Support for Digital VCR functionality over Network for H.264/AVC

Pengpeng Ni
Department of Informatics
The University of Oslo, Norway
pengpeng@ifi.uio.no

Damir Isovic
School of Innovation, Design and Engineering
Mälardalen University, Sweden
damir.isovic@mdh.se

Abstract

In this paper, we propose a scalable video coding scheme for H.264/AVC video that fully utilizes the benefits of Switching frames (SP/SI) to facilitate user-friendly remote multimedia stream browsing and editing. A content aware rate-control algorithm is developed to adaptively determine the optimal encoding configuration for balancing the trade-off between video coding efficiency and temporal scalability. Our temporal scalable coding scheme is useful for implementing DVCR-like interactive functionality such as fast-forwarding and reversing in multimedia streaming applications such as Video-On-Demand, video broadcasting and remote video editing.

I. Introduction

Digital Video Cassette Recording (DVCR) is a key technique to support user-friendly browsing of multimedia contents. The set of full DVCR functionality includes forward, backward, step-forward, step-backward, fast-forward, fast-backward, random access, pause, and stop operations. It allows the user to fully control the presentation of the multimedia content, supporting the applications such as video editing.

However, the realization of full DVCR functionality over network in applications with highly compressed digital multimedia streams, is challenging and not yet well resolved. The higher compression efficiency comes at the expense of higher computational requirement, which becomes a problem when the available decoding time for frames decreases due to a DVCR trick-mode.

Conveniently, the current state-of-the-art of MPEG coding [1], the standard H.264/MPEG-4 AVC, contains several new encoding features to support quick browsing. Among them, the new frame types, SI- and SP-frames can be used to implement DVCR functionality in streaming applications. One important property of SP-frames is that identical frames can be reconstructed from different reference pictures. Moreover, SP-frames have significantly better coding efficiency than I-frames while providing similar functionality [2]. Hence, they can replace I- and P-frames in applications to facilitate stream switching, splicing, error resilience and DVCR like functionality [3].

In this paper, we propose an approach for enabling the DVCR functionality in H.264 streaming applications by transforming the video streams to compliant bitstreams with user friendly syntax. We modify a subset of I- and P-frames in the original stream to primary SP-frames suitable for quick browsing, and create secondary SP/SI-frames to support different speed-ups at the decoder side. The transcoding is done in such a way that the target bit rate does not exceed a constant value which can be known in advance (e.g., the bitrate of the input stream). Given the target bit rate, we minimize the distortion by using different settings for transcoding parameters for extra SP-frames, and apply rate-distortion theory to find the optimized combination of the encoding options.

More specifically, we propose an adaptive frame-layer rate control for S-frames. The focus is how to minimize the variation of decoding latency while maintaining nearly constant quality among video frames in a video sequence. To achieve this, we jointly consider different coding factors, such as bit rate, distortion and content related complexity, when encoding video frames. We propose a rate control algorithm for S-frames based on quadratic rate distortion model that consists of a two phase encoding scheme, frame-layer bit allocation and quantization parameter estimation. We underline the effectiveness of our approach through a set of experiments.

II. Related work, motivation and approach

Several reverse-play transcoding algorithms that utilize the motion vectors to predict P- and B-frames backwards

have been presented and compared in [4], [5]. However, they are based on MPEG-2 and do not suit for current state-of-art standard which takes advantage of multiple reference pictures. Additionally, the problem of extra network traffic caused by fast-playback operation is not solved as well. In [6] a solution for DVCR with dual bitstreams has been presented. The idea is to offline re-encode an original MPEG video sequence in the reverse order to generate a secondary bitstream. This solution does not enable none-drift reconstruction for bitstream switching and it has large storage requirement.

H.264/AVC is a flexible coding standard with a potential to meet the manifold needs of applications. One of the key features is a new frame type, S-frames [2], [7], which leads to a significant increase in streaming flexibility. There are *primary* SP-frames, which use similar forward motion prediction technique as P-frames, and *secondary* SP- and SI-frames, which can be inter- and intra-coded. Each secondary frame is bounded to a primary SP-frame, and function as a substitute when functionalities like bitstream switching or random access are required. The usage of SP/SI-frames is exploited in different application scenarios. The most discussed scenarios are error resilience and bitstream switching, see [2], [8], [9] for some examples. However, the flexibility provided by S-frames comes at the price of decreased video coding efficiency. The performance impact need to be effectively controlled for dedicated application scenarios. In our current research, the target application scenario is quick browsing of remote video content by means of full DVCR functionality. With S-frame, we want to extend the temporal scalability of video streams so that efficient jump is enabled between video frames within a single stream.

Our contribution is a self-adaptive controlling mechanism for H.264/AVC encoder/transcoder. The aim is to find the optimal coding configuration for H.264/AVC bit stream with enhanced temporal scalability. Given the attractive features of S-frames, our optimal coding scheme is based on a multi-layered video structure. Different than standard single layer IPP or IBP GOP structure, a subset I- and P- frames are replaced by a set of S-frames. The primary SP frames belong still to the base layer while secondary S frames form enhancement layers. In the enhancement layer, secondary SP-frames are used for forward and backward DVCR modes, and SI-frames serve as random access points. We introduce our multi-layered encoding scheme in the next section.

III. Multi-layered Video Sequence

As what illustrated in figure 1, we encode video into multiple layered sequence consisting by one base layer and two enhancement layers. The base layer consists of regular P and primary SP frames. For each primary frame, SP, we generate two secondary SP-frames, $SP^{'}$ and $SP^{''}$, to form two enhancement layers respectively. Those secondary SP-frames are referenced by the nearest S-frames either on their left or right side while the secondary SI-frames do not use any references. In figure 1, the coding dependency between frames is illustrated by the arrows where the dashed arrows imply the coding bound between a primary and a secondary SP-frame. For the illustration simplicity, we neglect the B-frames.

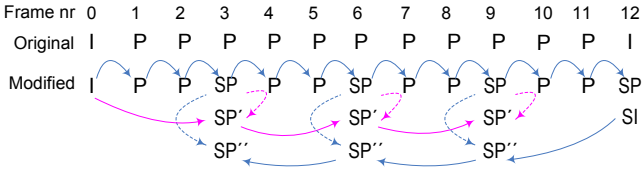

Fig. 1. Encoding to support DVCR

The S-frames in multiple layers are used jointly to support fast-forward and fast-backward playback modes. Assume, for example, a backward operation requests on a remote stream (e.g., the one from figure 1), as depicted in figure 2. Suppose that the currently decoded frame is P_8. Once when P_8 has been decoded and displayed, its reference frame P_7 is discarded from the frame buffers, which means that it must be decoded once again. In the original stream, the only way to decode P_7 is to decode the entire chain of reference frames starting with the I-frame (P_7 needs P_6 which needs P_5 which needs P_4 and so on). In the modified stream, however, P_7 can be decoded from the primary SP_6 frame, or from some of its secondary frames $SP_6^{'}$ or $SP_6^{''}$. So, the client sends the video segment $\{SP_9, SP_6^{''}, P_7\}$, compared to 7 frames in the original stream.

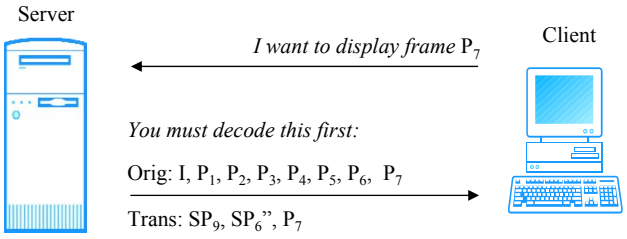

Fig. 2. Remote backward playback.

IV. Adaptive Encoding Scheme

In our encoding scheme, all I-frames in a video sequence, except the first one, can be replaced by primary SP-frames. Since an SP-frame is generally smaller than

an I-frame, bits can be saved and evenly allocated to the other inter-predicted frames. Notice that the replacement of an I-frame by a primary SP-frame makes most sense when the predecessors P-frame is similar to the replaced I-frame, which usually is the case in a video sequence. In other cases, such as at video shot boundaries, the difference between frames is most likely to be quite large, so we encode the starting pictures of a video shot as I-frames.

In video sequences encoded in traditional way, random access is enabled only by I-frames. This puts the limit of the maximum number of P-frames which can be anchored by I frames. In our encoding scheme, bounded to the primary SP-frames, secondary SI-frames can be adaptively generated based on the stream demands and random access requirement. Hence, a fixed and limited GOP length is not necessarily preserved. The GOP length can be theoretically extended as long as the video sequence lasts.

The periodicity of primary SP-frames decides the amount of switching points added into a video stream. It is noticeable that the number of frames between two adjacent primary SP-frames, Δn, has major effect to the decoding time of normal reverse playback operation. The larger Δn, the longer the decoding time. Combining the choice of suitable value for Δn with multiple reference picture control (see [3]), the decoding time jitter of backward playback can be minimized. Moreover, smaller value of Δn gives finer grained temporal scalability. The encoder can produce fast forwarding $SP^{'}$ or $SP^{''}$ built upon any pair of switching points, hence enable fast forwarding with different speed up factor.

The encoding process of S involves two quantization steps. We illustrate it briefly in Figure 3 (abstracted from the design introduced in [7]).

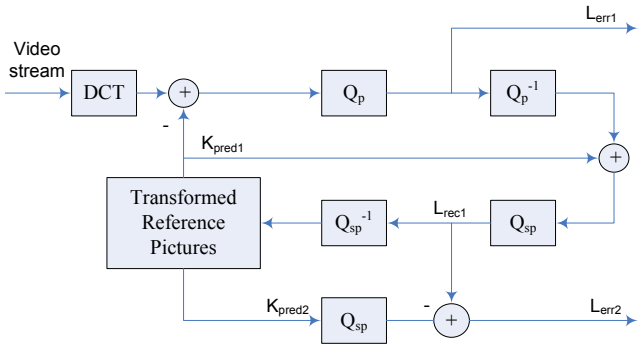

Fig. 3. Simplified SP-frame Encoder.

K_{pred1} and K_{pred2} are the prediction coefficients used for inter prediction. They are generated separately using different reference pictures (note that the coefficients are already transformed prior inter-prediction). For example, to encode primary SP-frame, K_{pred1} is subtracted from the original video frame to get the difference between the predicted picture and its reference picture. The difference is then quantized using the parameter Q_p. The quantization step will cause some quality loss. We denote the quantized error level by L_{err1} which is eventually entropy encoded to form the compressed bitstream of primary SP-frame. The encoding process is so far similar to that of a traditional P-frame. However, the difference is made by an additional quantization/dequantization step in the reconstruction procedure (Q_{sp},Q_{sp}^{-1} in the figure 3). L_{rec1} denotes the quantized coefficients of the video frame reconstructed from primary SP frame. It is compared with K_{pred2} to generate L_{err2} which represents the secondary SP-frame prior entropy encoding. Note that K_{pred2} needs also to be quantized by the same quantization parameter Q_{sp} before the substraction. Identical reconstruction can be obtained again at the decoder no matter whether the decoder receives primary or secondary SP frame.

Looking through the encoding process, we notice that the additional quantization may not affect direct the size of the primary SP-frame, and the identical reconstruction of the primary SP-frame using secondary SP-frame is also guaranteed regardless the value of Q_{sp}. However, the extra quantization decrease the coding efficiency of primary SP-frame and in its turn affects the encoding of its following video frames. A coarser Q_{sp} may lead to poorer prediction for the later encoded video frames, while a finer Q_{sp} gives larger secondary SP- and SI-frame size. These parameters can be adaptively changed so that bits allocated for a GOP can be evenly or strategically distributed among video segments separated by primary SP-frames, which we describe next.

V. Frame-layer rate control for S-frames

Pursuing optimal rate distortion performance is one of the main goals of an encoder. It is highly desired to find the optimal/suboptimal value for these configuration parameters. However, the optimization can be a very complex problem which sometimes is many times more complex than the video coding process itself. In addition, the choices of optimization strategy are usually application specific and dependent on the applications and encoding environments.

Our focus is how to minimize the variation of decoding latency while maintaining highest possible quality among video frames in a video sequence. Among the factors which may influence the coding latency, such as network bandwidth, available processing power, and the available memory, the network bandwidth is a bottleneck given its large range of variation. Hence, the total number of bits for every frame in a video sequence need to be controlled in an optimized way so that the video stream can be

self adaptive to the variations of network condition while avoiding quality degradation and abrupt quality variation.

A. Quadratic rate distortion model

Bit rate, distortion and content related complexity are three inter-related factors to be jointly considered in the concerns of coding performance and resource requirements of a video codec. Here, bit rate *R* refers to the total amount bit used to encode a single video frame, it defers to the available network bandwidth. However, for a video sequence with the average bit rate equal to the available bandwidth of *U*, we have $R = \frac{U}{F}$, where F refers to the frame rate of the video. Distortion refers to an objective video quality measure. It measures the quality loss introduced by the quantization technique, so that it is usually correlated to the quantization parameters *Q*. Finally, the content related complexity refers to the scene content characteristic. A typical measure of the complexity is mean average difference (MAD) between a video frame and its reference picture.

There are many theoretical rate distortion models. The most widely applied R-D model is a quadratic rate-quantization model expressed as.

$$\frac{R}{MAD} = a_1 \times Q_p^{-1} + a_2 \times Q_p^{-2} \tag{1}$$

where *a1* and *a2* are the first and the second-order model parameters; the distortion measure is represented by the quantization parameter Q_p of a video frame. This R-D model gives the theoretical foundation of a bit allocation scheme with rate-distortion optimum concern. For example, a frame layer rate control scheme can allocate a bit budget to each frame respect to their encoding complexity. In order to maintain constant video quality, more bits need usually to be allocated to video frames showing higher complexity, such as scenes containing actions and complex details. Once the target number bits for a frame is determined, the frame QP to meet the target bit allocation is determined according to the rate-distortion model.

B. Multi-layered rate control algorithm

In our approach, the two R-D model parameters $a1$ and $a2$ are justified by actual output bit-rate after encoding each frame. New values of the model parameters are derived using regression technique from the encoding history with the intension to minimize the deviation between the actual bits generated and the allocated bits. As long as more actual encoding results are collected, the accuracy of the R-D model is improved. We summarize our rate control procedure as follows.

- Initialization, the first two video frames are encoded using a fixed quantization parameter, *a1* and *a2* are initialized.
- Pre-encoding, the target bit rate is computed respect to different factors, such as the remaining available bits, the buffer status.
- Quantiser step size Q for a video frame is computed, the value of Q is found by solving the R-D model equation.
- The frame is encoded.
- The model parameters are updated based on the actual number of bits generated for the previously encoded video frames.

Under this rate control framework, algorithms with different bit allocation strategies can be developed. Our encoding scheme conforms also to the recommended framework, but we apply bit allocation strategies separated to different encoding layers.

As the figure 1 illustrates, our target video sequence is composed by three layers, one base layer and two enhancement layers with scaled temporal scalability. The secondary SP frames in enhancement layer belong to the same movie represented by base layer but being apart in time-line. We can actually regard the enhancement layers as two independent bitstreams with faster motions comparing to the base layer bitstream. Therefore, it should be able to apply rate control algorithm separately to each stand-alone bit-stream. However, to make the average video quality of separated video coding layers close to each other, the target bit-rate of the enhanced bitstream needs be higher than the target bit-rate of base layer given higher complexity. Hence, our rate control scheme uses two input parameters to represent the target bit-rates of base layer and enhancement layer respectively. Note, it can be assumed that the target bit-rates of the two enhancement layer being the same, give the video structure shown in 1.

$$\begin{cases} R_r = \frac{B_{low}}{Fr} * N_r \\ \\ \hat{R}_r = \frac{B_{upp}}{Fr} * \hat{N}_r \end{cases} \tag{2}$$

$$\begin{cases} \frac{R(i)}{MAD(i)} = a_1 \times Q_p^{-1}(i) + a_2 \times Q_p^{-2}(i) \\ \\ \frac{\hat{R}(i)}{\widehat{MAD}(i)} = b_1 \times Q_{sp}^{-1}(i) + b_2 \times Q_{sp}^{-2}(i) \end{cases} \tag{3}$$

- B_{low}, B_{upp}, the target average bit-rates of the base layer bit stream and enhancement layer bit stream.
- N_r, $\hat{N}_r$, the total number of video frames at base layer and enhancement layer. For base layer, N_r refer to the number of I, P and primary SP frames. For enhancement layer, $\hat{N}_r$ counts only on I frame and secondary SP in one single direction ($SP^{'}s$ or $SP^{''}s$).

- R_r,$\hat{R}_r$, the total bit budgets available for encoding the base-layer and one enhancement layer.
- Fr, the frame rate of the video.
- $MAD, \widehat{MAD}$, the content related complexities of frames in separate layers, $\widehat{MAD}$ is estimated by the difference of pairs of video frames apart by the distance of Δn.

The two input parameters, B_{low} and B_{upp}, indicate essentially a bounded range of bandwidth requirement to a network. By introducing the upper bound of the bit-rate requirement, we intend to limit the throughput variations of the streamed data. We believe this will help the network to make more effective resource planning and traffic control. For example, the network may adaptively control the accepted number of concurrent streams respect to their upper band of bandwidth requirement and provide them higher QoS guarantee in the form of smoother VCR functionality, meanwhile providing normal streaming services in a more common sense. Within each layer, we apply the same bit allocation strategy in [10]. The target bits allocated to each video frame is estimated jointly based on the current available bits and buffer fullness.

On the other hand, we need still consider the quality loss issue. Severe degradation of quality distortion is by all means not acceptable. However, the flexibility introduced by the S frames comes at the price of decreased rate distortion performance. We need limit the distortion to an extend so that abrupt quality degradation will be avoided. We solve this question by strategically set the window size of the allowed quantization parameter value for each frame so that the maximum changes of video quality between two consequent P- and/or SP-frames can be controlled. Further, it is noticeable that the P frame direct after a SP-frame will more likely become large considering its coarse reference picture. Moveover, the probability of a P-frame being dropped is related to its distance to its previous S-frame. Bandwidth saving will be more effective if we allocate more bits to frames with higher dropping probability.

Summarizing above, we limit the difference between the two quantization parameters used to encode a set of S frame to a constant value, Δq_{max}. The value of Δq_{max} can also be derived from B_{upp}, depending on system setup and user preference. In our experiment, we set it equal to 4 and $B_{upp} = 3 \cdot B_{low}$. By Δq_{max}, we try to prevent potential severe quality loss due to the limited available bits budget allocated to the enhancement layer.

For each frames in base layer, we shift the window of the possible parameter value as $[\tilde{Q}p(k-1) - \Delta ql, \tilde{Q}p + \Delta qr]$, where $\tilde{Q}p$ equals to Q_p for regular P frames; for primary SP frame, $\tilde{Q}p = \min\{Q_p + \frac{\Delta q_{max}}{2}, Q_{sp}\}$ is chosen. Δql and Δqr are derived from the current distance to the previous S frame. For $\Delta q_{max} = 4$, we use a simple way to calculate them as follows.

$$\begin{cases} \Delta ql = \lfloor \frac{k}{\Delta n} * \frac{\Delta q_{max}}{2} + \xi_1 \rfloor, \xi_1 = 0.4 \\ \Delta qr = \lfloor (1 - \frac{k}{\Delta n}) * \frac{\Delta q_{max}}{2} + \xi_2 \rfloor, \xi_2 = 0.8 \\ 0 \le k \le \Delta n \end{cases} \quad (4)$$

So far, we have introduced a single pass rate control algorithm for quantization parameter decision. In our scheme, another important parameter is the periodicity of S frames Δn. But Δn can only change within a small range, say $1 < \Delta n < 15$. As declared before, smaller Δn is more preferable. In addition, since we are performing off-line encoding, processing time is not a particular constraint. Multiple redundant encoding passes can therefore be employed. Hence, we use dynamic programming strategy to encode the video repeatedly with incremental Δn, starting from 2, i.e [2,4, 6...]. The encoding is done when the quality measure of the encoded video reaches its benchmark value or Δn exceeds its upper bound.

VI. Experiments

Experiments have been performed to verify our coding scheme. We implemented our rate control algorithm based on the H.264/AVC reference software JM 12.2 [11]. In our experiments, two test sequences *Mother and Daughter* and *Foreman* are encoded in CIF format at 30 frames per second. The target bit rate of the base layer is 512 Kbps. For space limitation reasons, we only show results for the *Mother and Daughter* sequence.

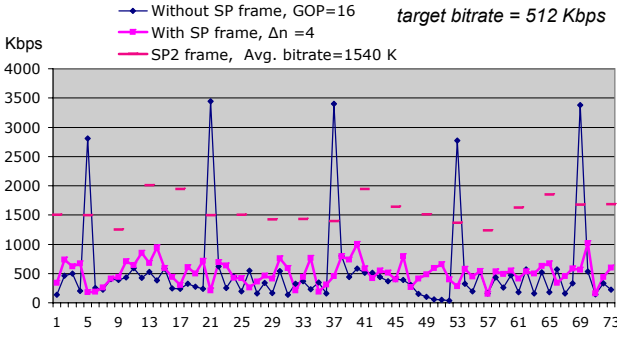

Fig. 4. Bitrate fluctuations for different coding methods

By our rate control algorithm, we intend to decrease the variation of video frame bit rates and narrow the gap between the maximum and minimum bit rate requirements. Figure 4 illustrates the bit-rate fluctuation of the same video encoded by different coding methods. Comparing to

fixed GOP length encoding, our coding scheme allocates bits more evenly among video frames. Meanwhile, the bit rate of enhancement layer is under control.

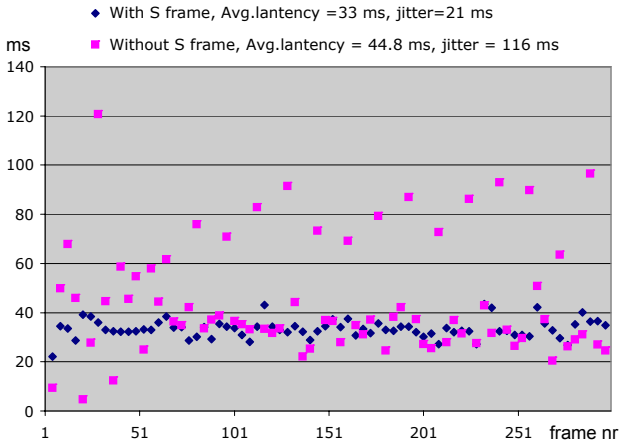

Fig. 5. Transmission delay comparison

The fluctuation of bit rates will mainly influence the end-to-end delay, and may result in end-to-end delay deadline miss. The problem will be intensified in VCR trick mode. Figure 5 shows the transmission time in fast forwarding playback mode. The transmission time is estimated assuming a constant available bandwidth. Although, the processing delay at the decoder side is not accounted in, the benefit of our coding scheme in jitter reduction is still attentive.

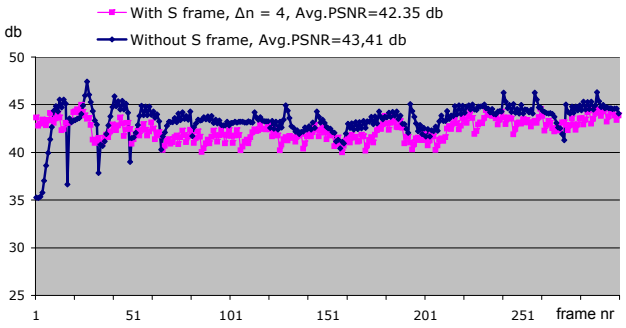

Fig. 6. PSNR comparison

Video quality of the test sequence is measured by taking the average of the peak-signal-to-noise-ratio over all the encoded/decoded frame. As figure 6 shows, the quality change between two consequent video frames is small for both two test sequences. However, the average PSNR of *Mother and Daughter* is about 1 db less than its fixed GOP variant. This is due to the shorter distance between S-frames and reflects the fact that more S-frames brings more distortion. It is believable that by increasing the value of Δn, the rate-distortion performance will get increased. On the other hand, we want to have small Δn to enable finer scalability. As other scalable coding techniques, the rate-distortion performance is still an issue.

VII. Conclusions

In this paper we presented a transcoding scheme for H.264/AVC video to support user-friendly real-time browsing. A sub-set of original I- and P-frames is transformed into SI/SP-frames suitable for quick browsing of local and remote video content. The transcoding parameters can be adaptively changed and we apply rate-distortion theory to find the optimized combination of the encoding options.

Given the target bit-rate we minimize the distortion by using different encoding parameters, such as unlimited GOP length with appropriate periodicity for S-frames and different quantization parameters. We apply rate-distortion theory to find the optimized combination of encoding parameters with respect to the target bit-rate. Our proposed rate-control framework and multi-layered rate control algorithm for S-frames minimizes the variation in the decoding latency while maintaining nearly constant quality among video frames. The effectiveness of our approach is underlined by a set of experiments.

References

[1] "ISO/IEC 13818-2: Information technology - generic coding of moving pictures and associated audio information, part2: Video," 1996.

[2] M. Karczewicz and R. Kurceren, "The sp- and si-frames design for h.264/avc," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, Jul. 2003.

[3] *Advanced Video Coding for Generic Audiovisual services, ITU-T Recommendation H.264*, ITU-T and ISO/IEC JTC 1, Apr. 2004, iSO/IEC 14496-10(AVC).

[4] S. J.Wee, "Reversing motion vector fields," in *Image Processing, 1998.ICIP 98.Proceeding*, vol. 2.

[5] S. J.Wee and B. Vasudev, "Compressed-domain reverse play of mpeg video streams," in *SPIE International Symposium on Voice,Video,and Data Communicatins*, Nov. 1998.

[6] C.-W. Lin, J. Zhou, J. Youn, and M.-T. Sun, "Mpeg video streaming with vcr functionality," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 3, Mar. 2001.

[7] X. Sun, S. Li, F. Wu, J. Shen, and W. Gao, "The improved SP frame coding technique for the jvt standard," in *Image processing, 2003. ICIP 2003 Proceedings. 2003 International Conference on*, vol. 3, 2003.

[8] E. Setton and B. Girod, "Video streaming with sp and si frames," in *Preceedings VCIP,Beijing*, Jul. 2005.

[9] W. tian Tan and G. Cheung, "Sp-frame selection for video streaming over burst-loss networks," in *Proceedings of the Seventh IEEE International Symposium on Multimedia*, 2005.

[10] Y. Wu, L. Shouxun, Z. Yongdong, and L. Haiyong, "Adaptive rate control with hrd consideration," 2005.

[11] J. J. of ISO/IEC MPEG and ITU-T VCEG, "H.264/avc jm reference software," URL: http://iphome.hhi.de/suehring/tml/.