Bit Stream Extraction Based on Video Content Method in the Scalable Extension of H.264/AVC

Daxing Qian, Hongyu Wang, Wenzhu Sun and Kaiyan Zhu
Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology,
Dalian 116023, P.R.China
Email: qiandaxing23@sina.com, whyu@dlut.edu.cn, sunwenzhu@mail.dlut.edu.cn, zkycat@126.com

***Abstract*—The bit stream extraction plays an important role in Scalable Video Coding (SVC). The video content is an important factor for video coding efficiency but it is ignored in current video coding methods. Therefore, in this paper, we proposed an equivalent MSE method to extract sub streams in the temporal and spatial enhancement layers. When the Motion Vectors (MVs) are large in one video, a larger frame rate is necessary to maintain the continuity of the object movement which makes no jump in visual sense. In this sense, sub streams extraction more temporal enhancement layers have to be satisfied. On the other hand, if there are some larger high-frequency components in a single frame of the video, it should try to meet the extraction requirement in spatial enhancement layer. This method has the advantage of considering the contents of the video, which can effectively improve the coding performance and quality. The experimental results demonstrate that the quality of reconstructed video has been improved significantly by using the equivalent MSE method when extracting bit stream arbitrarily at the same bandwidth.**

***Index Terms*—SVC, bit stream extraction, base on video content, temporal-spatial frequency**

I. INTRODUCTION

With the development of the network technology, the various devices have different demands on formats of multimedia video. From restricted small screen cell phone to the powerful high-definition display computer, it's hard to have a uniform video format to adapt all the various devices and different networks entirely. SVC standard [1] is the extension of the H.264/AVC [2] and developed by the ITU-T in collaboration with ISO. In addition, SVC includes a variety of scalabilities, like spatial, temporal, and quality scalabilities. In SVC, temporal scalability is provided by the concept of hierarchical B-pictures [3]. Spatial scalability, on the other hand, is achieved by encoding each supported spatial resolution into one layer. Quality scalability includes coarse grain scalability (CGS), medium grain scalability (MGS). CGS provides discontinuous decoding points. The number of points equal to the layers coded. MGS divides the refinement coefficients of each of enhancement layer into several fragments so that it can provide a progressive enhancement and graceful degradation. In this paper, we focus on discussing temporal and spatial enhancement layers extraction, so quality scalability is no longer discussed.

The design of SVC can create a video bit stream that is structured in layers, consisting of a base layer (BL) and one or more enhancement layers (EL). Each enhancement layer is able to improve the resolution (of spatially or temporally) or the quality of the video sequence. SVC encodes a video sequence into a bit stream. The scalability of this bit stream is achieved by discard the unimportant enhancement segment according to various device and different network. The more the bits received by terminal device, the better the quality of the reconstructed video will be.

SVC is researched in many fields. In [4], joint application physical-layer design (JAPLD) strategy to cost-effectively transmit scalable H.264/AVC video over multi-input multi-output (MIMO) wireless systems and adaptive channel selection (ACS) methods are proposed. In [5], an efficient inter-layer motion-compensation technique is proposed for enhancement layer in spatially scalable video coding. It exploits the prediction residue correlation between consecutive spatial layers. In [6], an adaptive filtering method in the type-II DCT up-sampling is introduced, which applies different weighting parameters to DCT coefficients.

The working platform mainly follows the 3 steps. Firstly, encoder makes an original video file be a scalable sequence. Secondly, extractor can be used to extract a sub stream which includes a BL and several ELs. Lastly, decoder is invoked to decode the sub stream which be extracted to get the reconstructed video.

Fig.1 shows SVC BL and Els, the smallest picture is reconstructed by base layer. Its format is QCIF (176 x 144), frame rate is 7.5 frames per second (fps). The middle picture is reconstructed by base layer and enhancement layer 1, its format is CIF (352 x 288), frame rate is 15 fps. The biggest picture is reconstructed by all the layers (1BL and 2 ELs) of the video coding sequence. Its format is 4CIF (704 x 576), frame rate is 30 fps. The BL can be decoded independent, while Els decoded always need its reference layers data. Every bit stream which include base layer is a sub stream of a scalable video sequence. Extraction different sub stream can get different reconstructed video quality. At a limited bandwidth, a video sequence extract different temporal and spatial enhancements will get different reconstructed

doi:10.4304/jsw.6.10.2090-2096

quality. Among all the sub streams, there will be an optimum reconstructed quality sub stream. We according to the video content choose the optimum sub stream to extract.

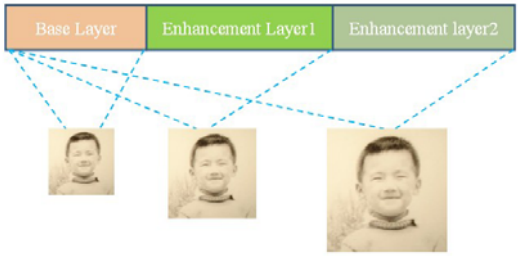


Figure.1 SVC base layer and enhancement layers

When the MVs are large in one video, a larger frame rate is necessary to maintain the continuity of the object movement which makes no jump in the visual sense. In this sense, sub streams extraction in temporal enhancement layer has to be satisfied. On the other hand, if there are some larger high-frequency components in a single frame of the video, that is to say, there are some higher spatial details in the video stream.

This paper is organized as follows. The section II is the related literature and contribution. The section III is the temporal frequency brought by motion. In section IV we introduce spatial frequency of general object in a picture. Experimental results are shown in Section V and finally conclusions are drawn in Section VI.

II. RELATED LITERATURE AND CONTRIBUTION

SVC has been an area to be researched and standardized at least 20 years. The prior international video coding standards H.262/MPEG-2 Video, H.263, and MPEG-4 already include several scalable characteristics. However, these scalable profiles have rarely been used for not only the spatial and quality scalability features came along with a significant loss in coding efficiency, but also a large increase in decoder complexity compared to the corresponding non-scalable profiles [7]. The temporal scalability features could bring more advantages than drawbacks. When temporal scalability is adopted, the coding efficiency can be improved significantly. However, the delay introduced to the encoding and decoding process is inevitable.

During the past two decades, lots of work has been done to have been a significant improvement of SVC. As a kernel module the bit stream extraction has been researched widely. The research work about the bit stream extraction is mainly focused on the quality scalable characteristic and independent of the content of video. A major drawback of the basic extraction method is that its prioritization policy is independent of the video content [8]. The efficiency of the bit extractor can be substantially improved by assigning a priority identifier to each NAL unit during the encoding or a post-processing operation [9]. When motion scalability [10] is taken into consideration, the bit stream extractor has an additional requirement, i.e. optimal bit allocation among motion and texture [11]. In this paper, we consider the method of extraction better sub stream based on the temporal-spatial scalability adaptively. Firstly, the concept Spatial Frequency (SF) is extended to the object of arbitrary size. Secondly, get the minimal frame rate (MFR) of the video that makes no jump in visual sense. Lastly, find the proper frame rate and spatial enhancement layer that is extracted according to the bandwidth.

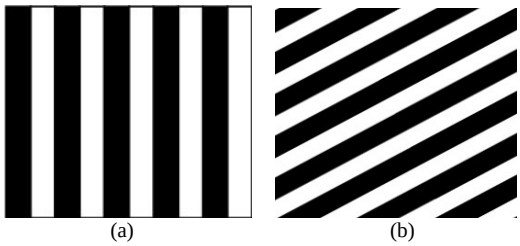
Few schemes of extraction bit stream based on the content of video are proposed. In this paper, we propose an extraction method which is based on equivalent MSE. It considers temporal and spatial ELs but not quality ELs for simplicity. We decide the content of video belong to high or low motion. The high motion video needs more temporal frequency to avoid looking like discontinuous, while the low motion video needs more spatial resolution to adapt human visual sense.

Due to PSNR is not able to represent the reconstructed video quality entirely, We commence the concept of SF of 2D sinusoidal signals according to human visual sense and then extend it to general situation by the equivalent MSE method. We use the SF in a picture and MVs to get the frame rate which makes no jump in visual sense. On this basis, we get the optimal tradeoff between temporal and spatial ELs extraction. The contents mentioned above constitute the main contribution of our paper.

III. THE TEMPORAL FREQUENCY BROUGHT BY MOTION

Spatial resolution is the resolution of a frame that is the total pixels in a picture. Temporal resolution is the refresh rate that is the number of frames refreshed in unit time. The spatial resolution is a 2D plane. It can be formed a 3D space with a time line, but the 2D plane is not orthogonal with the time line. in the following, we will introduce the relation between them.

In [12], the concept SF is introduced. See from Fig. 2.



(a) (b)

Figure.2 2D Sinusoidal signals.

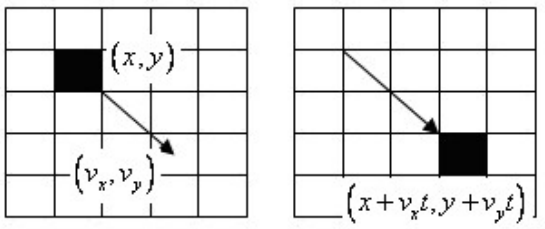
(a) (f_x, f_y) = (5, 0); (b) (f_x, f_y) = (3, 5).

The horizontal and vertical units are the width and height of the image, respectively.

Fig. 3 shows that every point (x, y) at t = 0 is shifted by $(v_x t, v_y t)$ to $(x + v_x t, y + v_y t)$ at time t. Alternatively, a point (x, y) at time t corresponds to a point $(x - v_x t, y - v_y t)$ at time 0. Let the image pattern of the object at time 0 be described by $\psi_0(x, y)$ and its

velocities in horizontal and vertical directions by v_x and v_y. The image pattern at time *t* will be:

$$\psi(x,y,t)=\psi_0(x-v_xt,y-v_yt). \quad (1)$$



(a) (b)

Figure.3 Object is moving at velocity of (v_x, v_y).

Performing CSFT for the above signal, we obtain

$$\begin{aligned}\Psi(f_x,f_y,f_t)&=\iiint\psi(x,y,t)\\&\cdot\exp(-j2\pi(f_xx+f_yy+f_tt))dxdydt\\&=\iint\psi_0(x-v_xt,y-v_yt)\\&\cdot\exp(-j2\pi(f_x(x-v_xt)+f_y(y-v_yt)))dxdy\\&\cdot\int\exp(-j2\pi(f_t+f_xv_x+f_yv_y)t)dt\\&=\Psi_0(f_x,f_y)\int\exp(-j2\pi(f_t+f_xv_x+f_yv_y)t)dt\\&=\Psi_0(f_x,f_y)\delta(f_t+f_xv_x+f_yv_y)\end{aligned} \quad (2)$$

where $\Psi_0(f_x,f_y)$ represents the 2D CSFT of $\psi_0(x,y)$.This means that a spatial pattern characterized by (f_x,f_y) in the object will lead to a temporal frequency.

$$f_t=-f_xv_x-f_yv_y. \quad (3)$$

From (3) we can draw a conclusion that the temporal frequency depends not only on the motion, but also on the spatial frequency of the object.

As shown in Fig. 2(a), the direction of spatial frequency is horizontal. If the plane moves vertically, then the eye will not perceive any change no matter how fast the plane moves. Once its motion is slightly tilted from the vertical direction, the eye will start to perceive temporal changes. The perceived change is most rapid when the plane moves horizontally.

Without loss of generality, we will consider the relation between maximum and minimum temporal frequency and movement of object at arbitrary SF.

$$\begin{cases}f_t=-f_xv_x-f_yv_y\\v_x=v\cdot\cos\theta\\v_y=v\cdot\sin\theta\end{cases}. \quad (4)$$

Where *v* is the velocity vector of the moving object, $|v|=\sqrt{v_x^2+v_y^2}$, θ is the angle between the direction of *v* and horizontal. Bring v_x and v_y to f_t, then $f_t=-f_xv\cdot\cos\theta-f_yv\cdot\sin\theta$, the derivative of the function with respect to the variable θ is $f_t'=f_xv\cdot\sin\theta-f_yv\cdot\cos\theta$, make $f_t'=0$, we can get $\tan\theta=\frac{f_x}{f_y}$,

$$\theta=\arctan\frac{f_x}{f_y}. \quad (5)$$

When directions of the object moving is identical to the SF's, it can gain the maximum f_t. Obviously, if they are orthogonal, the $f_t=0$.

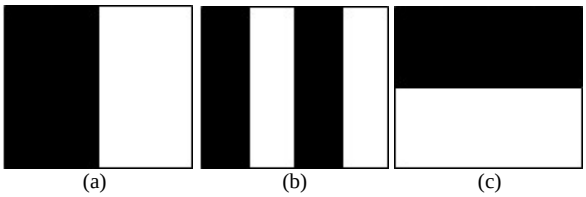
IV. SPATIAL FREQUENCY OF GENERAL OBJECT IN A PICTURE

The frame of video sequence is represented by the partition of MacroBlocks (MBs). SVC inherits the MB partition technique of H.264/AVC. The types of MB size include 16x16, 16x8, 8x16, 8x8. The 8x8 MB is partitioned further into sub MB: 8x8, 4x8, 4x4. In the following, we will introduce how to determine SF of general shape object.

A. The Same f_t and The Different MSE

We study the two special instances: the SF is (f_x, f_y) = (1, 0) and (f_x, f_y) = (2, 0), see from Fig. 4(a) and Fig. 4(b). The size of picture is $W\times H$. The objects of the two pictures are moving at the same speed: (v_x,v_y) = (*v*, 0), after enough short time Δt, we suppose the object of picture is $f_t(x,y)$ is changed to $f_{t+\Delta t}(x,y)$.

Use the equation Mean Square Error (MSE) between two pictures:



(a) (b) (c)

Figure.4 2D Sinusoidal signals.

(a) (f_x, f_y) = (1, 0); (b) (f_x, f_y) = (2, 0); (c) (f_x, f_y) = (0, 1).

$$MSE(x,y)=\frac{1}{WH}\sum_{x=1}^{W}\sum_{y=1}^{H}\left[f_t(x,y)-f_{t+\Delta t}(x,y)\right]^2. \quad (6)$$

We can learn the MSE value of Fig. 4(b) is twice as much as that of the Fig. 4(a). If the SF is (f_x, f_y) = (*n*, 0), it will bring n times MSE value to the Fig. 4(a)'s. So

$\left(f_x, f_y\right)$ = (0, *n*) is the *n* times MSE value to the $\left(f_x, f_y\right)$ = (0, 1).

We look the equation Peak Signal Noise Ratio (PSNR), which reflects the performance of image encode:

$$PSNR_{dB} = 10\lg^{\frac{\left(2^n-1\right)^2}{MSE}}. \quad (7)$$

$\left(2^n-1\right)^2$ is the square of maximum possible signal value, n is the number of bits to represent each pixel. The bigger MSE, the smaller PSNR.

From the Fig. 4(a) and Fig. 4(c), we can learn that the two pictures bring the same f_t and their MSE but PSNR is different. So if use MSE or PSNR as the adaptive parameter is not very reasonable. The Fig. 4(a) and Fig. 4(c) is moving at speed (*v*, 0) and (0, *v*), respectively. They bring the time frequency are both $f_t = -v$, but their MSE value is different obviously, ratio is $\frac{H}{W}$.

B. Equivalent MSE

We propose a method named equivalent MSE to calculate SF of general objects in a picture then get the appropriate frame rate.

$$f_{tB} = -\frac{MSEf\left(\frac{h}{H},0\right)}{MSEf\left(1,0\right)}\cdot v_x - \frac{MSEf\left(0,\frac{w}{W}\right)}{MSEf\left(0,1\right)}\cdot v_y. \quad (8)$$

As seen from Fig. 5, the size of black column is $w \times h$, we use f_{tB} represent the brought frame rate by the object. Let $MSEf\left(f_x,0\right)$ be the MSE between the SF is $\left(f_x,0\right)$ in a picture moving horizontally spend enough short time Δt and the original picture. And $MSEf\left(0, f_y\right)$ is the

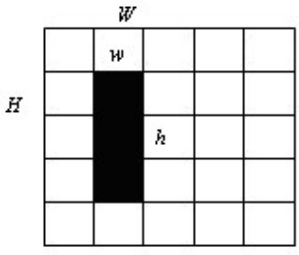


Figure.5 Object size is $w \times h$ and picture size is $W \times H$.

MSE between the SF is $\left(0, f_y\right)$ in a picture moving vertically spend enough short time Δt and the original picture. We regard the SF of a general picture is:

$$\begin{aligned}\left(f_x, f_y\right) &= \left(\frac{MSEf\left(\frac{h}{H},0\right)}{MSEf\left(1,0\right)}, \frac{MSEf\left(0,\frac{w}{W}\right)}{MSEf\left(0,1\right)}\right). \quad (9)\\ &= \left(\frac{h}{H},\frac{w}{W}\right)\end{aligned}$$

The objects in a picture bring frame rate from moving from arbitrary directions is:

$$\begin{aligned}f_t &= \sum f_{tb}\\ &= \sum\left(-\frac{MSEf\left(\frac{h}{H},0\right)}{MSEf\left(1,0\right)}\cdot v_x - \frac{MSEf\left(0,\frac{w}{W}\right)}{MSEf\left(0,1\right)}\cdot v_y\right). \quad (10)\\ &= \sum\left(-\frac{h}{H}\cdot v_x - \frac{w}{W}\cdot v_y\right)\end{aligned}$$

$\sum$ is all the MBs in the picture.

We get the MFR that makes the video seems continuous, then the spatial resolution of the sequence according to the bandwidth. We expect to get an optimal tradeoff between the temporal and spatial scalability.

V. EXPERIMENTAL RESULTS

We use two CIF test sequences (i.e., Mobile and Bridge-far) to compare the performance of the proposed scheme and use [13, 14] reference software codec. Each GOP size is 16 and structured to provide 5 temporal layers. Each video sequence we coded the first 2 GOPs (32 frames).

The Mobile is dynamic while the Bridge-far is static. So the Mobile required more frame rate, but less spatial resolution than Bridge-far. It will bring better visual effect.

We use (D, T) to represent the layer to be extracted. D is the stage of spatial scalability and T is the stage of temporal scalability. As shown from Tab. 1 and Tab. 2.

TABLE I.
SPATIAL SCALABILITY STAGE CORRESPONDING TO RESOLUTION

Spatial stage	D = 0	D = 1
resolution	QCIF	CIF

TABLE II.
TEMPORAL SCALABILITY STAGES CORRESPONDING TO FRAME RATE

Temporal stage	T = 0	T = 1	T = 2	T = 3	T = 4
Frame rate[fps]	1.875	3.75	7.5	15	30

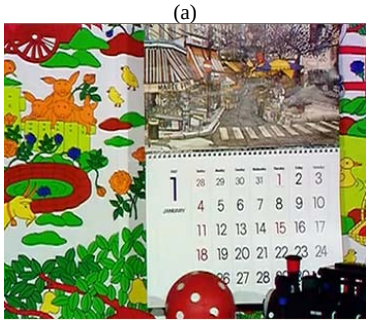
Each test sequence is partitioned into two spatial layers and five temporal layers. We extracted sub stream from different temporal and spatial layers at the same bit-rate as possible as we can.

A. Compare Subjective Performance

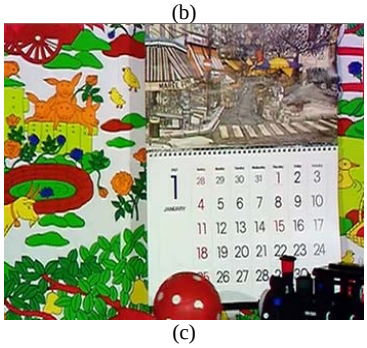
In the following, Mobile and Bridge-far at various spatial and temporal resolution are shown.

Seen from Fig. 6 and Fig. 7, the single picture quality in the extracted sub stream (D, T) = (0, 4) looks worse while the object move continuously. The single frame quality in the sub stream (D, T) = (1, 2) is close to the original sequence while the object movement looks like jumping (observe between ball and calendar) and that makes awful in vision.

(a)



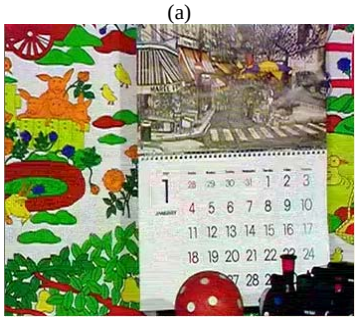
(b)



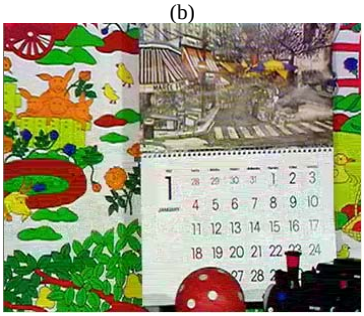
(c)

Figure.6 Mobile decoded at (D, T) = (1, 2): (a) The first frame in the test video;
(b) The second frame in the test video; (c) The third frame in the test video

(a)



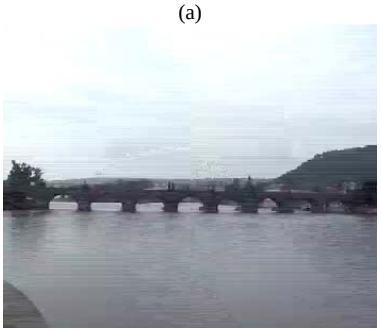
(b)



(c)

Figure.7. Mobile decoded at (D, T) = (0, 4): (a) The first frame in the test video;
(b) The second frame in the test video; (c) The third frame in the test video

(a)



(b)

Figure.8 Bridge-far decoded at (D, T) = (0, 4): (a) The first frame in the test video;
(b) The second frame in the test video

(a)

(b)

Figure.9 Bridge-far decoded at (D, T) = (1, 3): (a) The first frame in the test video;
(b) The second frame in the test video

In the Fig. 8 and Fig. 9, there are few moving objects between the two frames and the MVs are small. The human visual is more willing to accept more spatial resolution but not the temporal resolution. As is shown in

(D, T) = (0, 4) and (D, T) = (1, 3) of Fig. 8 and Fig. 9, obviously the (D, T) = (1, 3) is better.

B. Compare Objective Performance

We will show the extracted layers, bit-rate and PSNR of the section IV.A in Tab. III.

TABLE III.
EXTRACTION VIDEO SEQUENCE AT BIT-RATE AND Y-PSNR

	Mobile		Bridge-far	
(D, T)	(0, 4)	(1, 2)	(0, 4)	(1, 3)
Y-PSNR[dB]	28.4753	17.7137	33.1305	37.5356
Bit-rate[kbps]	1344.80	1025.30	197.70	156.30

Obviously, different video sequences and different sub streams can bring different PSNR and subjective feelings. Extract more temporal ELs in Mobile can get better visual effect, while in Bridge-far extract more spatial Els that the visual effect can be better.

VI. CONCLUSION

In this paper, we propose a method equivalent MSE to determine how to extract temporal and spatial enhancement layer. It may get the SF in a picture of a video sequence and calculate the MFR which makes it like to be continuous. It makes up the drawback of the current extraction method that is independent of the video content. Experimental results show that a significant gain can be achieved with the method. In this paper, quality scalability has not been considered for simplicity. In the future, we will extend our work to quality layers research.

REFERENCES

[1] Text of ISO/IEC 14496-10:2005/FDAM 3 Scalable Video Coding, Joint Video Team (JVT) of ISO-IEC MPEG & ITU-T VCEG, Lausanne, N9197, Sep. 2007.
[2] ISO/IEC ITU-T Rec. H264: Advanced Video Coding for Generic Audiovisual Services, Joint Video Team (JVT) of ISO-IEC MPEG & ITU-T VCEG, Int. Standard, May 2003.
[3] H. Schwarz, D. Marpe, and T. Wiegand, in: Hierarchical B pictures. Joint Video Team, Doc. JVT-P014, July 2005.
[4] Daewon Song, Chang Wen Chen, Scalable H.264/AVC Video Transmission Over MIMO Wireless Systems With Adaptive Channel Selection Based on Partial Channel Information, IEEE Transactions on Circuits and Systems For Video Technology, Vol. 17, No.9, September 2007.
[5] Rong Zhang, Mary L. Comer, Efficient Inter-Layer Motion Compensation for Spatially Scalable Video Coding, IEEE Transactions on Circuits and Systems For Video Technology, Vol.18, No.10, October 2008.
[6] IlHong Shin, Hyun Wook Park, Adaptive Up-Sampling Method Using DCT for Spatial Scalability of Scalable Video Coding, IEEE Transactions on Circuits and Systems For Video Technology, Vol. 19, No. 2, February 2009.
[7] Heiko Schwarz, Detlev Marpe, Thomas Wiegand, in: Overview of the Scalable Video Coding Extension of the H.264/AVC Standard, IEEE Transactions on Circuits and Systems For Video Technology, Vol. 17, No. 9, September 2007.
[8] Ehsan Maani, Aggelos K. Katsaggelos, Optimized Bit Extraction Using Distortion Modeling in the Scalable Extension of H.264/AVC, IEEE Transactions on Image Processing, Vol. 18, No.9, September 2009.
[9] I. Amonou, N. Cammas, S. Kervadec, and S. Pateux. Optimized rate-distortion extraction with quality layers in the scalable extension of H.264/AVC. IEEE Trans. Circuits Syst. Video Techn., 17(9):1186–1193, 2007.
[10] M.-P. Kao and T. Nguyen, A fully scalable motion model for scalable video coding, IEEE Trans. Image Process., vol. 17, no. 6, pp. 908 – 923, Jun. 2008.
[11] J. Barbarien, A. Munteanu, F. Verdicchio, Y. Andreopoulos, J. Cornelis, and P. Schelkens, Motion and texture rate-allocation for predictionbased scalable motion-vector coding, EURASIP Signal Processing: Image Communication, vol. 20, pp. 315 – 342, Apr. 2005.
[12] Yao Wang, Jorn Ostermann, Ya-Qin Zhang, in: Video Processing and Communications.(2001)
[13] Joint Scalable Video Model JSVM 9_12_2 .
[14] JSVM Software Manual. JSVM 9.12.2 (CVS tag: JSVM_9_12_2) ,Ap

Daxing Qian received his B.S. degree in Information Science and Engineering from Northeastern University, Shenyang, P.R. China in 2002. He is now a Ph.D. candidate in the Department of the Faculty of Electronic Information and Electrical Engineering at Dalian University of Technology, Dalian, P.R. China. His research interests are scalable video coding, joint source and channel coding and information theory.

Hongyu Wang received the B.S. and M.S. degrees respectively from Jilin University of Technology , Jilin, Tianjin, P.R. China and Graduate School of Chinese Academy of Sciences in 1990 and 1993, both in Electronic Engineering. He received the Ph.D. in Precision Instrument and Optoelectronics Engineering from Tianjin University, Tianjin, P.R. China in 1997. Currently, he is a Professor in the institute of Information Science and Communication Engineering, Dalian University of Technology, P.R. China, he had been an Assistant Professor and Associate Professor in the Department of Electronic Engineering, Zhejiang University, Zhejiang，P.R. China from 1997 to 2004. Dr. Wang's research interests include algorithmic, optimization, and performance issues in wireless ad hoc, mesh and sensor networks, cross-layer design and optimization, and multimedia communications.

Wenzhu Sun received the B.S degree in information and electrical engineering from Shandong Jianzhu University, Jinan, P.R. China in 2006. He is currently pursuing the Ph.D. degree from the Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian, P.R. China. His research interests include joint source and channel coding, scalable video coding, and convex optimization.

Kaiyan Zhu received the B.S. and M.S. degrees in information and electrical engineering from Jilin University, Changchun, P.R. China in 2002, 2005, respectively. She is now a ph.D. candidate in the Department of the Faculty of Electronic Information and Electrical Engineering at Dalian University of Technology, Dalian, P.R. China. Her research interests are cooperative diversity and network coding.