), Block-split bidirectional motion estimation, 3D lifted wavelet transform
doi:10.1631/jzus.2006.A0857 **Document code:** A **CLC number:** TN919.8

INTRODUCTION

With the expansion of multimedia applications, video transmission is becoming an important issue in the research area of communication. It is very important to compress a large amount of video data effectively, which has already attracted considerable attention recently. To make source coders both error-resilient and network-adaptive are two main challenges facing the video compression techniques. It is necessary to make the bitstreams temporal, spatial and SNR scalable. A single bitstream can be decoded at different bitrates to meet different constraints on transmission bandwidth and decoding complexity in a heterogeneous networking environment. Moreover, the bitstreams are expected to have superior transmission robustness over unreliable networks.

Multiple description coding (MDC) has recently emerged as an attractive framework for robust transmission over unreliable channels (Yu *et al.*, 2005). A multiple description coder divides the original bitstream into two or more correlated bitstreams (descriptions), which are then separately transmitted over networks. Each description can be decoded independently so that loss of some of the descriptions will not jeopardize the decoding of correctly received descriptions. The fidelity of the received message is further improved as the number of received descriptions increases. MDC provides a solution to reduce the degradation of video signal resulted from packet losses, bit error and burst (erasure) error during transmission, and it guarantees the

* Project supported by the National Natural Science Foundation of China (No. 60472100), the Natural Science Foundation of Zhejiang Province (Nos. RC01057, Y105577, 601017), the Ningbo Science and Technology Project (Nos. 2003A61001, 2004A610001, 2004A630002), and the Zhejiang Science and Technology Project (No. 2004C31105), China

demand of real-time services. Many MDC schemes have been proposed (Wang *et al*., 2001; Reibman *et al*., 2002). However, a drawback of these techniques is that they are aimed solely at increasing the transmission robustness of video, and do not address other important transmission challenges, such as bandwidth variations, and receiving device characteristics. Over unreliable, heterogeneous and dynamic networks, MDC and scalable coding should complement each other to provide an efficient solution (ISO/IEC JTC1/SC29/WG11 N5540, 2003). One example of MDSC scheme was proposed (Bajic and Woods, 2003), and is referred to as domain-based MDC, which partitioned the wavelet coefficients into maximally separated sets, and packetized. However, due to such partition the conventional zero-tree structures cannot be exploited, and it is more difficult to interpolate the lost descriptions because the wavelet coefficients have less correlation compared to temporal decomposition. In addition, the spatiotemporal correlations are not taken into consideration in their scheme, and it is not efficient for video.

In this work, motion-compensated lifted 3D wavelet coding is studied, and a new scheme is proposed to deal with the unconnected pixels in MCTF, which improves the coding efficiency. Furthermore, we present a new approach to MDSC based on 3D lifted wavelet coding to enhance the transmission robustness of video over unreliable networks, and to provide scalable bitstreams to adapt heterogeneous networks.

3D LIFTED WAVELET CODING WITH A NEW MCTF SCHEME

3D wavelet coding has come into being recently as a promising alternative to hybrid DPCM video coding techniques. It provides high scalable bitstream for network and user adaptation, and resilience to transmission errors. The diagram of the 3D wavelet coding scheme adopted in this work is shown in Fig.1. The MCTF is implemented on the input video first, followed by spatial wavelet analysis on the temporal stage to complete the 3D wavelet decomposition. Then the spatio-temporal decomposition subbands and the motion vectors are encoded.

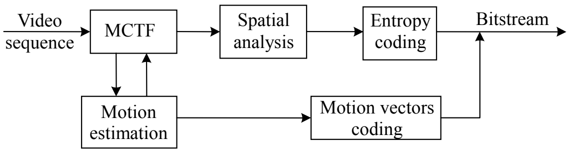


Fig.1 Diagram of 3D wavelet coding scheme

MCTF implemented through lifting is proved to be an effective and efficient temporal decomposition tool (Pesquet-Popescu and Bottreau, 2001; Secker and Taubman, 2003). In this work, we implement MCTF using a lifting scheme, which can achieve perfect reconstruction with sub-pixel accurate motion estimation. Fig.2 shows the implementation of Harr-based lifting MCTF. Frames *A* of the video sequence are displaced by the estimated value $\hat{d}_{2k,2k+1}$ to predict Frames *B*. The prediction step is followed by an update step with the displacement $-\hat{d}_{2k,2k+1}$.

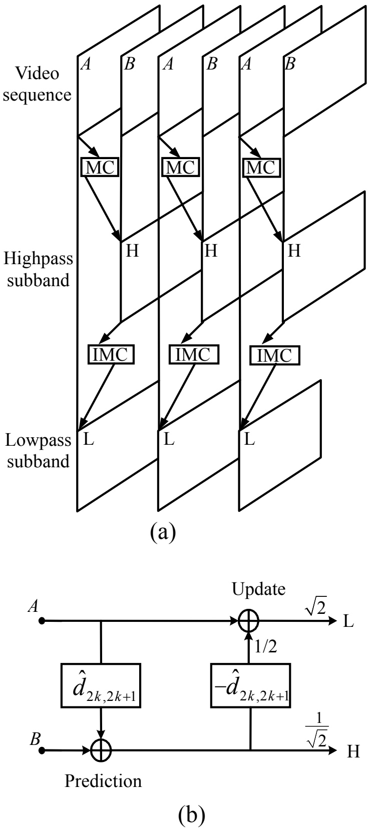


Fig.2 (a) Implementation of Harr-based lifting MCTF; (b) Harr lifting scheme

In MCTF, the block-based motion compensation with block-size of 16×16 is used, that is, video frames are divided into several blocks, and motion vectors of the blocks in the current frame point to the closest matching blocks in the reference frame. Fig.3 shows the state of connection of each pixel in the reference frame and in the current frame. If there is a one-to-one connection between the pixels, they are connected pixels. If several pixels in the current frame are connected to the same pixel in reference frame, only one of them is classified as a connected pixel, the others are listed as unconnected. Here, we do not use the scan-order rule to discriminate, but according to the absolute *DFD* value with each of them, regarding only the one with minimum *DFD* value as connected pixel. Conversely, some pixels in reference frames are not used as reference for current frames, which are regarded as unconnected pixels too. The unconnected pixels can detrimentally affect both overall coding efficiency and subjective video quality since they cannot be directly included in MCTF. In conventional MCTF methods, for the unconnected pixels, the corresponding temporal lowpass and highpass subbands are generated as follows:

$$\begin{aligned} &L[m,n]=2A[m,n]/\sqrt{2},\\ &H[m,n]=(B[m,n]-\tilde{A}[m-d_m,n-d_n])/\sqrt{2}, \end{aligned} \quad (1)$$

where $L[m,n]$ and $H[m,n]$ are temporal lowpass and highpass subbands, respectively; $A[m,n]$ is the reference frame, $B[m,n]$ is the current frame, and (d_m,d_n) is the motion vector; $\tilde{A}$ is the interpolated reference frame.

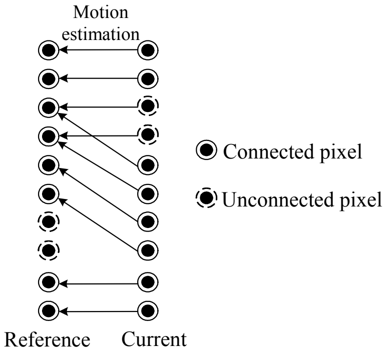


Fig.3 State of pixel's connection

A block-split bidirectional motion estimation (ME) scheme is proposed to deal with unconnected pixels, with the algorithm being as follows:

(1) Obtain the motion vectors between reference frame I_t and current frame I_{t+1} with block-based ME.

(2) Decide the unconnected pixels in the current frame I_{t+1}. If more than half of the pixels in a block of frame I_{t+1} are unconnected, we call this block an unconnected block.

(3) The block-size used in the block-based ME is 16×16. For more accurate motion estimation for the unconnected blocks, we split the 16×16 block into four 8×8 blocks.

(4) For each unconnected split 8×8 block, forward and backward ME are used to obtain the good matches, as shown in Fig.4.

(5) For the split blocks, if forward ME has the smallest DFD, the corresponding highpass subband is generated by

$$H[m,n]=(I_{t+1}[m,n]-\tilde{I}_t[m-d_m,n-d_n])/\sqrt{2}, \quad (2)$$

where (d_m, d_n) is the forward motion vector.

If backward ME has the smallest DFD, the corresponding highpass subband is generated by

$$H[m,n]=(I_{t+1}[m,n]-\tilde{I}_{t+2}[m-d_m,n-d_n])/\sqrt{2}, \quad (3)$$

where (d_m, d_n) is the backward motion vector.

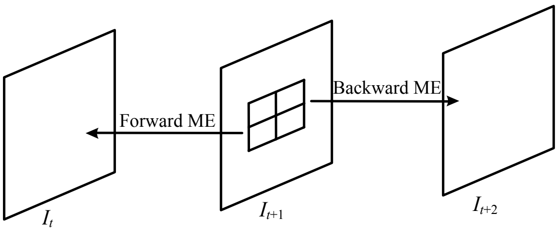


Fig.4 Block-split bidirectional motion estimation

PROPOSED MDSC SCHEMES

The proposed MDSC schemes are shown in Fig.5, the input video sequence is first decomposed into two descriptions, with the two descriptions being encoded with motion-compensated lifted 3D wavelet coding method respectively. 3D wavelet coding scheme has high compression efficiency while providing truly scalable bitstreams. In this work, two different multiple description decomposition schemes, based on spatial-domain and temporal-domain, are studied.

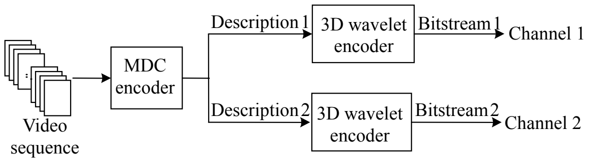


Fig.5 MDSC based on 3D wavelet coding

MDSC schemes based on spatial-domain

There is strong correlation between inter-pixels in raw image data, the value of any given pixel can be reasonably predicted by the value of its neighbors. Thus, it is possible to exploit this feature to create a multiple description coder. The video frames are split into even and odd descriptions by row-downsampling or column-downsampling in the spatial-domain. Here, we refer to row-downsampling method as spatial scheme-1, and column-downsampling method as spatial scheme-2. Then, each description is encoded with 3D wavelet coding, and the generated embedded bitstreams are transmitted to different channels. The texture characteristics of each video sequence are different, which results in different performances of the two spatial schemes.

The spatial-domain MDSC schemes utilize the natural correlations of pixels, which are stronger than those in the transform domain. And simple error concealment methods can yield good estimates of lost data. If the channels drop one description, linear filter error concealment is adopted to recover the lost description. For spatial scheme-1, the error concealment method is as follows:

$$\bar{F}_1(x,y)=\begin{cases} F_2(x,y), & \text{if } y=0,\\ \dfrac{F_2(x,y-1)+F_2(x,y)}{2}, & \text{otherwise,}\end{cases} \tag{4}$$

$$\bar{F}_2(x,y)=\begin{cases} F_1(x,y), & \text{if } y=\dfrac{row}{2}-1,\\ \dfrac{F_1(x,y)+F_1(x,y+1)}{2}, & \text{otherwise.}\end{cases} \tag{5}$$

For spatial scheme-2, the error concealment method is as follows:

$$\bar{F}_1(x,y)=\begin{cases} F_2(x,y), & \text{if } x=0,\\ \dfrac{F_2(x-1,y)+F_2(x,y)}{2}, & \text{otherwise,}\end{cases} \tag{6}$$

$$\bar{F}_2(x,y)=\begin{cases} F_1(x,y), & \text{if } x=\dfrac{col}{2}-1,\\ \dfrac{F_1(x,y)+F_1(x+1,y)}{2}, & \text{otherwise,}\end{cases} \tag{7}$$

where $F_1(x,y)$ and $F_2(x,y)$ are the frames for description one (even-row or even-column frames) and description two (odd-row or odd-column frames), $\bar{F}_1(x,y)$ and $\bar{F}_2(x,y)$ are the reconstructed frames for description one and description two, respectively.

MDSC scheme based on temporal-domain

According to the statistical characteristics of video, the neighboring frames have strong correlation because of the short temporal distance between them. Motion estimation and motion compensation can remove temporal redundancy effectively, and based on which temporal-downsampling MDC is adopted in this scheme. The video sequence is split into even and odd frames (descriptions) in temporal domain. And the two generated descriptions are encoded by 3D wavelet coding methods separately. Then each embedded bitstream is transmitted to different channels.

Error concealment methods should be used to recover the lost description at the receiver. Here, according to the temporal-domain MDSC scheme, bidirectional motion compensation interpolation (BMC-I) error concealment scheme is used, where the lost frame is estimated using both previous and future reference frames. BMC-I is more effective than single motion compensation interpolation (SMC-I) method, where the lost frame is estimated only using previous or future reference frames. BMC-I method used in this paper can be described as follows:

(1) Obtain the motion vectors of frame n+1, with reference frame n−1. As shown in Fig.6, the current block (x, y) has motion vector (Δx, Δy).

(2) Bidirectional motion estimation. For current frame n, the motion vector is ($\Delta x/2$, $\Delta y/2$) to frame n−1, while the motion vector is ($-\Delta x/2$, $-\Delta y/2$) to frame n+1, as shown in Fig.7.

(3) Bidirectional motion compensation,

$$\begin{aligned}\tilde{\psi}_n(x,y)&=a\psi_{n-1}(x+\Delta x/2,y+\Delta y/2)\\ &\quad+b\psi_{n+1}(x-\Delta x/2,y-\Delta y/2),\end{aligned} \tag{8}$$

where $\psi_{n-1}(x, y)$ and $\psi_{n+1}(x, y)$ are Frame $n-1$ and Frame $n+1$ respectively. $\tilde{\psi}_n(x,y)$ is the estimation frame with BMC-I scheme, a and b are two estimation coefficients, generally, both set to 0.5.

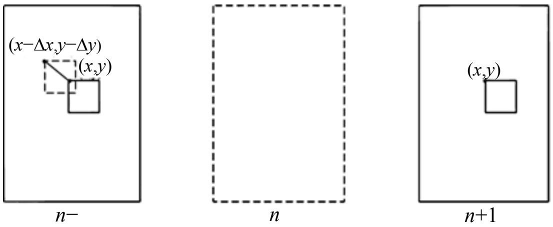


Fig.6 Motion estimation between frames

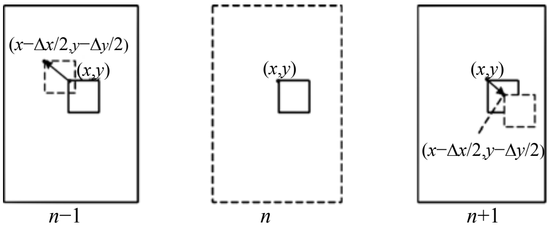


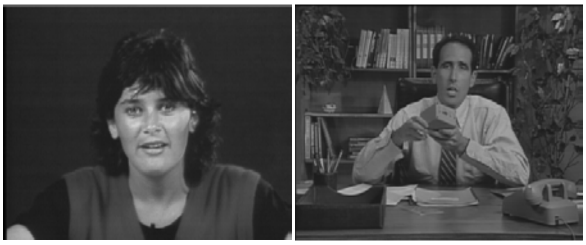
Fig.7 Bidirectional motion estimation for the current frame *n*

The two proposed MDSC schemes make use of the spatial-correlation and temporal-correlation adequately, and the corresponding error concealment methods have low complexity. With a 3D transform performed on each description, reconstruction of the reference frame is not necessary and the closed prediction loops do not exist. Hence, the mismatch issue in some multiple description video coders does not occur in the proposed schemes. And more consistent reconstructed video quality can be achieved at the receiver.

EXPERIMENTAL RESULTS

To evaluate the performance of the proposed coding scheme, ‘Missa’ (CIF, 30 fps) and ‘Salesman’ (CIF, 30 fps) video sequence in Fig.8 are adopted as test video sequences, 80 frames of each sequence are used in the experiments.

First, the proposed block-split bidirectional motion estimation scheme is tested. Block-based motion estimation was performed with half-pixel accuracy in MCTF, and temporal subbands have been spatially decomposed over 4 levels using biorthogonal 9/7 Daubechies filters. The resulting spatio-temporal wavelet coefficients are encoded using the EZBC algorithm in (Hsiang and Woods, 2001). Motion fields are encoded using lossless DPCM and adaptive arithmetic coding. The results on Missa and Salesman sequences are shown in Fig.9 with the conventional MCTF method and the proposed block-split MCTF method. Coding performance is expressed in terms of Y component $PSNR$, calculated by averaging the Y



(a) (b)

Fig.8 Test video sequence using (a) Missa and (b) Saleman

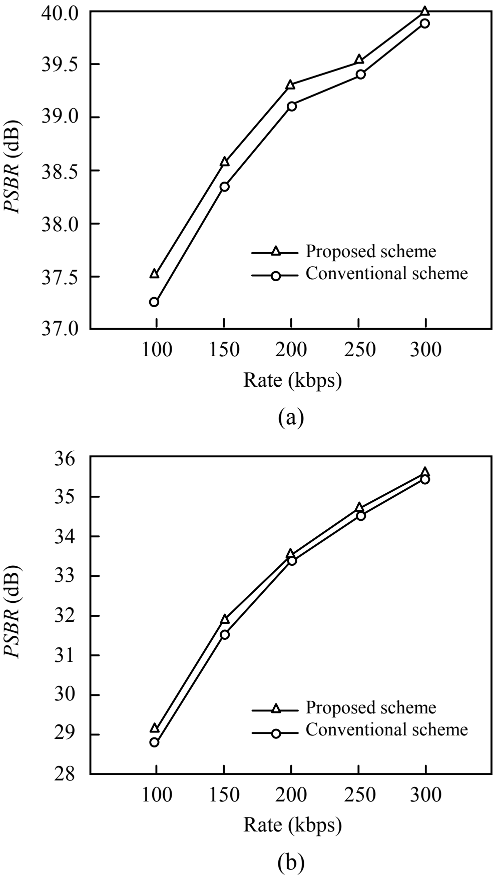


Fig.9 Rate-distortion performances for (a) Missa; (b) Salesman

component *PSNR* over all decoded frames. The simulation results showed that the proposed MCTF scheme achieves better rate-distortion performance.

Second, the proposed MDSC schemes, spatial-domain based and temporal-domain based, are tested. It is assumed that two descriptions are sent over separate channels, and only one description may be completely lost.

Coding efficiency and redundancy analysis

The coding efficiency of Salesman sequence of the proposed MDSC and 3D lifted wavelet based single description coding (SDC) is shown in Fig.10. As may be seen from the curves, SDC scheme achieves better rate-distortion performance than MDSC schemes. The lower efficiency of MDSC schemes results from their more redundancy compared with SDC scheme. Fig.11 shows the corresponding redundancy-ratio-distortion (RRD) performance for Salesman sequence. The temporal-domain MDSC scheme has lower redundancy than the schemes of spatial-domain, and has better coding performance. For the temporal-domain scheme, the redundancy comes from the longer distance of temporal domain, which can utilize MCTF to reduce the impact to coding efficiency. While for spatial-domain schemes, row-downsampling or column-downsampling reduces the spatial-correlation of images, which cannot depend on MCTF to retain the coding efficiency. Similar experimental results are also achieved for Missa sequence.

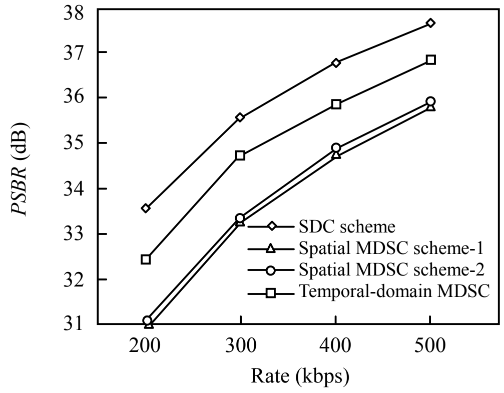


Fig.10 Comparison of coding efficiency

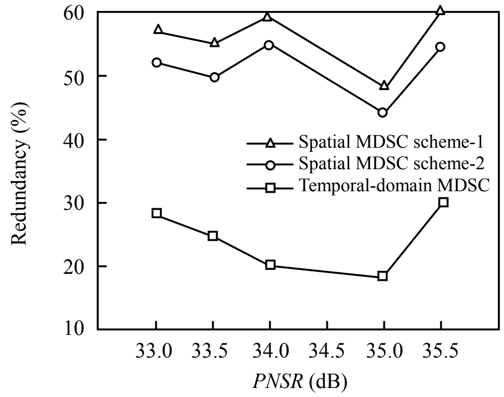


Fig.11 RRD performances of MDSCs

Analysis of robustness and scalability

Fig.12 shows the robustness performance of the spatial-domain MDSC schemes with respect to Salesman sequence. At the receiver, if only one description is received, the schemes can also reconstruct the accepted video with the error concealments. For Salesman sequence, the qualities of reconstructed video with spatial scheme-1 is better than with spatial scheme-2, which is due to the stronger row-correlation than column-correlation. Contrarily, for Missa sequence, spatial scheme-2 has better performance because of its stronger column-correlation. Thus, for the spatial-domain MDSC schemes, the MD schemes based on the veins characteristic of video should be selected.

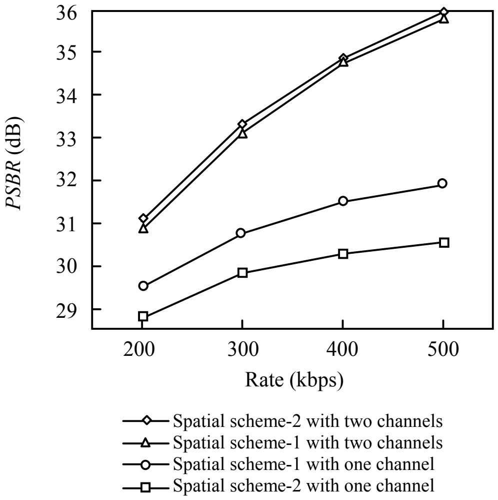


Fig.12 Robustness performance of spatial MDSCs

Fig.13 shows the robustness performance of the temporal-domain MDSC scheme. With the error concealment, the scheme can reconstruct accepted video in the case of one description received. BMC-I scheme has better performance than SMC-I scheme, and the proposed BMC-I error concealment is more effective for the temporal-domain MDSC scheme. In particular, the quality of frames reconstructed with one description does not drop much compared to that of both descriptions received.

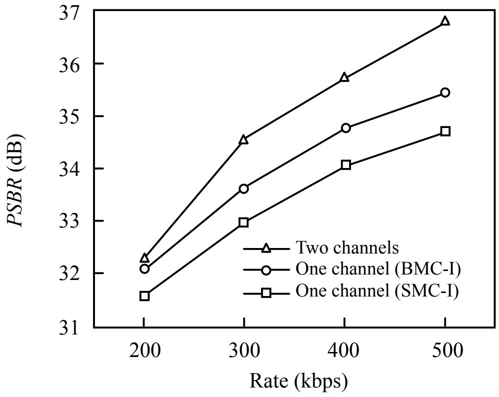


Fig.13 Robustness performance of temporal MDSCs

The proposed MDSC schemes can not only increase the transmission robustness, but also introduce a high degree of scalability into the coding scheme so that one compressed representation can be decoded at a variety of rates and fidelities. Fig.12 and Fig.13 show the average PSNRs of reconstructed frames at different bit rates.

From the simulation results, the proposed MDSC schemes have excellent robustness and scalability performance. The temporal-domain MDSC scheme demonstrates its superior performance with low redundancy.

CONCLUSION

In this work, a new method called block-split bi-directional motion estimation is proposed to deal with the unconnected pixels in MCTF, and based on which new MDSC schemes are proposed. The proposed MDSC schemes can provide truly scalable and highly error resilient video transmission over heterogeneous and unreliable networks. Moreover, the mismatch-related issues between the encoder and the decoder are avoided, and more consistent reconstructed video quality can be obtained at the receiver.

References

Bajic, I.V., Woods, J.W., 2003. Domain-based multiple description coding of images and video. *IEEE Trans. on Image Processing*, **12**(10):1211-1225. [doi:10.1109/TIP.2003.817248]

Hsiang, S.T., Woods, J.W., 2001. Embedded video coding using invertible motion compensated 3-D subband/wavelet filter bank. *Signal Processing: Image Communications*, **16**(8):705-724. [doi:10.1016/S0923-5965(01)00002-9]

ISO/IEC JTC1/SC29/WG11 N5540, 2003. Applications and Requirements for Scalable Video Coding.

Pesquet-Popescu, B., Bottreau, V., 2001. Three-dimensional Lifting Schemes for Motion Compensated Video Compression. Proc. ICASSP. Salt Lake City, **3**:1793-1796.

Reibman, A.R., Jafarkhani, H., Wang, Y., Orchard, M.T., Puri, R., 2002. Multiple-description video coding using motion-compensated temporal prediction. *IEEE Trans. on Circuits Syst. Video Technol.*, **12**(3):193-204. [doi:10.1109/76.993440]

Secker, A., Taubman, D., 2003. Lifting-based invertible motion. Adaptive transform (LIMAT) framework for highly scalable video compression. *IEEE Trans. on Image Processing*, **12**(12):1530-1542. [doi:10.1109/TIP.2003.819433]

Wang, Y., Orchard, M.T., Vaishampayanm, V., Reibman, A.R., 2001. Multiple description coding using pairwise correlating transforms. *IEEE Trans. on Image Processing*, **10**(3):351-366. [doi:10.1109/83.908500]

Yu, M., Jiang, G., He, S., 2005. Review on multiple description coding of video. *Journal of Circuits and Systems*, **10**:76-84.