Fussion Based Side Information Creation Method for Distributed Scalable Video Coding

Phương pháp tạo thông tin phụ trợ dựa trên kỹ thuật kết hợp ảnh trong mã hóa video phân tán liên lớp

Nguyen Thi Huong Thao**, Vu Van San, Vu Huu Tien

Posts and Telecommunications Institute of Technology, Nguyen Trai, Hanoi, Viet Nam

Received: December 20, 2016; accepted: September 5, 2017

Abstract

In recent years, video entertainment demand has significantly changed. Video content is transmitted through different bandwidth connections and played on many devices that have different processing capabilities and screen sizes. For this reason, scalable extensions of video coding standards have been released, e.g SHVC, scalable extension of HEVC. Beside high compression efficiency, SHVC has disadvantages including high encoder complexity and weakness in error resilience. These are not suitable for emerging applications such as wireless sensor networks, video surveillance systems or remote sensing that have limited processing capabilities, low energy and low network bandwidth. A potential solution supporting for these systems is Distributed Scalable Video Coding (DSVC). In DSVC system, Side Information (SI) creation plays a critical role in deciding system overall performance. Therefore, this paper proposes a spatially scalable DSVC architecture and a new side information creation technique for this DSVC system. Results show that the proposed method generates better quality SI when compared to some previous SI creation methods. Consequently, the system performance is improved when compared with the previous methods.

Keywords: DVC, DSVC, Side information

Tóm tắt

Trong những năm gần đây, nhu cầu sử dụng video đã thay đổi rất nhiều. Nội dung video được truyền qua các kết nối có băng thông khác nhau và được hiển thị trên các thiết bị có khả năng xử lý khác nhau và kích thước màn hình khác nhau. Vì lý do này, các chuẩn mở rộng khả năng liên lớp của các chuẩn video truyền thống đã ra đời nhằm đáp ứng nhu cầu trên, ví dụ SHVC của HEVC hoặc SVC của H.264/AVC. Tuy nhiên, các chuẩn mở rộng này lại không phù hợp với các dịch vụ mới như mạng cảm biến không dây, mạng giám sát video hay mạng cảm biến từ xa. Một giải pháp thay thế đầy tiềm năng là hệ thống mã hóa video liên lớp phân tán (DSVC). Đối với DSVC, thông tin phụ trợ (SI) đóng một vai trò quan trọng trong việc quyết định hiệu năng hệ thống. Vì vậy bài báo này đề xuất một kiến trúc DSVC liên lớp không gian mới và một kỹ thuật tạo thông tin phụ trợ mới cho hệ thống DSVC này. Các kết quả đã chỉ ra rằng phương pháp đề xuất tạo ra SI chất lượng tốt hơn khi so sánh với các phương pháp tạo SI trước đó. Do đó, hiệu năng hệ thống cũng được cải thiện khi so với các phương pháp trước đó.

Từ khóa: DVC, DSVC, thông tin phụ trợ

1. Introduction

There will be 50 billion connected devices by 2020 [1], and more than 15 billion of these will be video enabled. Such an amazing increased use of video has an important contribution of video compression. Current video standards such as H.264/AVC, HEVC bring high compression performance with good video quality. But today, there is a large variety of video devices with the diversity of screen size, bandwidth and processing power. In order to address this, scalable extensions are introduced along with current video standards. Scalable Video Coding (SVC) and Scalable High efficiency Video Coding (SHVC) are scalable extensions of H.264/AVC and HEVC respectively. These extensions enable transmission and decoding of partial bit streams to provide video services with lower temporal or spatial resolutions or reduced fidelity while retaining a reconstruction quality that is high relative to the rate of the partial bit streams. Structure of scalable video content is defined as a combining of one base layer (BL) and several enhancement layers (EL). BL corresponds to lowest video performance, ELs improve the quality of the BL. The main types of scalability are spatial, temporal and quality scalabilities. Spatial scalability and temporal scalability describe cases in which sub-bitstream represents the video content with a reduced picture size (or spatial resolution) and frame rate (or temporal resolution), respectively. With quality

* Corresponding author: Tel.: (+84) 915009199
Email: thaotb07@gmail.com

scalability, the sub-bitstream provides the spatial and temporal resolution as the complete bitstream but with a lower reconstruction quality. Main characteristics of these video standards are that complex computational operations such as motion estimation and compensation are performed at the encoder in order to achieve high rate-distortion performance, while the decoder can directly use the motion vectors to decode the sequence. Therefore, the encoder is much more complicated than the decoder. This architecture is not suitable for emerging applications such as wireless video surveillance, multimedia sensor networks where video sequence is encoded many times and decoded once and devices require a low complexity encoding while affording a high complexity decoding. An alternative solution is Distributed Video Coding (DVC) [2]. DVC is a new video coding paradigm where temporal correlation among successive frames is exploited at the decoder instead of the encoder, therefore, encoder complexity is much lighter than decoder complexity. In other words, the encoder complexity is shifted to the decoder. DVC is based on two information theory results, the Slepian-Wolf theorem for lossless compression and the Wyner-Ziv (WZ) theorem for lossy compression. Therefore, DVC encoder/decoder is named Wyner-Ziv encoder/decoder. In DVC codec, the video sequence is split into key frames (KF) and Wyner-Ziv frames (WZF). KFs are independently encoded and decoded by predictive video approach such as H.264/AVC or HEVC while WZFs are block based transformed and quantized and syndrome encoded. Syndrome encoder encodes quantized coefficients with fewer bits for delivering. To help the decoder to check the correctness of decoded frames, some auxiliary information is sent to the decoder. At the decoder, a side information (SI) frame, an estimation of WZF, is generated by using previously decoded KFs and even WZFs. This SI frame and syndrome bits and auxiliary information are used to decode WZFs. This DVC architecture is ideal for emerging applications mentioned above because almost complicated operations are performed at the decoder. With benefits received from DVC architecture and desire for providing scalabilites, the Distributed Scalable Video Coding (DSVC) solution [3,4] is proposed. DSVC refers to a new video coding paradigm in which DVC principles are used while providing scalabilities. By combining predictive and distributed video coding prinpciples, this DSVC architecture makes use of advantages of two coding paradigms, high compression efficiency of predictive coding and low encoding complexity, robustness to error of distributed coding while still providing scalable characteristics. In addtion, due to base layer is encoded by predictive coding, DSVC have backward compatibility with current video standards. DSVC becomes a potential solution for scalable video coding. So far, some DSVC architectures [3,4,5,6] have been proposed and basically, DSVC includes one base layer which is coded by current video standards, e.g H264/AVC or HEVC, and one or more enhancement layers which are coded by distributed video coding principles. In DSVC, SI plays an important role because the more alike to orginal WZF the generated SI, the fewer the bits that need to be sent to the decoder and the quality of the WZF at the decoder is better. Thus, a spatially scalable DSVC architecture and fusion based side information creation for spatially scalable DSVC are proposed. Moreover, adaptive block size selection at the encoder is also proposed to improve compression efficiency.

The rest of the paper is organized as follows. Section 2 introduces related work. Spatially scalable DSVC architecture, fusion based side information creation technique and adaptive block size selection are depicted in Section 3. Experimental results and discussion are shown in Section 4 in order to evaluate the efficiency of proposed methods. Finally, conclusions are presented in Section 5.

2. Related background work

2.1 Wyner - Ziv coding architecture

Based on Slepian-Wolf and Wyner-Ziv theorems, there are two main practical DVC implementations which are the DVC Stanford solution and the DVC Berkeley solution named PRISM (Power-efficient, Robust, hIgh-compression, Syndrome-based Multimedia coding). The transform-domain WZ (TDWZ) codec architecture is illustrated in Fig.1.

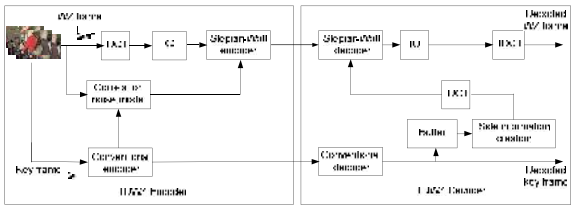

Fig.1. Wyner-Ziv codec architecture

At the encoder, the input video sequence is split into WZFs and KFs. KFs are conventionally encoded while WZFs are WZ encoded. For simplicity, Group of Pictures (GOP), which refers to the number of the frames from one KF to the next KF, is set to 2. This means that the odd frames are KFs and the even frames are WZFs. The WZFs (typically only luminance but the same solution could be applied also to the chrominances) are divided into non-overlapping blocks. Each block is then independently coded using the following coding modules: Discrete

Cosine Transform (DCT), Quantizer (Q), Correlation noise model, Slepian-Wolf encoder.

At the decoder, KFs and WZF are conventionally decoded and Wyner-Ziv decoded, respectively. Buffer module stores next and previous KFs that are used by side information creation module. The detail of this architecture can be found at [7].

In DVC architecture, the more accurate the estimated side information, the fewer the bits need to transmit and the better the reconstructed WZF. Therefore, the bit rate may be reduced for the same quality.

2.2 Side Information Creation

In both DVC and DSVC, the choice of the techniques to generate the side information significantly influences the rate-distortion performance. The SI is more similar to the original WZF, the necessary bit rate for the same quality is reduced. Therefore, many SI creation techniques have been proposed so far. SI creation techniques can be classified into extrapolation, interpolation and hash based techniques [8]. With extrapolation methods, SI is gererated by using only past reference frames [9,10]. The authors in [9] proposed a extrapolation based SI generation method including four steps. First, motion vectors are estimated by using two previously decoded frames. Then, a new motion vector for each block is calculated by averaging all neighboring motion vectors. The next step is motion projection. The pixels from the last decoded frame (or other) are projected to the next time instant using the motion field obtained above assuming that the motion is linear. The last step is to process overlapping and uncovered areas. The results show that this method produces SI with low delay but quite low quality. [10] proposed an extrapolation method combining with motion estimation algorithm. Before generating side information by extrapolating previously decoded frames, a 3-D Recursive Search motion estimation algorithm is performed to get a good estimation of motion. On average the proposed extrapolation scheme comes very close to the performance of some interpolation schemes. As seen, the extrapolation methods only use past temporal adjacent frames of a WZF, so extrapolation based SI generation methods generally achieve poor performance although they have low delay.

Another approach is based on interpolation. The SI frame in this case is estimated based on neighboring frames using both next and previous reference frames. Assuming that Y_{2i}, X_{2i-1}, X_{2i+1} are SI, previous and next key frames, respectively. The simplest frame interpolation technique is to make Y_{2i} equal to X_{2i-1}. Another simple technique is to perform bilinear interpolation between the KFs, X_{2i-1}, X_{2i+1}. However, these techniques are only appropriate to low motion video sequences where the similarity between adjacent frames is quite high. In order to achieve a more accurate SI, more complicated techniques are used. The most used SI generation method is Motion-Compensated Temporal Interpolation (MCTI) [11]. Firstly, key frames are low pass filtered to improve the reliability of the motion vectors. Then, forward motion estimation is performed to creat the motion vector for each block in the interpolated frame. These motion vectors are refined by bidirectional motion estimation module. In order to reduce the number of the false motion vectors, the weighted median filter is used to improve spatial smoothing or to remove outlier motion vector. Finally, two motion compensated blocks in KFs are averaged to produce the side information. MCTI can be used together global motion estimation to creat the better side information [12]. In this paper, an another side information is generated named the global motion compensation side information (GMC SI). The GMC SI and MCTI SI are fused at the decoder to create SI. Authors in [4] also proposed a new SI creation architecture for DSVC. This architecture exploits temporal correlations between the neighboring EL decoded frames and inter-layer correlation from the BL decoded frames in order to generate two SI candidates. Together SI candidate created by MCTI, they are fed to SI fusion module to choose the best SI.

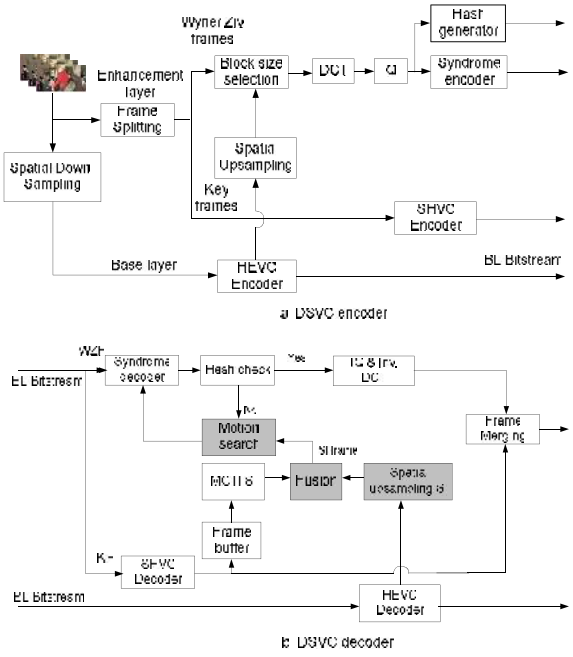

Fig. 2. DSVC encoder and decoder architecture

In DSVC framework, not only the KFs at the EL are used but also BL frames and their motion information are used. Therefore, the next section will introduce a spatially scalable DSVC architecture with a novel SI creation method.

3. Proposed side information creation solution for spatially scalable DSVC

Architecture of DSVC codec with spatial scalability is illustrated in Fig.2. The detail of this architecture is described in [14].

3.1 Generating SI candidates

In this DSVC architecture, there are two SI candidates: one SI is estimated using the MCTI [11] algorithm, SI_{MCTI} and another SI is generated by upsampling the BL decoded frame, SI_{BL}. MCTI technique includes the following steps: 1) Low pass filter: The next and previous key frames are low pass filtered to improve the reliability of the motion vectors; 2) Forward motion estimation: A motion vector is estimated for every block in the next KF with reference to the previous KF; 3) Bidirectional motion estimation: Using the trajectories of the motion vectors, a symmetric motion vector is selected for each block in the SI by selecting the one with intersection in SI closest to the center of the block. Then, the symmetric motion vector is split into a forward motion vector, and a backward one, assuming constant motion. The two motion vectors are further refined in a small area, keeping symmetric during this refinement process; 4) Spatial smoothing: In order to remove outlier motion vectors, weighted vector median filter is applied to the two motion vector fields; 5) Bidirectional motion compensation: the two motion compensated blocks in the next and previous KFs are averaged to produce the side information.

SI_{BL} is generated by using the same upsampling filter in SHVC [14]. Each frame in the BL are upsampled with a finite impulse response (FIR) filter. When upsampling a frame by the ratio N, the concept used is to interpolate the frame to 16 times it's size and then decreasing the size by a ratio M where $M = 16/N$, in both x and y directions. This is done with a 16-phase filter with 8 and 4 taps for luma and chroma respectively. For more details, see [14].

3.2 Improving qualtiy of SI_{MCTI}

In this section, a method to improve MCTI frame quality is proposed by using upscaled frame as an oracle frame. In the proposed method, upscaled frame is selected as an oracle frame because of two reasons: (1) The quality of upscaled frame is higher the MCTI frame and (2) The upscaled frame is available at the decoder.

After MCTI frame is generated from previous and next key frames by bidirection motion compensation, pixels of three frames are compared to appropriated pixels in upscaled frame. The pixels, which have the minimum difference values, are assigned to the pixels value of MCTI frame.

Assumed that P_{ij}^{B}, P_{ij}^{F}, P_{ij}^{M} and P_{ij}^{U} are values of pixels in the backward, forward, MCTI and upscaled frame, appropriately. The value of new MCTI frame, SI_{New_MCTI}, is computed as:

$$M = MIN\left(\left|P_{i,j}^{B} - P_{i,j}^{U}\right|, \left|P_{i,j}^{F} - P_{i,j}^{U}\right|, \left|P_{i,j}^{M} - P_{i,j}^{U}\right|\right) \quad (3)$$

$$P_{i,j}^{M} = \begin{cases} P_{i,j}^{B} \ \ if\ M = \left|P_{i,j}^{B} - P_{i,j}^{U}\right| \\ P_{i,j}^{F} \ \ if\ M = \left|P_{i,j}^{F} - P_{i,j}^{U}\right| \end{cases} \quad (4)$$

3.3 Fusion of SI_{New_MCTI} and SI_{BL}

After improving the quality of MCTI, the fusion frame is generated by combining the upscaled frame, SI_{BL}, and the improved MCTI frame generated in the previous step, SI_{New_MCTI}. In particularly, the value of pixels in fusion frame, SI_{Fusion}, is computed as:

$$P_{ij}^{Fusion} = \alpha\ P_{ij}^{M} + (1-\alpha)\,P_{ij}^{U} \quad (5)$$

In Eq.(5), α is weight measuring the contribution of SI_{New_MCTI} *and* SI_{BL} to the quality of SI_{Fusion}. In this work, α is selected exhaustively equal to 0.1.

4. Experimental results

4.1. Test conditions

To evaluate the performance of the proposed side information creation solution regarding other methods, the following video sequences are used: BasketballDrill, BQMall, PartyScene and RaceHorses with characteristics presented in Table 1.

4.2. Results

Table 2 illustrates PSNR of SI_{New_MCTI} *and* SI_{BL}, SI_{Fusion}. The results show that the quality of SI_{BL} is better than the SI_{New_MCTI} because SI_{BL} is interpolated from the BL frame that is the closest to the original WZ frame and SI_{New_MCTI} is interpolated from the adjacent EL key frames. It is clear that the fusion of SI_{New_MCTI} *and* SI_{BL} achieves the best quality SI for four sequences.

Table 1. Test conditions

Test sequence	Spatial Resolut-ion	Tem-poral Resol-ution	Num-ber of frames	Quantizati on parameter
Basketball Drill	EL: 832 x 480 BL: 416 x 240	50 Hz	50	EL: 32 BL: 30
BQMall		60 Hz	50	EL: 32 BL: 30
PartyScene		50 Hz	50	EL: 32 BL: 30
RaceHorses		30 Hz	50	EL: 32 BL: 30

Table 2. PSNR of SI_{New_MCTI}, SI_{BL} and SI_{Fusion}

	SI_{New_MCTI}	SI_{BL}	SI_{Fusion}
BasketballDrill	26.33	31.67	31.95
BQMall	27.70	28.12	28.36
PartyScene	25.27	24.99	25.31
RaceHorses	21.84	29.94	29.97

As shown in the Table 2, MCTI method gives the lowest PSNR and the quality of the fusion frame significantly depends on the quality of the SI_{BL}. However, SI_{New_MCTI} also contributes to the quality of the fusion frame.

To validate the efficiency of the proposed SI creation method, PSNR of EL layer in this system is measured and compared with the method in [14]. The results are depicted in Table 3 and figures 3, 4 in which $DSVC_{Prev}$ is the method using the previous frame as SI frame and $DSVC_{Fusion}$ is method using the fusion frame as SI frame. The results show that the $DSVC_{Fusion}$ is better than $DSVC_{Prev}$. The reason is that the PSNR of the fusion frame is higher than the previous frame. Therefore, LSB bits of each block is more exactly decoded. Consequently, PSNR of the proposed method using fusion frame is better.

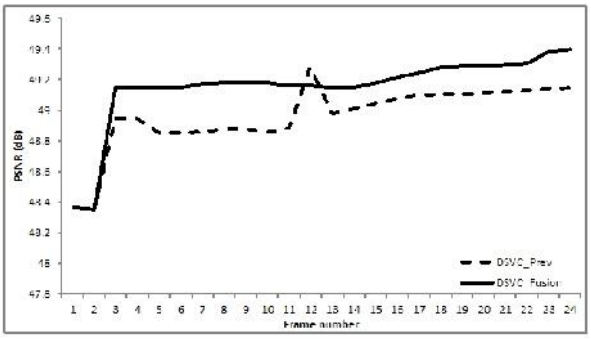

Fig. 3. PSNR of system using previous frame as SI frame for BasketBall Drill sequence

Table 3. PSNR comparison of $DSVC_{Prev}$ and $DSVC_{Fusion}$ methods

	$DSVC_{Prev}$	$DSVC_{Fusion}$
BasketballDrill	48.95	49.14
BQMall	50.30	50.49
PartyScene	49.91	50.04
RaceHorses	51.92	51.79

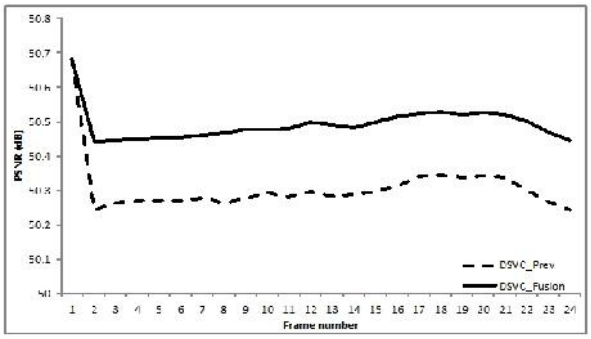

Fig. 4. PSNR of system using previous frame as SI frame for BQMall sequence

5. Conclusion

In this paper, a new SI creation method is proposed for spatially scalable DSVC. In this DSVC architecture, the BL is coded by predictive video coding and the EL is coded by distributed video coding. So, the proposed SI creation technique is applied to EL of the DSVC system. The SI frame is created by fusing two SI candidates, SI_{MCTI} and SI_{BL}. The experimental results show that the proposed SI creation method can significantly improve the SI quality comparing with the separate algorithms such as MCTI and upsampling methods. Applying the SI creation method to the DSVC, the PSNR of the proposed method using the fusion frame also give the better results.

References

[1] https://www.ericsson.com/openarticle/mwc-connected-devices_1686565587.

[2] P.L Dragotti and M. Gastpar, Distributed Source Coding: Theory, Algorithms and Applications, Academic Press, Feb. 2009.

[3] X. HoangVan, J. Ascenso, and F. Pereira, HEVC backward compatible scalability: A low encoding complexity distributed video coding based approach, Signal Process.: Image Commun., vol. 33, no. 4, pp. 51-70, Apr. 2015.

[4] X. HoangVan, J. Ascenso, and F. Pereira., Adaptive Scalable Video Coding: a HEVC based Framework Combining the Predictive and Distributed Paradigms, IEEE TCSVT, vol. 99, no. 00, pp. 1-14, Mar. 2016.

[5] Myeong-jin Lee, "Side information generation for frame rate scalable distributed video codecs", Electronics Letters, vol. 50, no. 5, pp. 370-372, Feb. 2014.

[6] BrunoMacchiavello, Fernanda Brandi, Eduardo Peixoto, Ricardo L. de Queiroz, and Debargha Mukherjee, "Side-Information Generation for Temporally and Spatially scalable Wyner-Ziv Codecs, EURASIP Journal on Image and Video Processing, Volume 2009, Article ID 171257.

[7] B. Girod, et al., Distributed video coding, Proc.IEEE93(1) (2005) 71–83.

[8] Yuan Jia, Yangli Wang, Rui Song, Jiandong Li: Decoder side information generation techniques in Wyner-Ziv video coding: a review, Springer Science+Business Media New York 2013.

[9] Natario L, Brites C, Ascenso J, Pereira F: Extrapolating side information for low-delay pixel-domain distributed video coding. In 9th international workshop on visual content processing and representation, VLBV 2005, 15–16 Sept 2005.

[10] Borchert, R.P. Westerlaken, R. Klein Gunnewiek, and R.L. Lagendijk. On extrapolating side information in distributed video coding. In 26th Picture Coding System, Lisbon, Portugal, November 2007.

[11] Ascenso J, Brites C, Pereira F (2005) Improving frame interpolation with spatial motion smoothing for pixel domain distributed video coding. In: 5th EURASIP conference on speech and image processing, multimedia communications and services. Smolenice, Slovak Republic, pp 1–6.

[12] Abou-Elailah A, Dufaux F, Farah J, Cagnazzo M, Pesquest-Popescu B (2013) Fusion of global and local motion estimation for distributed video coding. IEEE Trans Circuits Syst Video Technol 23(1):158–172.

[13] AaronA, Rane S,Girod B (2004) Wyner-ziv video coding with hash-based motion compensation at the receiver. In: IEEE International Conference on Image Processing, ICIP 2004. vol 2. IEEE, pp 3097–3100.

[14] Thao N.T.H, Tien.V.H, Xiem H.V, Duong D.T, Side information creation using adaptive block size for distributed video coding. Advanced Technologies for Communications (ATC), 2016 International Conference, pp 339-343.