Temporal Scalability Comparison of the H.264/SVC and Distributed Video Codec

Xin Huang [1], Anna Ukhanova [2], Eugeniy Belyaev [2], Søren Forchhammer [1]
[1] DTU Fotonik, Technical University of Denmark
[2] State University of Aerospace Instrumentation, Saint-Petersburg, Russia

Abstract— **The problem of the multimedia scalable video streaming is a current topic of interest. There exist many methods for scalable video coding. This paper is focused on the scalable extension of H.264/AVC (H.264/SVC) and distributed video coding (DVC). The paper presents an efficiency comparison of SVC and DVC having reduced encoder complexity. Moreover, temporal scalability is described for these two algorithms, and it is analyzed and compared.**

I. Introduction

Scalable video coding is very interesting for multimedia networks. Various clients might require decoding of the same video at different resolutions and qualities. Therefore, scalable coding encodes the video only once and enables decoding at different qualities, spatial and temporal resolutions. It makes scalable video coding attractive for different applications. The Moving Picture Experts Group (MPEG) has recently introduced the Scalable Video Coding (SVC) standard [1], which is an extension of the H.264/MPEG-4 Advanced Video Coding (AVC) standard [2]. SVC achieves very good compression performance. On the other hand, SVC entails a higher complexity at the encoder side. Another approach is taken in the field of Distributed Video Coding [3] as a new video coding paradigm to deal with lossy source coding using side information to exploit the statistics at the decoder to reduce computational demands at the encoder. Using DVC, for example, the burden of motion estimation and compensation can be shifted from the encoder to the decoder. This implies low power / low complexity encoders.

The paradigm of distributed source coding (DSC), which has its roots in the theory of coding correlated sources developed by Slepian and Wolf [4] for the lossless case and Wyner and Ziv [5] for the lossy case, has recently become the focus of different kinds of video coding schemes [6], [7]. DVC is promising in creating reversed complexity codecs for power constrained (handheld) devices. Unlike regular broadcast oriented video codecs with high encoding complexity and low decoding complexity, reversed complexity codecs have low encoding complexity but high decoding complexity.

SVC could be used in the situation when we have many receivers and it is needed to receive the data at different bitrates. This can be used for the following:

- video transmitted over Internet for the users with different receiving rate;
- digital TV (DVB-T, DVB-H, ATSC, DTMB, ISDB, SBTVD);
- wireless transmission (on the base of Wi-MAX, WiFi).

Another case is when we have to control the transmission rate depending on the situation in the channel. If the channel becomes worse, it is possible to use scalable stream for power saving [11]. As for DVC, it will suit the situation better, when there are limitations for the complexity and memory of the encoder, and also for power consumption. In a number of resource critical applications, a complex video encoder is a disadvantage in terms of physical size and power consumption. DVC is proposed to apply in areas, where the cost of separated video encoders is the primary concern:

- wireless video surveillance;
- low-power video sensors;
- wireless digital cameras and camera embedded mobile phones.

The goal of this paper is to explore the efficiency of the temporal scalability of DVC and SVC for reduced encoder complexity. By comparing the coding performance, the advantages and disadvantages of scalable DVC and H.264/SVC are analyzed and discussed. The rest of this paper is organized as follows: Section II briefly describes different types of scalabilities in video coding. In Section III, temporal scalability in H.264/SVC is introduced. In Section IV, coding procedures of state-of-the-art DVC is described. Temporal scalablities and complexity of H.264/SVC and DVC are compared

9781–4244–3941–6/09/$25.00

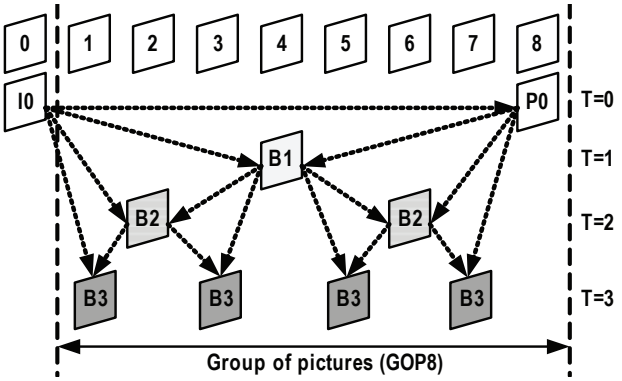

Fig. 1. Temporal scalability scheme.

in Section V.

II. Types of scalability

Scalable extension of the H.264/AVC standard is a highly attractive solution to the problems posed by the characteristics of modern video transmission systems. "Scalability" in this paper means removal of parts of the bit stream to adapt it to the different needs or preferences of end users as well as to the network conditions.

The main idea of scalable coding is that coder forms the bit-stream from several layers: base layer and enhancement layers. The base layer of a bit stream is always coded in compliance with a non-scalable profile of H.264/AVC (single-layer coding). For next enhancement layers encoding the previous layers (that may include the base layer) is needed. Each layer is characterized by its own bit rate and visual quality. Thus, receivers could decode the necessary layers to provide with the necessary bit rate and visual quality.

There exist different ways of the video data processing to form the streams with the properties described above:

- *Temporal scalability.*
- *Spatial scalability.*
- *SNR-scalability.*
- *Combined scalability.*

Spatial scalability and temporal scalability describe cases in which subsets of the bit stream represent the source content with a reduced picture size (spatial resolution) or frame rate (temporal resolution), respectively. With SNR (quality) scalability, the substream provides the same spatial-temporal resolution as the complete bit stream, but with a lower fidelity where fidelity is often informally referred to as signal-to-noise ratio (SNR). The different types of scalability can also be combined, so that a multitude of representations with different spatial-temporal resolutions and bit rates can be supported within a single scalable bit stream [10]. As temporal scalability is the most obvious scalability type, we only focus on this case.

III. Temporal scalability in H.264/SVC

Temporal scalability in H.264/SVC is achieved by using hierarchical coding structures with B-pictures [8]. The pictures of the temporal base layer are only predicted from previous pictures of this layer. The enhancement layer pictures can be bidirectionally predicted by using the two surrounding pictures of a lower temporal layer as references. A picture of the temporal base layer and all temporal refinement pictures between the base layer picture and the previous base layer picture build a group of pictures (GOP). In each GOP, the frame at the lowest level is called the key frame and it is encoded as I- or P-frames. Each temporal layer is marked by an additional identifier T. T is equal to 0 for pictures of the temporal base layer and is increased by 1 from one temporal layer to the next.

Figure. 1 shows an example of building hierarchical B-picture structure for the case of GOP containing 8 frames. In this case base temporal layer $T = 0$ consists of only the single key (frame 8) of this GOP. Next layer $T = 1$ consists of single B-picture (frame 4) that requires two reference frames in forward and backward direction (frame 0, frame 8) from layer $T = 0$. In the same manner B-picture (frame 2) in the layer $T = 2$ also requires two reference frames (frame 0, frame 4) from layers $T = 0$ and $T = 1$ accordingly. The following steps are done in a similar manner. For the implementation of this type of scalability it is necesary to store all 8 frames in the encoder memory. This brings additional delay and increases the size of the memory used. Therefore, if it is needed to decrease the memory size and delay, temporal scalability could be used in low-delayed mode. However, this will lead to a efficiency degradation.

Temporal scalable bit-stream can be generated by using hierarchical prediction structures without any changes to H.264/MPEG4-AVC. The encoding process for each frame includes the following operations:

1) Inter-frame prediction
   - Motion estimation (4x4, 4x8, 8x4, 8x8, 8x16,16x8,16x16 inter-block search). For each block in the current frame it is necessary to make the search for the most similar block in the previous frame(s).
   - Motion compensation. This means the difference calculation between the current block and blocks found in the reference frame(s).

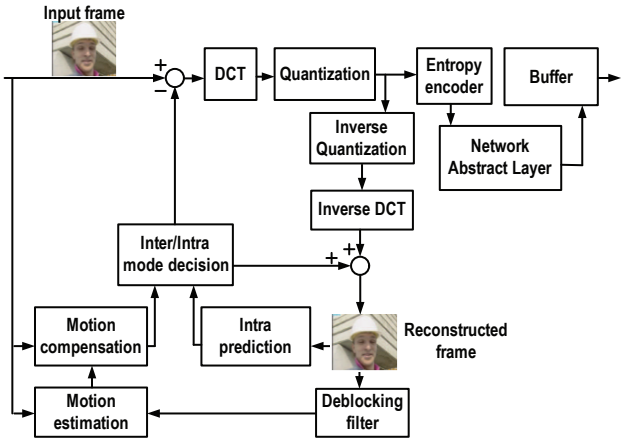

Fig. 2. Main operarions for SVC encoding [2]

2) Intra-frame prediction (DC-prediction, Vertical, Horizontal, Diagonal and others predictions). The prediction of the current block is done by the pixels of the left and upper blocks.
3) Deblocking filter. This filter is used to remove the blocking effect for the improvement of the motion compensation.
4) Discrete Cosine Transform (4x4 DCT, 8x8 DCT, 16x16 DCT).
5) Scalar quantization.
6) Intra/Inter prediction mode decision - in this stage the best prediction mode and best DCT type for current macroblock is chosen.
7) Entropy encoding
   - CABAC (Context-Adaptive Binary Arithmetic Coder);
   - CAVLC (Context-Adaptive Variable Length Coder).
8) Network Abstract Layer - it forms a H.264/SVC compatible stream which can be transmitted over any network.

The decoding process using H.264/SVC includes:

1) Entropy decoding (depending on what was used for encoding).
2) Scalar dequantization.
3) Inverse Discrete Cosine Transform.
4) Motion compensation.
5) Error resielence algorithm - describes what to do if some part of the bit stream was damaged.
6) Deblocking filter.

IV. DVC AND ITS TEMPORAL SCALABILITY

Feedback channel based transform domain Wyner-Ziv video coding is one DVC approach. The architecture of transform domain Wyner-Ziv video codec [6] is depicted in Fig. 3. The encoding procedure includes the following main operations:

1) A fixed Group of Pictures (GOP=*N*) is adopted to split video sequences into two kinds of frames, i.e. Key frames and Wyner-Ziv frames. Periodically one frame out of N in the video sequence is named as key frame and intermediate frames are WZ frames. The key frames are Intra coded by using a conventional video coding solution such as H.264/AVC Intra while the Wyner-Ziv frames are coded using a Wyner-Ziv video coding approach.
2) Each Wyner-Ziv frame X_i is partitioned into non-overlapped 4×4 blocks and a DCT [2] is applied to each of them.
3) The transform coefficients within a given band $b_k, k \in \{0...15\}$, are grouped together and then quantized. DC coefficients are uniformly scalar quantized and AC coefficients are dead zone quantized, respectively.
4) After quantization, the coefficients are binarized. The binary bits with the same significance are formed to a bitplane, which is given to a rate compatible Low Density Parity Check Accumulate (LDPCA) encoder [12]. Starting from the most significant bitplane, each bitplane is independently encoded by the LDPCA encoder, the corresponding accumulated syndrome is stored in a buffer together with an 8-bit Cyclic Redundancy Check (CRC). The amount of transmitted bits depends on the requests made by the decoder through a feedback channel. Although latency is introduced by a feedback channel, encoder complexity can be minimized with this feedback channel based rate control mechanism.

The decoding procedure is described as follows:

1) A side information frame Y_i and its corresponding noise residual frame R are created in the side information generation module [13] by using previously decoded frames. The side information frame Y_i is seen as a 'noise' version of the encoded Wyner-Ziv frame X_i, the estimated noise residual frame R is utilized to express the correlation noise between the Wyner-Ziv frame X_i and the side information frame Y_i.
2) The estimated noise residual frame R and side information frame Y undergo the DCT to obtain the coefficients C_R and C_Y. Taking C_R and C_Y as inputs of a noise model module [15], the noise distribution between corresponding frequency bands

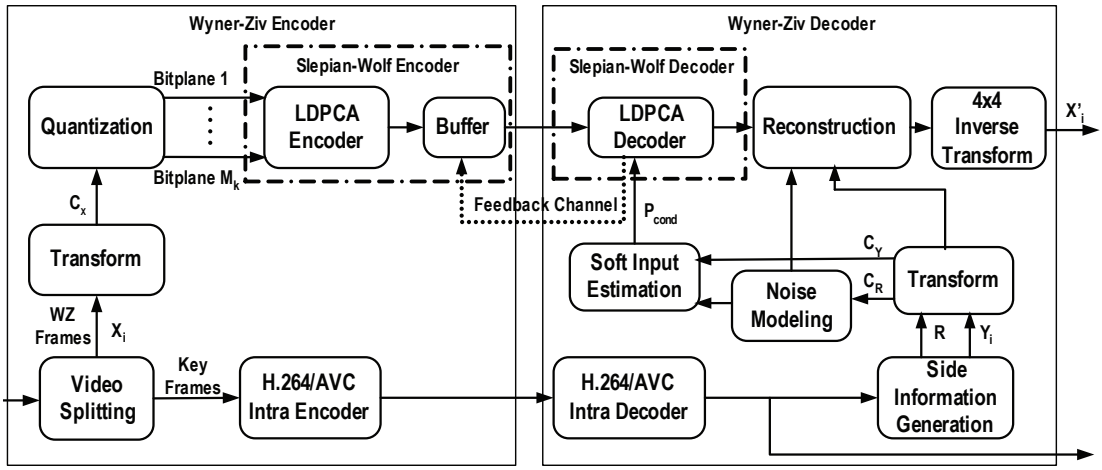

Fig. 3. Feedback channel based transform domain Wyner-Ziv video codec architecture

of the side information frame Y_i and the Wyner-Ziv frame X_i is modeled.

3) Using a modeled noise distribution, the coefficient values of the side information frame C_Y and the previous successfully decoded bitplanes, soft-input P_{cond} (conditional bit probabilities) for each bitplane is calculated.
4) With the obtained soft-input P_{cond}, the LDPCA decoder starts to process various bitplanes to correct bit errors. Convergence is tested by the 8-bit CRC sum and the Hamming distance. If the Hamming distance is different from zero or the CRC sum is incorrect after a certain amount of iterations, the LDPC decoder requests more accumulated syndrome bits from the encoder buffer via the feedback channel to correct the existing bit errors. If both the Hamming distance and CRC sum are satisfied, convergence is declared, guaranteeing a very low error probability for the decoded bitplane.
5) After successful LDPCA decoding, the obtained bitplanes are grouped together to form a set of decoded quantization symbols for each band b_k. With the received quantization information, the decoded quantized symbols are used to calculate the correct intervals in which the Wyner-Ziv coefficients are located. Together with side information coefficients C_Y, noise distribution parameter α and the interval information, decoded coefficients within band b_k of the Wyner-Ziv frame are reconstructed.
6) After all the coefficients bands are reconstructed, 4×4 block inverse transform is performed to obtain the reconstructed Wyner-Ziv frame X_i'.

Compared with the DISCOVER DVC codec in [14], the novelty of the implemented DVC codec is combining an improved Overlapped Block Motion Compensation (OBMC) based side information generation module [13] and an adaptive virtual channel noise model module [15]. Beside the novel DVC aspects, our DVC implementation is also extended with temporal scalability in this paper according to GOP size 8 example as shown in Fig. 1. Each temporal layer in DVC can be encoded independently without storing any reference frames. The temporal layer $T = 0$ consists of the H.264/AVC Intra coded frames (frame 0 and frame 8), while the other layers are Wyner-Ziv coded frames. During the decoding, Wyner-Ziv frames in the next layer $T = 1$ (frame 4) needs two previous decoded key frames in forward and backward direction (frame 0, frame 8) from layer $T = 0$ for decoding. Similarly, Wyner-Ziv frames (frame 2) in layer $T = 2$ utilize two frames (frame 0, frame 4) from layers $T = 0$ and $T = 1$ for decoding.

Due to the feedback channel based rate control mechanism in our DVC implementation, the *coded data* (e.g. the coded frame 4) in higher layers still needs to be stored in a buffer before lower layer frames (e.g. frame 0 and frame 8) is successfully decoded. The size of *coded data* to be stored may be equivalent of up to 1.5 frames with simple rate control [16]. Ideally, if an efficient rate control is employed, it may be possible to avoid store these data for the realization of temporal scalability in DVC.

V. Comparison of Scalability Performance

In order to make fair scalability comparisons between DVC and H.264/SVC, the Joint Scalable Video Model (JSVM) reference software v.9.15 [9] which has processed video stream in temporal scalable mode is used. It is important to note, that the comparison was made for reduced encoder complexity. SVC worked in the

Intra mode without memory for the frames. For this the most complicated blocks were turned off (e.g. motion estimation). In the differential coding mode the encoding complexity was also minimal but the additional memory for the frames was needed. The test conditions adopted in this paper are the DISCOVER project test conditions, commonly used in the DVC literature [13][14]. The test sequences "hall monitor" and "coastguard" are coded at QCIF, 15 frames per second (fps). The key frames are encoded using H.264/AVC Intra and the QPs are chosen so that the average PSNR (Peak Signal-to-Noise Ratio) of the WZ frames is similar to the average PSNR of the key frames. The RD performance is evaluated for the luminance component of the key frames, WZ frames and hierarchical B frames. GOP consists of 8 frames: IWWWWWWWI for Wyner-Ziv encoding and IBBBBBBBI for SVC encoding (taking into account that I frames were encoded in a similar manner). The temporal scalability results are shown in Figs. 4–9.

TABLE I

COMPLEXITY AND MEMORY SIZE COMPARISON FOR ENCODER FOR GOP8

Encoder type	Computation complexity	Memory
H.264/SVC Intra	Intra prediction, DCT, Quantization, Entropy encoding, IDCT, Dequantization	Less than 1 frame
DVC	DCT, Quantization, LDPCA encoder, CRC	Equivalent to 1 frame
H.264/SVC Differential frame coding	Inter/Intra mode decision, DCT, Quantization, Entropy encoding, IDCT, Dequantization	more than 8 frames

If there are no restrictions on the complexity of the decoder, then as shown in Figs. 4–9 the use of state-of-the-art DVC is preferable in the case when we need to have minimal memory and encoder complexity. The efficiency of state-of-the-art DVC is better than SVC for the same memory size and complexity. If the size of the memory at the encoder is not limited, the H.264/SVC has better results (see Table I). If there is a limitation on the encoder complexity then simplified H.264/SVC (e.g. without motion compensation) is better.

In this work, we evaluate temporal scalability. It is straightforward possible for our DVC scheme to provide SNR-scalability by selecting of bitplanes. Furthermore the basic layer (or key frames) could be lower resolution, thus also providing spatial scalability. The choice of scalability using DVC may be made without changing the DVC encoder, thus only the decoder needs to be modified.

VI. CONCLUSION

The efficiency of the temporal scalability of state-of-the-art DVC and SVC with reduced complexity encoding are discussed in this paper. If there are the strong restrictions on the encoder memory and complexity then only H.264 in the Intra-frme mode can provide temporal scalability. If the encoder memory is close to one frame and we have complexity restrictions at the encoder then DVC shows better results. If there are no encoder memory restrictions, but only restriction for the complexity, it is better to use H.264 in the Differential Frame Coding mode. Thus, it is shown that with the encoder memory restrictions and availability of the temporal scalability the best method of the encoding should be chosen taking into account the memory restrictions Due to the existing performance gap, it is necessary to further improve the coding efficiency of DVC. The minimization of the encoder complexity overhead for scalable coding without sacrificing coding efficiency has become an active research area in the video coding community. As a continuation of this work in the future, additional research for spatial and SNR scalability will be conducted.

REFERENCES

[1] International Organization for Standardization, "Introduction to SVC Extension of Advanced Video Coding", ISO/IEC JTC1/SC29/WG11, International Organization for Standardization, Coding of Moving Pictures and Audio, Pozna , Poland, July 2005. URL: http://www.chiariglione.org/mpeg/technologies/mp04-svc/svc/.

[2] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC Video Coding Standard", IEEE Trans. on Circuits and Systems for Video Technology, vol. 13, no. 7, July 2003.

[3] B. Girod, A. Aaron, S. Rane and D. Rebollo-Monedero, "Distributed Video Coding". Proceedings of the IEEE, vol. 93, no. 1, pp. 71-83, January 2005.

[4] J. Slepian and J. Wolf, "Noiseless coding of correlated information sources," IEEE Trans on Inf. Theory, vol. 19, no. 4, pp. 471-480, Jul 1973.

[5] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," IEEE Trans on Inf. Theory, vol. 2, no. 1, pp. 1-10, Jan 1976.

[6] A. Aaron, R. Zhang, and B. Girod, "Transform-domainWyner-Ziv codec for video," In Proc. SPIE Visual Com. and Img. Proc., vol. 5308, pp. 520-528, January 2004.

[7] H. Wang, N. M. Cheung, and A. Ortega, "A framework for adaptive scalable video coding using Wyner-Ziv techniques," EURASIP Journal on Applied Signal Proc., pp. 1-18, 2006.

[8] S. Lim, J. Yang, and B. Jeon,"Fast Coding Mode Decision for Scalable Video Coding", *10th Int'l Conf. on Advanced Communication Technology*, vol. 3, pp. 1897-1900, 2008.

[9] JSVM 9.15 software package, CVS server for the JSVM software. http://iphome.hhi.de/

[10] H. Schwarz, D. Marpe, and T. Wiegand "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

[11] E.Belyaev, V. Grinko, A. Ukhanova, "Power Saving Control for the Mobile Receivers in the DVB-H based on the Scalable Extension of H.264/AVC Standard" *Wireless Telecommunications Symposium*, 2009

[12] D. Varodayan, A. Aaron, and B. Girod, "Rate-adaptive distributed source coding using low-density parity-check codes," *EURASIP Signal Process. Journal, Special Section on Distributed Source Coding*, vol. 86, pp. 3123-3130, Nov. 2006.

[13] X. Huang, and S. Forchhammer, "Improved side information generation for distributed video coding," *IEEE Int'l Workshop om Multimedia Signal Processing*, pp. 223-228, Oct. 2008.

[14] Available on: www.discoverdvc.org.

[15] X. Huang, and S. Forchhammer, "Improved virtual channel noise model for transform domain Wyner-Ziv video coding," *IEEE Int'l Conf. on Acoustics, Speech and Signal Processing*, pp. 921-924, April 2009.

[16] M. Morbee, J. Prades-Nebot, A. Pizurica, and W. Philips, "Rate allocation algorithm for pixel-domain distributed video coding without feedback channel," *IEEE Int'l Conference on Acoustics, Speech and Signal Processing* , pp. 521-524, April 2007.

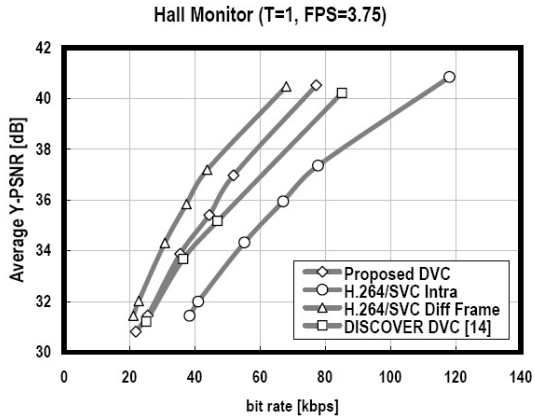

Fig. 4. RD comparison for SVC and DVC, "hall" (T = 1)

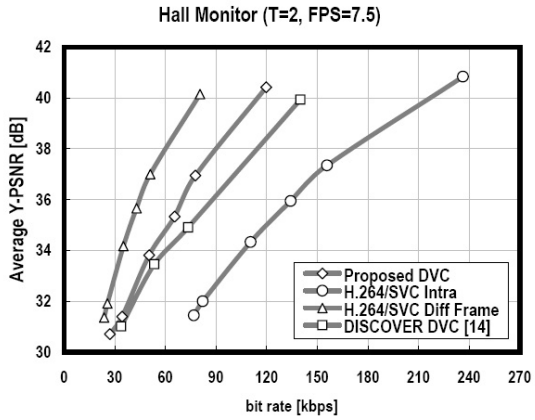

Fig. 5. RD comparison for SVC and DVC, "hall" (T = 2)

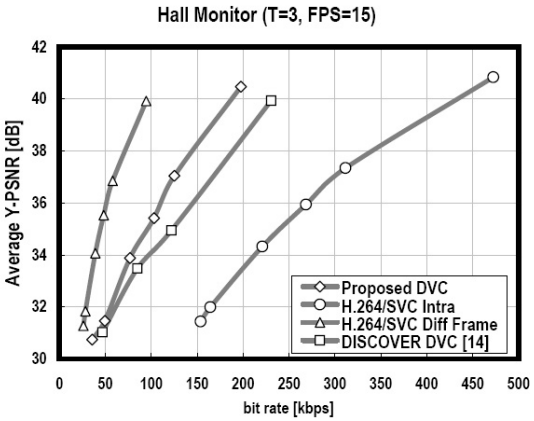

Fig. 6. RD comparison for SVC and DVC, "hall" (T = 3)

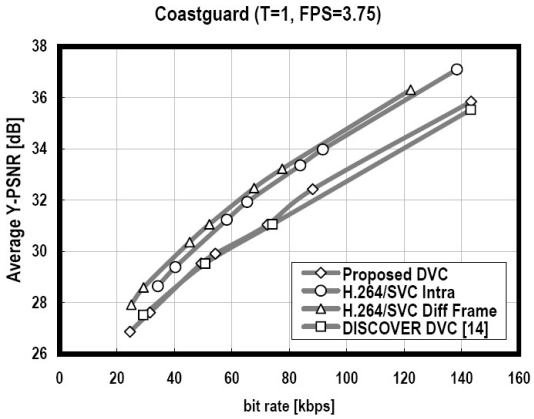

Fig. 7. RD comparison for SVC and DVC, "coastguard" (T = 1)

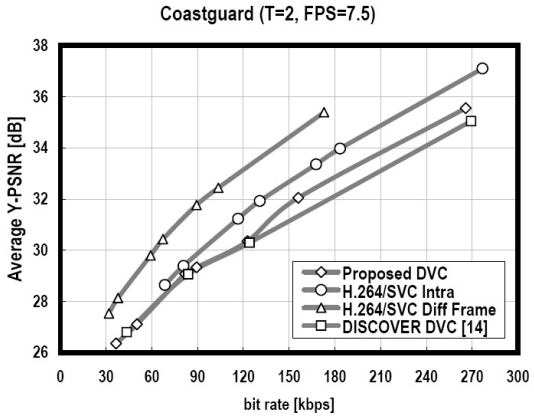

Fig. 8. RD comparison for SVC and DVC, "coastguard" (T = 2)

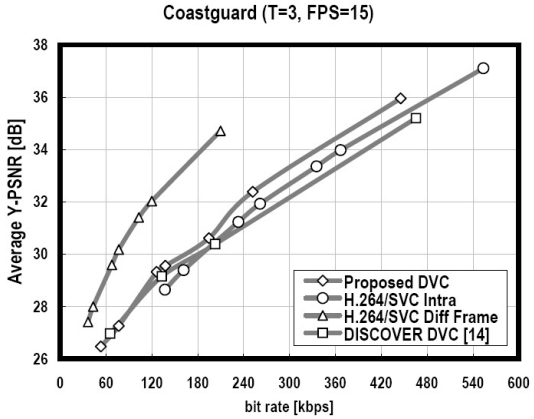

Fig. 9. RD comparison for SVC and DVC, "coastguard" (T = 3)