Journal of Electronic Imaging

JElectronicImaging.org

Scene-library-based video coding scheme exploiting long-term temporal correlation

Xuguang Zuo
Lu Yu
Hualong Yu
Jue Mao
Yin Zhao

SPIE. IS&T imaging.org

Xuguang Zuo, Lu Yu, Hualong Yu, Jue Mao, Yin Zhao, "Scene-library-based video coding scheme exploiting long-term temporal correlation," *J. Electron. Imaging* **26**(4), 043026 (2017), doi: 10.1117/1.JEI.26.4.043026.

Scene-library-based video coding scheme exploiting long-term temporal correlation

Xuguang Zuo,[a] Lu Yu,[a,*] Hualong Yu,[a] Jue Mao,[a] and Yin Zhao[b]
[a]Zhejiang University, Institute of Information and Communication Engineering, Hangzhou, China
[b]Huawei Technologies Co., Ltd., Central Research Institute, Hangzhou, China

Abstract. In movies and TV shows, it is common that several scenes repeat alternately. These videos are characterized with the long-term temporal correlation, which can be exploited to improve video coding efficiency. However, in applications supporting random access (RA), a video is typically divided into a number of RA segments (RASs) by RA points (RAPs), and different RASs are coded independently. In such a way, the long-term temporal correlation among RASs with similar scenes cannot be used. We present a scene-library-based video coding scheme for the coding of videos with repeated scenes. First, a compact scene library is built by clustering similar scenes and extracting representative frames in encoding video. Then, the video is coded using a layered scene-library-based coding structure, in which the library frames serve as long-term reference frames. The scene library is not cleared by RAPs so that the long-term temporal correlation between RASs from similar scenes can be exploited. Furthermore, the RAP frames are coded as interframes by only referencing library frames so as to improve coding efficiency while maintaining RA property. Experimental results show that the coding scheme can achieve significant coding gain over state-of-the-art methods. [DOI: 10.1117/1.JEI.26.4.043026]

Keywords: video coding; repeated scene; long-term temporal correlation; scene library; clustering.

Paper 16832 received Oct. 1, 2016; accepted for publication Aug. 3, 2017; published online Aug. 30, 2017.

1 Introduction

In video coding, the key to achieve high video coding efficiency is to make full use of correlations in video. The short-term temporal correlation has been well exploited by current video coding. However, a substantial part of videos is characterized with long-term temporal correlation since they contain scenes that appear repeatedly. For example, in news programs, the scenes of studio and logo clips may emerge at intervals. In talk shows, the images of the hosts, the guests, and the audience repeat alternately. In movies and TV series, many scenes in dialogue clips and flashback episodes appear repeatedly. The video coding efficiency would be highly improved if the long-term correlation is well exploited.

During the development of video coding, the temporal correlation in video has been more and more adequately used owing to the improvement of temporal reference techniques. In early video coding standards, such as MPEG1, H.261, H.262/MPEG2,[1] and H.263,[2] only short-term temporal correlation was exploited because the compression algorithms only made reference to one previous decoded picture. Later, a multiple reference frame technique was introduced by providing a long-term memory that stores seconds of previously decoded frames (up to 50) to the codec so that temporal dependencies in video containing repetitive motion, uncovered background, noninteger pixel displacement, etc. can be exploited.[3,4] The multiple reference frame idea was further developed and adopted in the later video coding standards H.264/Advanced Video Coding[5] and High Efficiency Video Coding (HEVC),[6] in which two kinds of frames, short-term reference (STR) frames and long-term reference (LTR) frames, are stored in the decoded picture buffer and used for motion compensation. STR frames, the neighboring frames of the encoding frame, are used to eliminate short-term temporal correlation while LTR frames, which are from distant past, are employed to make use of long-term temporal correlation. In practical application, the numbers of both STR frames and LTR frames are limited because of: (1) the overhead of syntax to signal reference frames; (2) exhaustive computation complexity cost introduced by motion estimation (ME). Typically, the number of LTR frames is set as one or two. For example, in a well-known codec called VP8,[7] only one LTR frame called golden frame is enabled. In the latest Chinese video coding standard AVS2,[8] a long-term background reference frame is used to assist the coding of surveillance video.

Although LTR frames have been supported by many video coding standards, the long-term temporal correlation in videos with repeated scenes still exists. Previous works always focus on exploiting the long-term temporal correlation in a long scene.[9–15] For example, in Refs. 9–12, a background frame was generated for a surveillance video and used as an LTR frame to improve background prediction. In Refs. 13 and 14, the optimal LTR frame interval was investigated and LTR frames were adaptively selected according to the accumulation of change in a scene. In Ref. 15, the former two ideas were integrated, a background frame was generated for LTR and then updated when more background regions were exposed. However, the method of selecting LTR frames for multiple repeated scenes in a video to exploit their correlation has not been well explored. In addition, random access (RA) is desirable in many applications. It enables seek, fast-forward, and fast-backward operations in locally stored video streams (e.g., DVD, BD, etc.). In video-on-demand (VOD) streaming, it allows the servers

*Address all correspondence to: Lu Yu, E-mail: yul@zju.edu.cn

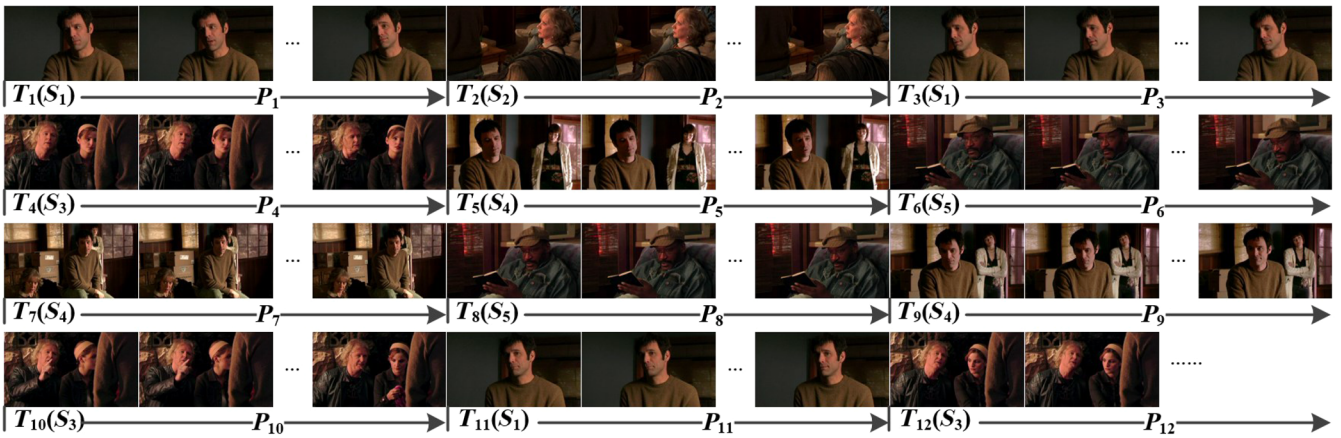

Fig. 1 An example video with repeated scenes.

to respond to client requests. To support RA functionality, the encoding video is divided into independent RA segments (RASs) by RA points (RAPs) and the correlation between RASs cannot be exploited. Moreover, the RAP frames are all intracoded. An intraframe typically requires several times more bits compared to an interframe with the same quality since it takes no advantage of temporal correlation. Because of above reasons, the rate-distortion performance of current video coding is degraded.

In Ref. 16, we proposed a method for the coding of RAP frames in video with repeated scenes. The representative frames of the video were extracted and stored in a scene library in advance. Then, the RAP frames were coded by only referencing the scene library. The method in Ref. 16 can improve the coding efficiency of RAP frames and somewhat employ the correlation between repeated scenes. However, there are still some problems. First, the construction of scene library was based on a front-to-back analysis of the video. The first frame of a nonrepetitive scene was certainly chosen as library frame, which might not be efficient for reference. Second, the method did not take advantage of the scene library to optimize the coding of non-RAP frames, thus the correlation between RASs was not fully used.

To make the best use of long-term temporal correlation in videos containing repeated scenes, a scene-library-based video coding (SLBVC) scheme is proposed in this paper. The main contributions of this study are as follows: (1) A video coding framework, in which the reference frames are stored in a scene library not cleared by RAPs, is introduced to exploit the correlations between RASs, as well as similar scenes. (2) A method to build the scene library is designed. The method is based on clustering and is able to construct a compact scene library to facilitate the video coding efficiency. (3) A layered coding structure based on the scene library is proposed for the coding of each RAS. In the layered coding structure, the library frames are intracoded. Each RAS is coded by referencing only one library frame to improve coding efficiency while preserving the RA capability. The proposed scheme is suited to stored video and VOD streaming.

The rest of this paper is organized as follows: Sec. 2 presents the motivation and theoretical performance analysis on the proposed scheme. The SLBVC scheme is described in Sec. 3. The experimental results are discussed in Sec. 4. Finally, Sec. 5 concludes the paper.

2 Motivation and Theoretical Analysis

It has been said that many videos are composed of repeated scenes. Figure 1 gives an example, which is extracted from movie "The Man from Earth."[17] The example video can be divided into 12 clips by scene change, which are signaled as P_i, $i = 1,2,\ldots,12$. The appearance time points of each clip are signaled as T_i, $i = 1,2,\ldots,12$. Visually as some clips belong to the same scene, there are five scenes in total, which are marked as S_i, $i = 1,2,\ldots,5$. In current video coding, to support RA, the encoding video is divided into independent RASs by RAPs. Without loss of generality, here we discuss the coding of S_1, which appears at a set of time points $\{T_1, T_3, T_{11}\}$. The coding structure of S_1 is shown in Fig. 2. It can be seen that P_1, P_3, and P_{11}, the clips of S_1, are divided into RASs, which are signaled as $\mathbf{X}_{1,i}$, $\mathbf{X}_{3,i}$, and $\mathbf{X}_{11,i}$, $i = 1,2,\ldots,$ by RAPs. Not only the temporal correlation between consecutive RASs in P_1, P_3, and P_{11} cannot be used, the long-term temporal correlation between P_1, P_3, and P_{11} cannot be used, either. As a result, the coding efficiency of S_1 is limited. The RASs are independent from each other because the decoded picture buffer is cleared up by RAP picture. Naturally, we introduce a scene library that stores reference frames and does not clear up at RAPs, as shown in Fig. 3. Then, the quarantine between RASs is broken. For S_1 in example video, the long-term temporal correlation between consecutive RASs, as well as clips P_1, P_3, and P_{11} can be eliminated by referencing the library. As a result, the coding efficiency of S_1 can be improved.

More generally, for sequence $\mathbf{X}$, assume the RAPs are inserted evenly according to the specified RA interval (RAI). As shown in Fig. 4, sequence $\mathbf{X}$ is divided into N independent RASs $\mathbf{X}_n$, $n = 1,2,\ldots,N$. $\mathbf{X}$ and $\mathbf{X}_n$, $n = 1,2,\ldots,N$ can be regarded as vector sources. According to rate-distortion theory,[18] the rate-distortion function of $\mathbf{X}$ in current video coding $R^C_{\mathbf{X}}(\mathbf{D}_{\mathbf{X}})$ can be written as

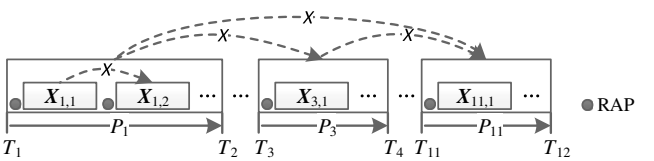

Fig. 2 Illustration of encoding P_1, P_3, and P_{11} into RASs.

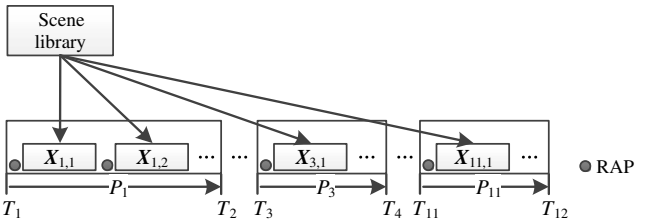

Fig. 3 Illustration of using a scene library to exploit long-term temporal correlation between RASs.

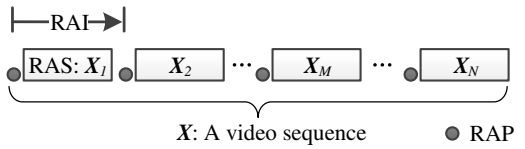

Fig. 4 An illustration of video coding with uniform RAI.

$$R^C_{\mathbf{X}}(\mathbf{D}_{\mathbf{X}}) = \sum_{n=1}^{N} R_{\mathbf{X}_n}(\mathbf{d}_{\mathbf{X}_n}) \geq R_{\mathbf{X}_1\mathbf{X}_2\ldots\mathbf{X}_N}(\mathbf{d}_{\mathbf{X}_1}, \mathbf{d}_{\mathbf{X}_2}, \ldots, \mathbf{d}_{\mathbf{X}_N}) = R_{\mathbf{X}}(\mathbf{D}_{\mathbf{X}}), \quad (1)$$

where $R_{\mathbf{X}_n}(\mathbf{d}_{\mathbf{X}_n})$ and $R_{\mathbf{X}}(\mathbf{D}_{\mathbf{X}})$ denote the classical Shannon rate-distortion function of $\mathbf{X}_n$ and $\mathbf{X}$, respectively. It can be seen that the lower bound to rate of $\mathbf{X}$ is higher than the Shannon lower bound because the correlations between RASs cannot be exploited.

In SLBVC, we employ a scene library that does not clear at RAP to exploit the long-term temporal between RASs from similar scenes. Use $\mathbf{Y}$ to represent the reference frames in the scene library. $\mathbf{Y}$ is extracted from a part of RASs, which can form a set signaled as Φ. The coding process of SLBVC is analogous to encoding RASs $\mathbf{X}_n \in \Phi$ first and then letting the extracted reference frames $\mathbf{Y}$ referenced by the rest RASs. The rate-distortion function of $\mathbf{X}$ in SLBVC can be written as

$$R^L_{\mathbf{X}}(\mathbf{D}_{\mathbf{X}}) = \sum_{\mathbf{X}_n \in \Phi} R_{\mathbf{X}_n}(\mathbf{d}_{\mathbf{X}_n}) + \sum_{\mathbf{X}_n \in \Phi} R_{\mathbf{X}_n|\mathbf{Y}}(\mathbf{d}_{\mathbf{X}_n}), \quad (2)$$

where $R_{\mathbf{X}_n|\mathbf{Y}}(\mathbf{d}_{\mathbf{X}_n})$ is the conditional rate-distortion function of $\mathbf{X}_n$ given $\mathbf{Y}$. Clearly, since reference frames $\mathbf{Y}$ can help the coding of $\mathbf{X}_n$,[18,19] we have $R_{\mathbf{X}_n|\mathbf{Y}}(\mathbf{d}_{\mathbf{X}_n}) \leq R_{\mathbf{X}_n}(\mathbf{d}_{\mathbf{X}_n})$. Finally, $R^L_{\mathbf{X}}(\mathbf{D}_{\mathbf{X}})$ can be expressed as

$$R^L_{\mathbf{X}}(\mathbf{D}_{\mathbf{X}}) \leq R^C_{\mathbf{X}}(\mathbf{D}_{\mathbf{X}}). \quad (3)$$

Compared with current video coding, the proposed scheme can push the rate-distortion bound of $\mathbf{X}$ downward to approach the Shannon rate-distortion bound.

3 Scene-Library-Based Video Coding Scheme

The framework of the SLBVC scheme is shown in Fig. 5. It is composed of a traditional codec and two scene libraries. The scene library at the encoder is built by extracting representative frames of similar scenes in a video. Then, the library frames are encoded into a unique stream and transmitted to the decoder, so that the scene library can be reconstructed in the decoder. In the encoding/decoding process, each frame is coded using its most similar library frame (MSLF) for LTR.

3.1 *Clustering-Based Scene Library Construction*

In practical implementation, we first extract library frames $\mathbf{Y}$ from sequence $\mathbf{X}$. Then, $\mathbf{Y}$ is coded and the reconstruction $\hat{\mathbf{Y}}$ is used for the reference of all RASs in $\mathbf{X}$. So the rate–distortion function of $\mathbf{X}$ is

$$R^{\mathrm{PL}}_{\mathbf{X}}(\mathbf{D}_{\mathbf{X}}) = R_{\mathbf{Y}}(\mathbf{d}_{\mathbf{Y}}) + \sum_{n=1}^{N} R_{\mathbf{X}_n|\hat{\mathbf{Y}}}(\mathbf{d}_{\mathbf{X}_n}), \quad (4)$$

where $R_{\mathbf{Y}}(\mathbf{d}_{\mathbf{Y}})$ denotes the rate–distortion function of $\mathbf{Y}$ and $R_{\mathbf{X}_n|\hat{\mathbf{Y}}}(\mathbf{d}_{\mathbf{X}_n})$ represents the conditional rate-distortion function of $\mathbf{X}_n$ given $\hat{\mathbf{Y}}$. The aim of library construction is to minimize the rate-distortion function shown in Eq. (4). However, it is known that the intra-RAP frames use no temporal correlation and their coding efficiency is not high. So, the scene library has a much stronger impact on the RAP frames. It is more efficient to build a library carrying high correlation with the RAP frames rather than with the entire encoding video. In addition, to simplify the construction of the scene library, we also ignore the impact of distortion. As a result, the goal of library construction can be modified as

$$\min\left\{H(\mathbf{Y}) + \sum_{n=1}^{N} H(L_n|\mathbf{Y})\right\}, \quad (5)$$

where $H(\mathbf{Y})$ is the entropy of $\mathbf{Y}$, L_n, $n = 1,2,\ldots,N$ are the RAP frames, and $H(L_n|\mathbf{Y})$ is the conditional entropy of L_n given $\mathbf{Y}$. To minimize Eq. (5), on one hand, $\mathbf{Y}$ should share as much correlations as possible with the RAP frames. On the other hand, the entropy of $\mathbf{Y}$ should be limited as low as possible. As the library frames are used for the reference

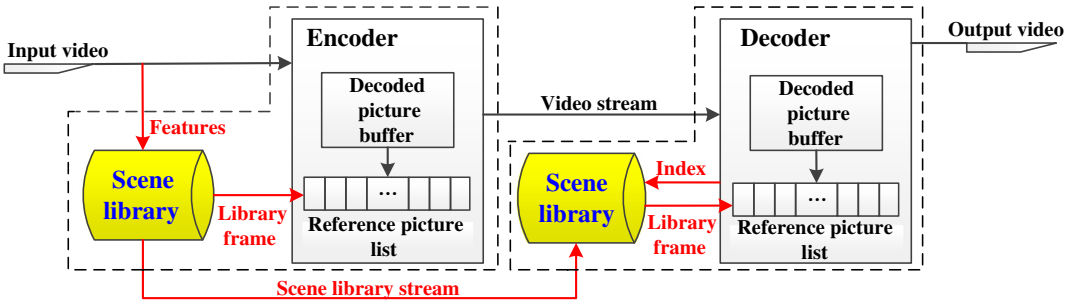

Fig. 5 The framework of the SLBVC scheme.

of RAP frames, they should be all intracoded to ensure the RA capability of RAP frames. So, we directly extract the library frames from the RAP frames. Then, the problems become: (1) How many RAP frames should be extracted? (2) Which RAP frames should be extracted?

If similar RAP frames can be clustered together, a single cluster representative can serve as a good reference frame of the others. This is a good optimization solution to Eq. (5). To build the library, we classify the RAP frames into different clusters, and the center frames of each cluster are specified as library frames. K-means algorithm is employed for clustering because of its simplicity and efficiency.[20] Also, RAP frames from the same scene have similar color histogram features. So, the clustering is based on color histogram feature.[21,22] For an RAP frame L_n, $1 \le n \le N$, it can be described by a color histogram as

$$L_n = \{f_n^0, f_n^0, \ldots, f_n^{767}\}, \tag{6}$$

where f_n^p, $0 \le p < 256$ represents the number of Y-component with value p, f_n^p, $256 \le p < 512$ represents the number of U-component with value $p - 256$, and f_n^p, $512 \le p < 768$ represents the number of V-component with value $p - 512$. Accordingly, the difference between two RAP frames L_i and L_j, $1 \le i, j \le N$ can be calculated as

$$D(L_i, L_j) = \sum_{p=0}^{767} |f_i^p - f_j^p|. \tag{7}$$

The K-means algorithm is implemented with given clustering number K and initial clustering centers μ_k, $k = 1, 2, \ldots, K$ as follows:

Step 1: Classify each RAP frame L_n, $1 \le n \le N$ to the cluster $\hat{k}$ with minimum difference, as

$$\hat{k} = \arg_{1 \le k \le K} \min[D(L_n, \mu_k)]. \tag{8}$$

Step 2: Update the clustering centers. For cluster k, the center frame is updated as the frame with the minimum sum of differences with other frames in cluster k, which is expressed as

$$\mu_k = L_{j^*}^k = \arg_{1 \le j \le n_k} \min\left[\sum_{i=1}^{n_k} D(L_i^k, L_j^k)\right], \tag{9}$$

where n_k is the number of frames belonging to cluster k, L_i^k and L_j^k are the i'th and j'th frames, respectively, in cluster k.

Repeat step 1 and step 2 until clustering centers μ_k, $k = 1, 2, \ldots, K$ do not change any more.

The number of clusters K needs to be explicitly specified for K-means algorithm. However, how many library frames should be extracted is not known as *a priori*. To find the optimal number of clusters, we traverse all clustering options with K varying from 1 to N. The clustering cost of each option is calculated and the minimum clustering cost corresponds to the optimal clustering number.

We define the clustering cost of each clustering option as the sum of information content of all clusters. For K-cluster clustering, the clustering cost is computed as

$$\text{Cost}(K) = \sum_{k=1}^{K} C_k, \quad C_k = H(\mu_k) + \sum_{i=1}^{n_k} H(L_i^k|\mu_k), \tag{10}$$

where C_k is the cost of k'th cluster, $H(\mu_k)$ is the entropy of μ_k, while $H(L_i^k|\mu_k)$ represents the conditional entropy of L_i^k given μ_k. Use $H(L_i^k, \mu_k)$ to represent the joint entropy of L_i^k and μ_k. The relationship between $H(L_i^k|\mu_k)$ and $H(L_i^k, \mu_k)$ is given by

$$H(L_i^k|\mu_k) = H(L_i^k, \mu_k) - H(\mu_k). \tag{11}$$

We use the luma histogram of μ_k and the joint luma histogram of L_i^k and μ_k to estimate $H(\mu_k)$ and $H(L_i^k, \mu_k)$, respectively.[23] They are calculated as

$$H(\mu_k) = -\sum_{p=0}^{255} \frac{f_{\mu_k}^p}{TN} \log\left(\frac{f_{\mu_k}^p}{TN}\right), \tag{12}$$

$$H(L_i^k, \mu_k) = -\sum_{p=0}^{255}\sum_{q=0}^{255} \frac{f_{L_i^k,\mu_k}^{p,q}}{TN} \log\left(\frac{f_{L_i^k,\mu_k}^{p,q}}{TN}\right). \tag{13}$$

In Eqs. (12) and (13), TN is the total number of pixels in a picture, $f_{\mu_k}^p$ represents the number of pixels with luma value p in μ_k, $f_{L_i^k,\mu_k}^{p,q}$ represents the number of pixel pair (p, q), which means the luma values of the same position in L_i^k and μ_k are p and q, respectively.

By integrating Eqs. (11)–(13) into Eq. (10), the clustering cost $\text{Cost}(K)$ can be derived. Traverse K from 1 to N to calculate the cost of all clustering options. The clustering number is chosen as K_{opt} which corresponds to the minimum clustering cost. After classifying the RAP frames into K_{opt} clusters, the center frames are extracted as library frames. Note that there may be some clusters containing only one frame, which has little correlation with other RAP frames. Therefore, we only extract the center frames of clusters containing multiple frames as library frames.

The value of K_{opt} is in the range of 1 to N. In extremes, when K_{opt} is equal to 1, it means all RAP frames are similar to each other. So, only one library frame is extracted. In contrast, when K_{opt} is equal to N, it means the RAP frames are different to each other. It is not efficient to use any frames as the reference of the others. Thus, no scene library frames are extracted.

Use the clustering of the scene change frames of the example video in Fig. 1 as an example. There are 12 scene change frames, which are signaled as L_i, $i = 1,2, \ldots, 12$. After traversing all clustering options, a curve of clustering cost $\text{Cost}(K)$ relative to K is drawn as Fig. 6. It can be seen that the optimal clustering number is five, and the corresponding clustering result is shown in Fig. 7. In the example, frames L_1, L_2, L_6, L_9, and L_{10} are the center frames. As frame L_2 is the only frame in cluster 2, finally frames L_1, L_6, L_9, and L_{10} are extracted as the library frames.

3.2 *Coding Structure Based on Scene Library*

Given the specified RAI, the distance of two consecutive RAP pictures should be less than or equal to the specified interval. Simply, RAP pictures are inserted with fixed RAI

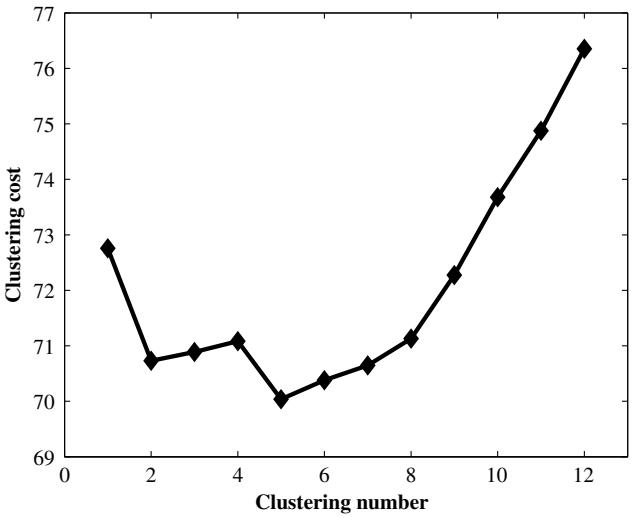

Fig. 6 The curve of clustering cost relative to cluster number of example video.

(FRAI). However, RAP picture has another function that is used as the reference frame of its following interframes in the same RAS directly or indirectly. So, RAP pictures are usually coded with higher quality for better interprediction efficiency. For a video with multiple scenes, if scene change occurs in an RAS, frames in the old scene are not suitable to be used as reference frames of the new scene because of their low similarity. Thus, the coding efficiency of the RAS is limited. As a result, we adopt an adaptive RAI (ARAI) coding structure[24] in the SLBVC scheme. When scene change happens, the scene change frame is coded as a RAP frame. The later RAP frames are inserted at the specified interval until the next scene change occurs. The ARAI structure can facilitate inter prediction efficiency in the RAS at scene change position so as to improve coding efficiency. Table 1 shows the coding efficiency of ARAI structure compared to FRAI structure in terms of Bjøntegaard delta bit rate

Table 1 The performance of ARAI over FRAI with RAI 32.

Sequence	BD-rate Y (%)	BD-rate U (%)	BD-rate V (%)
Bigbang	−2.1	−3.7	−4.4
Cards	−1.8	−2.3	−2.4
Sherlock	−2.2	−3.8	−3.8
Earthman	−2.3	−3.3	−3.1
Girls	−2.2	−3.0	−2.9
Throne	−2.3	−3.4	−3.1
Average	−2.1	−3.2	−3.3

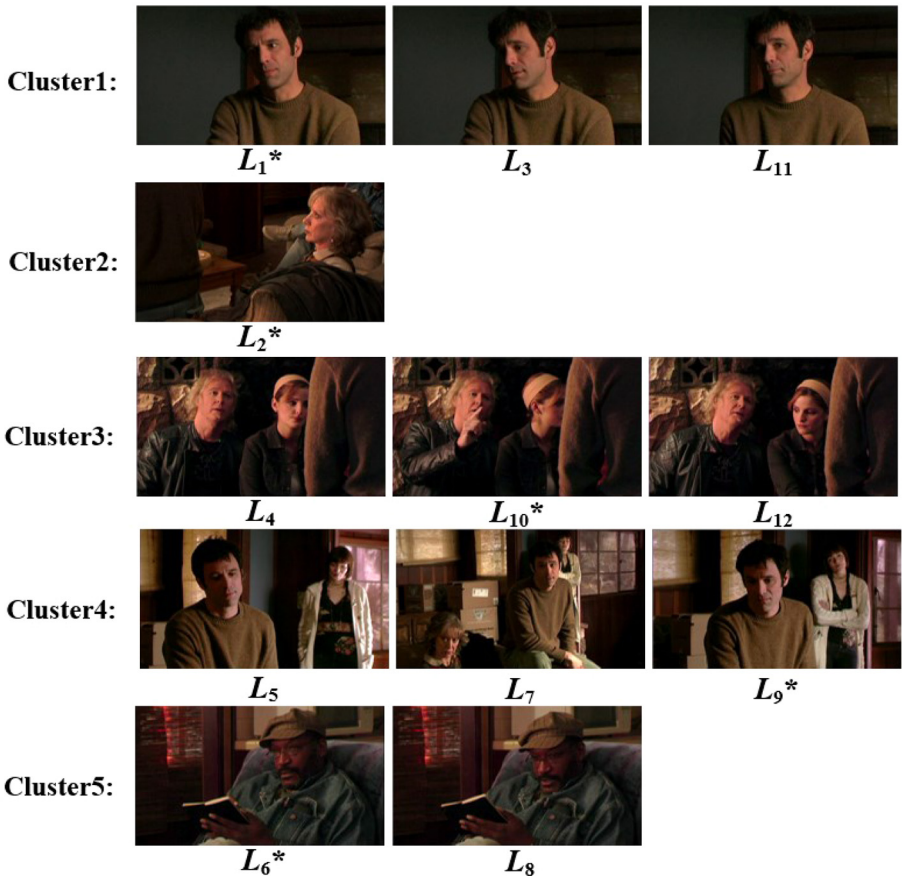

Fig. 7 The clustering result of example video corresponding to the optimal clustering number (frames with "*" are center frames).

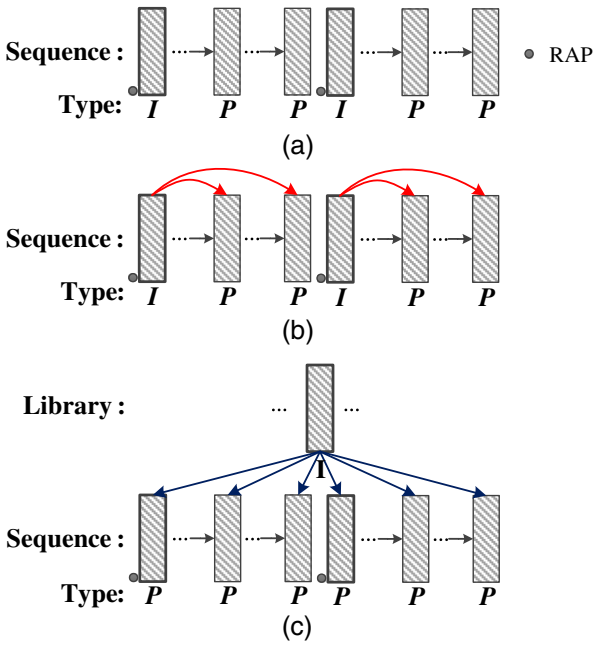

Fig. 8 The coding structures of (a) conventional video coding, (b) LTR, and (c) SLBVC.

(BD-Rate).[25] We employed six sequences containing multiple scene changes for test. The details of the test sequences can be found in Sec. 4. The test is conducted under RA common test condition[26] and the RAI is specified as 32. It can be seen that 2.1% coding gain can be achieved by the ARAI coding structure, which confirms the adoption of ARAI structure in the proposed scheme.

Figure 8(a) shows the coding structure of two RASs from the same scene in conventional video coding (for simplicity we did not show multiframe reference and B-frames). The RAP frames are all intracoded while the non-RAP frames are coded as interframes by referencing neighboring frames to exploit short-term correlation. A traditional LTR scheme is implemented using RAP frame as the LTR frame of following non-RAP frames in the same RAS, which is shown in Fig. 8(b). In this way, the long-term correlation in an RAS can be exploited. Now with the availability of the scene library, the long-term correlation between RASs can also be exploited.

In SLBVC, a layered coding structure is proposed, as shown in Fig. 8(c). The library frames are coded as intraframes. Each RAP frame is coded as an interframe by only referencing its MSLF, which is retrieved by color histogram comparison.[27] Then, the MSLF of the RAP frame is used as the LTR frame of the following non-RAP frames. We do it this way because frames in an RAS belong to the same scene with ARAI coding structure. So, the complexity for retrieving the MSLFs of non-RAP frames can be saved. Meanwhile, the RA property of the RAS is guaranteed. As a result, the long-term temporal correlation between RASs can be exploited clearly. In addition to, the long-term temporal correlation in RASs can also be removed to some extent.

3.3 *Scene Library Management*

Due to storage capacity of codec, the scene library is built for video clip with several minutes' length to limit the number of library frames. In other words, the frames in the scene library are not infinitely accumulated. After encoding/decoding a video clip, the scene library is cleared. Then, a new scene library will be built for the next clip. The video clip and corresponding scene library are encoded into two streams. In video stream, a signal for each frame is added to indicate its MSLF in the library stream. Here, we only discuss the library management strategies in VOD streaming application, because for locally stored video streams, the strategies are all the same except for omitting the stream request operation.

In decoder, the video stream and the library stream are downloaded and decoded synchronously. When the decoder is capable of storing all the decoded library frames, they are stored into the scene library and not to be removed. In this way, the stored library frames can be directly reused by later frames in the video clip. In contrast, when the decoder can only store a part of the decoded library frames, e.g., mobile devices, the downloaded library stream is stored in local. The scene library works in a first-in, first-out manner to keep the most recently decoded library frames. Then, later frames can find their MSLFs either in the scene library or in the library stream. For RA within current video clip, if the RAP frame to be decoded cannot find its MSLF in the scene library, it will search the stored library stream. If the MSLF cannot be found in the stored library stream either, it will be requested from the server. For RA across video clips, the scene library and the stored library stream will be cleared. Then, the streams of the RAP frame and its MSLF will be requested from the server.

4 Experimental Results

4.1 *Experimental Set-up*

Experiments are conducted on six sequences containing repeated scenes to evaluate the performance of the SLBVC scheme. The test sequences were extracted from six movies and TV series: "The Big Bang Theory," "House of Cards," "Sherlock," "The Man from Earth," "2 Broke Girls," and "Game of Thrones." The details of the sequences are shown in Table 2. They all have a length of >4000 frames. All sequences are composed of multiple scenes, a large portion of which appear repeatedly.

The proposed scheme is implemented on an HM12.1[28] encoder. We compare it with HEVC, the LTR scheme, and the method in Ref. 16 to demonstrate its performance.

Table 2 Description of test sequences.

Sequence	Resolution	Fps (Hz)	Length	Scene change times
Bigbang	1280 × 720	30	4080	52
Cards	1280 × 720	30	4337	23
Sherlock	1280 × 720	30	4553	69
Earthman	640 × 360	30	4339	37
Girls	640 × 360	30	4567	44
Throne	640 × 360	30	4336	46

Table 3 The random-access configurations of HM.

Configuration	Value	Configuration	Value	Configuration	Value
Profile	Main	LCU Size	64	Fast search	Enable
Frame structure	Hierarchical B	Search range	64	SAO	Enable
		AMP	Enable	RDOQ	Enable
GOP size	8	Hadmard ME	Enable	Rate control	Disable

The sequences are encoded under RA common test condition, which is designed for RA applications. Table 3 shows the details of RA common test condition. The quantization parameter (QP) of test sequences QP_S is set as 22, 27, 32, and 37 while the QP of scene library frames is empirically set as $QP_S - 6$, which usually leads to the best performance among all QP values. In our experiments, the ARAI coding structure is also employed by HEVC and the LTR scheme for fair comparison. Two different RAI values are tested: 32, as specified in RA common test condition; and 152 (5 s for 30 fps), which is the typical RAI used by online video websites. In the following description, we use RAI 32 and RAI 152 to represent the conditions that the RAI is set as 32 and 152, respectively.

Table 4 The numbers of RAP frames, clusters, and library frames with RAI 32 and 152.

	RAI 32			RAI 152		
Sequence	RAP frame	Cluster	Library frame	RAP frame	Cluster	Library frame
Bigbang	150	41	26	56	13	10
Cards	147	27	21	40	13	9
Sherlock	175	42	27	74	24	13
Earthman	152	23	21	46	12	10
Girls	166	53	29	53	14	9
Throne	158	47	28	50	10	8

4.2 *Results of Library Construction*

The numbers of RAP frames and clusters of test sequences are shown in Table 4. It can be seen that the RAP frames are classified into a smaller number of clusters. The optimal clustering number depends on the video content and the number of RAP frames. An illustration of the curves of clustering cost relative to clustering number of Bigbang is shown in Fig. 9. The optimal clustering numbers are, respectively, 41 and 13 with RAI 32 and 152. After clustering, the scene library is constructed by extracting the center frame of each cluster. As only the center frames of clusters containing more than one frames are chosen as library frames, so the number of library frames is less than or equal to the number of clusters, as also shown in Table 4.

4.3 *Performance of Proposed Scheme*

BD-Rate is used to evaluate the rate-distortion performance of the proposed scheme. Note that each non-RAP frame also uses the library frame as LTR frame, so the coding performance is evaluated on the whole sequence. The bits brought by the scene library are taken into consideration when calculating the bitrate. The achieved gains of the LTR scheme, the scheme in Ref. 16, and the proposed scheme with respect to HEVC are shown in Table 5. It can be seen that with RAI 152, there is some long-term temporal correlation in RAS that can be exploited, so the LTR scheme can achieve

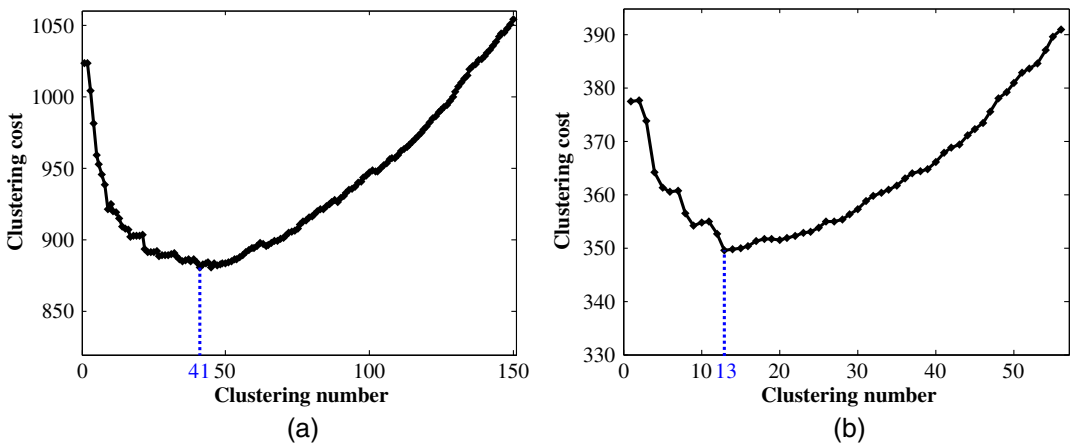

Fig. 9 The curves of clustering cost relative to clustering number of Bigbang with (a) RAI 32 and (b) RAI 152.

Table 5 The performances of LTR scheme, scheme in Ref. 16, and SLBVC scheme compared to HEVC.

	RAI 32			RAI 152		
Sequence	LTR (%)	Ref. 16 (%)	SLBVC (%)	LTR (%)	Ref. 16 (%)	SLBVC (%)
Bigbang	−0.2	−16.2	−19.5	−0.5	−6.6	−10.8
Cards	0.1	−30.9	−37.4	−0.8	−9.8	−16.3
Sherlock	−0.1	−19.2	−26.5	−0.7	−3.8	−16.2
Earthman	0.2	−28.0	−36.2	0.1	−9.9	−12.5
Girls	0.0	−24.4	−34.9	−0.9	−3.6	−15.9
Throne	0.3	−23.2	−35.8	−0.1	−9.8	−15.0
Average	**0.1**	**−23.6**	**−31.7**	**−0.5**	**−7.2**	**−14.4**

0.5% coding gain. However, with RAI 32, few long-term temporal correlation is left because of the short length of RAS, thus the LTR scheme cannot bring any coding gain. However, the scheme in Ref. 16 and the proposed scheme can achieve much higher coding gain because they are capable of removing the long-term temporal correlation between RASs and repeated scenes. Also for the both schemes, the performance with RAI 32 is better than RAI 152. This is because much more long-term temporal correlation between RASs remains with shorter RAI structure but now can be removed. Compared to Ref. 16, the bit-savings of the proposed scheme are 8.1% with RAI 32 and 7.2% with RAI 152. These coding gains are obtained due to the more efficient scene library built with the cluster-based method and the better coding of non-RAP frames.

Figures 10 and 11 show the rate-distortion curves for all six test sequences by four schemes: HEVC, the LTR scheme, the scheme in Ref. 16, and the SLBVC scheme with RAI 32 and 152, respectively. The rate-distortion curves of the LTR scheme and HEVC are hardly distinguishable because of their close performance. However, the rate-distortion curves of the SLBVC scheme are obviously above that of the other three schemes, which reveals the remarkable performance of the proposed scheme. For sequences Cards (1280×720) and Throne (640×360), the maximal bitrate is around 200 and 100 kbps because the images contain a lot of dark regions and lack high efficiency components. It does not take too many bits to encode them even with very high quality. For the other sequences with rich colors and textures, the bitrate can be up to hundreds of kbps. In summary, the proposed scheme can deal with different types of video contents and show advantage over a wide range of bitrate.

In addition to the improvement in objective performance, visual quality is also significantly improved by the proposed scheme. When the SLBVC scheme is applied, coding artifacts caused by aggressive compression, such as blockiness and blurriness, will be greatly alleviated. Figure 12 shows a set of images cropped from the middle frame of an RAS in

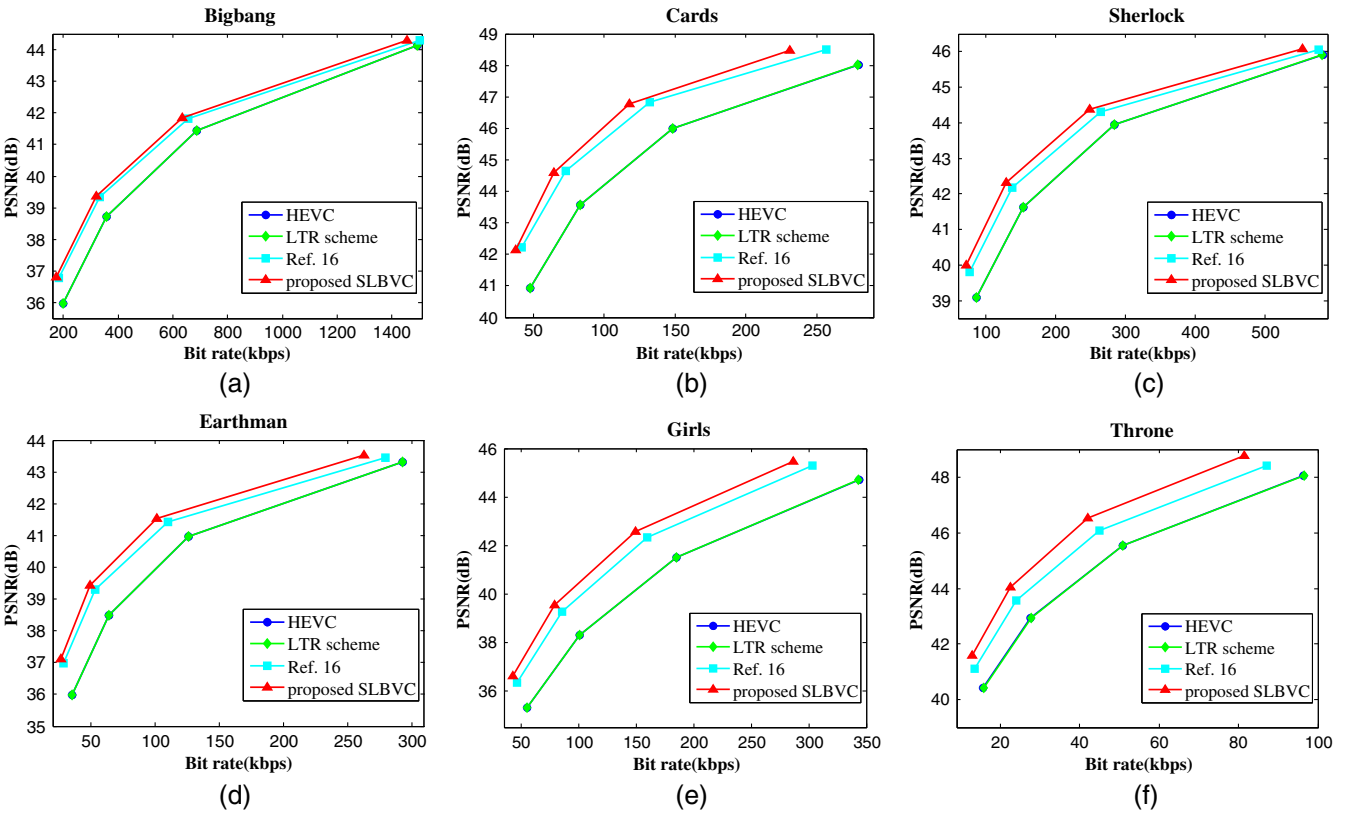

Fig. 10 The rate-distortion curves of (a) Bigbang, (b) Cards, (c) Sherlock, (d) Earthman, (e) Girls, and (f) Throne with RAI 32.

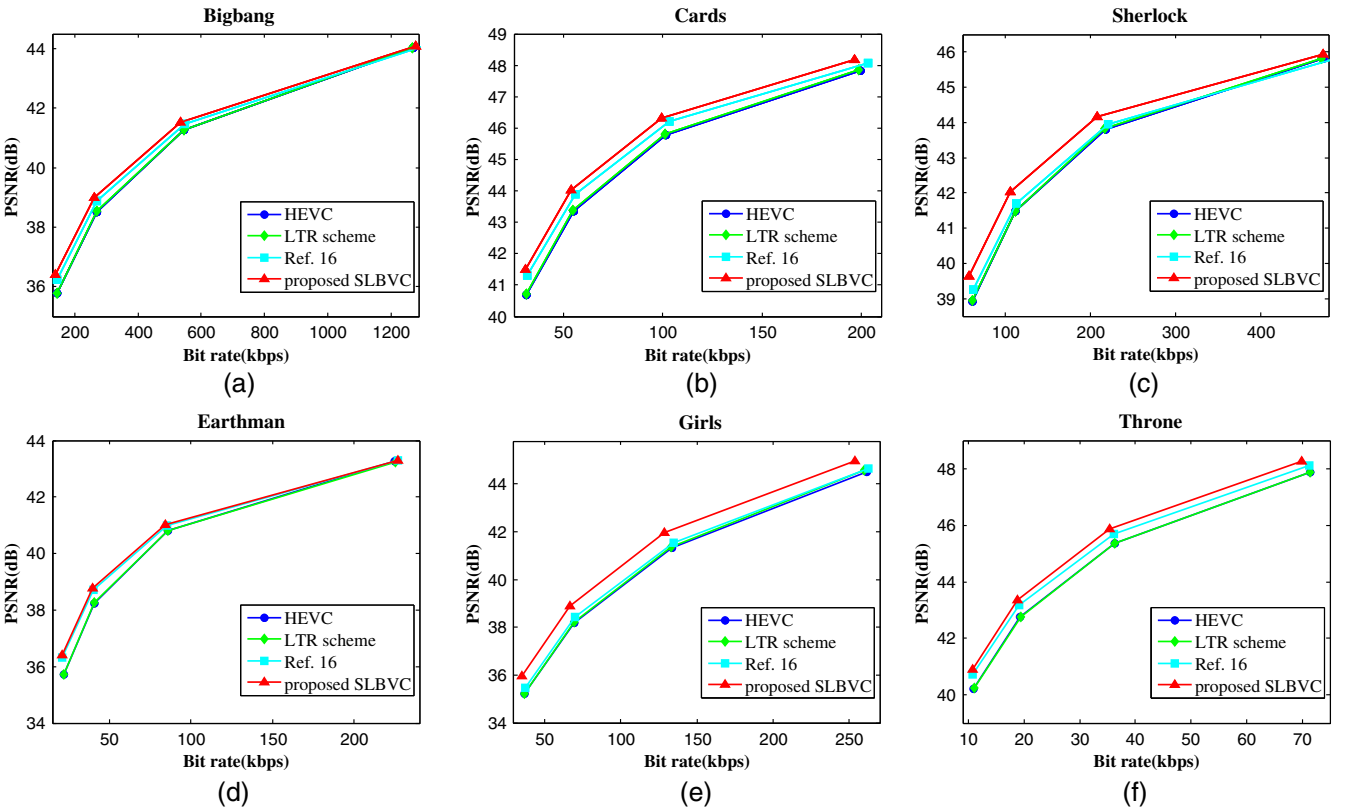

Fig. 11 The rate-distortion curves of (a) Bigbang, (b) Cards, (c) Sherlock, (d) Earthman, (e) Girls, and (f) Throne with RAI 152.

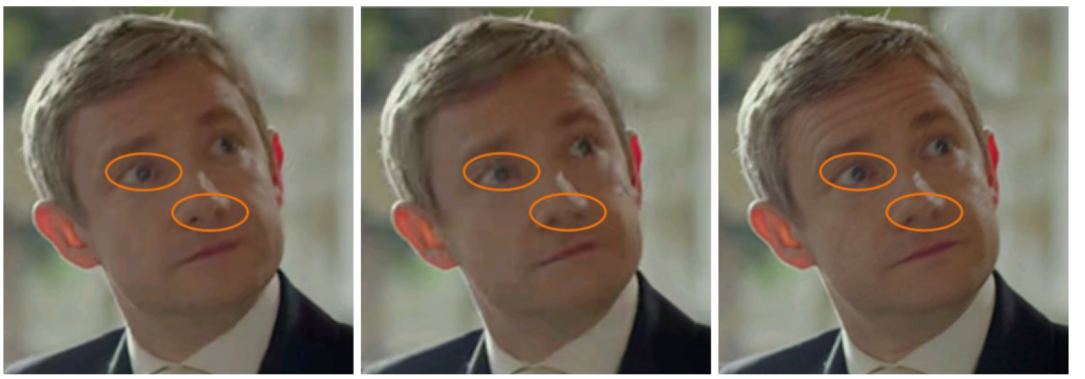

Fig. 12 Images cropped from the middle frame of an RAS in sequence Sherlock with RAI 32. (a) By SLBVC (72 kbps), (b) by HEVC (86 kbps), and (c) original image.

Sherlock coded with RAI 32. The first image is coded by the proposed scheme, the second image is coded by HEVC, and the third one is the original image used for comparison. The bitrate of the proposed scheme and HEVC is, respectively, 72 and 86 kbps, the PSNR values of the images range between 39 and 41 dB. It can be seen that the contour of the actor's face (eyes and nose) is much better preserved by the proposed scheme. Similarly, Fig. 13 shows the images cropped from Bigbang coded with RAI 152. Also, the textures of the clothes are visually clearer with the proposed scheme. As a result, we can conclude that the subjective quality of proposed scheme is much better than that of HEVC.

4.4 *Discussion of Random Access*

In VOD streaming, sometimes the users do not want to watch the entire video sequence. They may request only part of the sequence from the server by RA. When only a video part is transmitted to the client, only the referenced library frames instead of the whole library need to be synchronously transmitted. It is feasible because the library frames are intra-frames and independent from each other. With the length of the transmitted video part decreasing, the number of frames that a library frame is referenced by also decreases. The performance of the SLBVC scheme may deteriorate. However, since about 75% users watch the entire video in

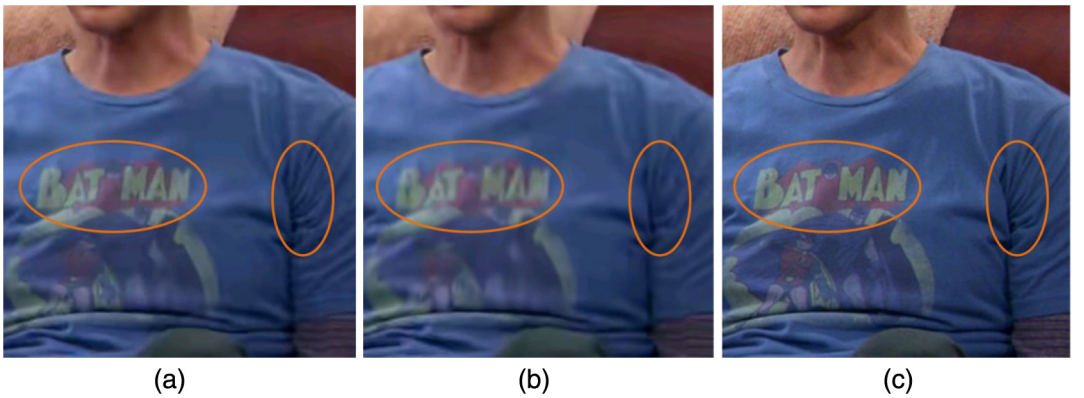

Fig. 13 Images cropped from the middle frame of an RAS in sequence Bigbang with RAI 152. (a) By SLBVC (139 kbps), (b) by HEVC (145 kbps), and (c) original image.

practice,[29,30] the proposed scheme is still valuable for VOD streaming application.

We simulate the performance when part of the sequence transmitted in VOD streaming. Assume the test sequence contains N RASs with RAI 152. We test a series of transmitted lengths round$(\frac{N}{2^m})$, $m = 0,1,\ldots,m*$, where round$(*)$ is the function that rounds the variable to the nearest integer, and $m*$ is the least integer satisfying round$(\frac{N}{2^{m*}}) = 1$. In each test option, the test sequence is divided into homogeneous parts at the specified transmitted length. The coding performance (BD-rate) of each part is calculated with considering the bits of the referenced library frames. Then the performances of all parts are averaged to derive the final result. Figure 14 shows the curves of coding performance relative to m of all sequences. The value of $m*$ for Cards and Earthman is 5, whereas for the other sequences is 6. There is performance deterioration with the increase of m and even performance loss when m is too large. However, performance improvement can still be observed over a large range of transmitted length. For Bigbang, Sherlock, Girls, and Throne, when m is between 0 and 3, which corresponds to that 12.5% to 100% of the whole sequence is transmitted, there is coding efficiency improvement. For Cards, a larger range of m between 0 and 4 is observed to show coding gain. For Earthman, although the range of m with coding gain is narrower (0 to 2), the proposed scheme still outperforms HEVC if only no <25% of the sequence is transmitted.

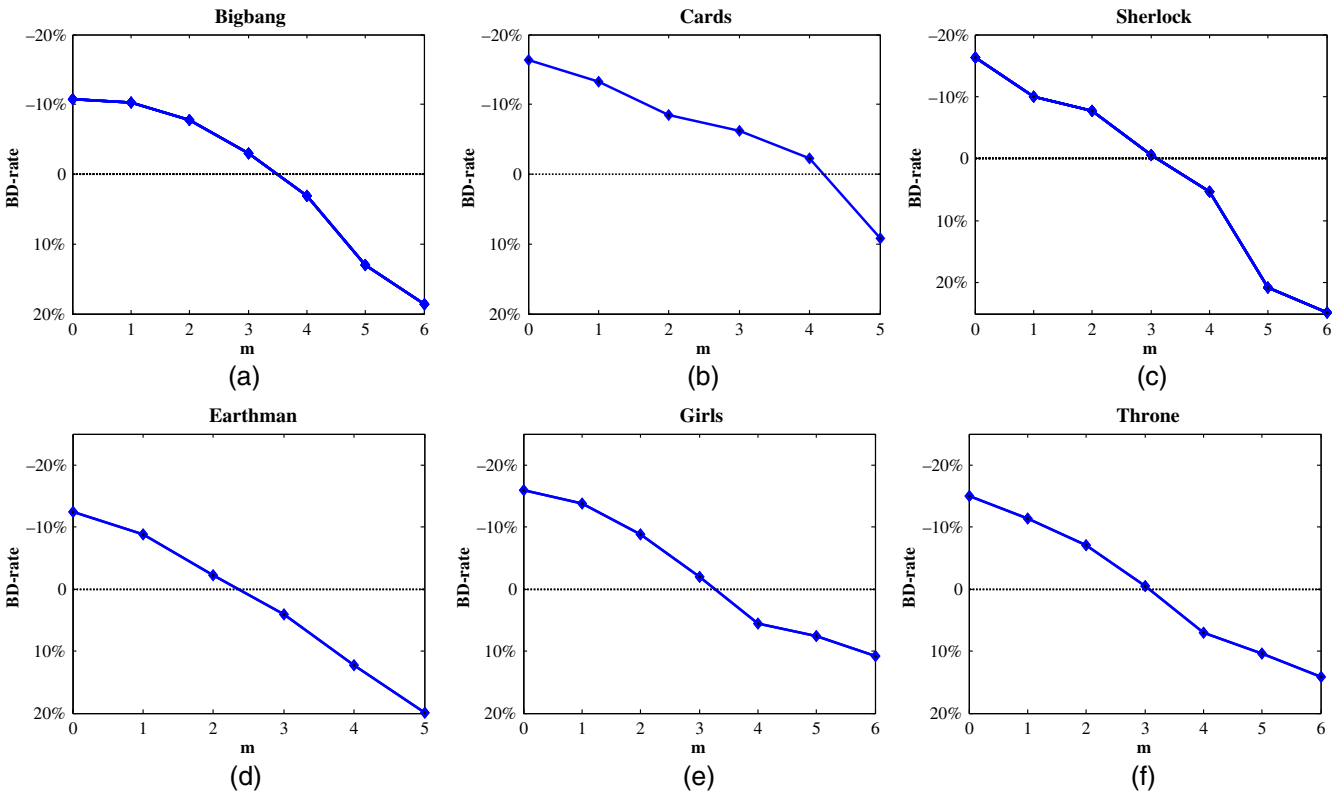

Fig. 14 The curves of coding performance (BD-Rate) relative to m of (a) Bigbang, (b) Cards, (c) Sherlock, (d) Earthman, (e) Girls, and (f) Throne with RAI 152.

4.5 *Complexity Analysis*

The complexity of the proposed scheme is imposed by scene library construction (RAP frame clustering and library frame encoding), MSLF selection, and video coding. As the HM encoder is written in C++, the algorithms of scene library construction and MSLF selection are also implemented in C++ for fair comparison. The experiments are conducted on Intel(R) Xeon (R) CPU E5- 2690 0 @2.90 GHz with 190G RAM memory. The complexity of each process is measured by the running time.

The computational complexity of SLBVC T_{SLBVC} can be expressed as

$$T_{\mathrm{SLBVC}} = T_1 + T_2 + T_3 + T_4, \tag{14}$$

where T_i, $i = 1, \ldots, 4$ represent the running time of RAP frame clustering, library frame encoding, MSLF selection, and video coding, respectively. The complexity distributions of each process with RAI 32 and 152 are presented in Tables 6 and 7. In RAI 152, T_1, T_2, and T_3 occupy about 0.04%, 0.05%, and 0.001% of the total time. While in RAI 32, the percentages of T_1, T_2, and T_3 increase to 0.79%, 0.13%, and 0.002% for the reason of more RAP frames. Especially for T_1, assume the number of RAP frames is N, the number of times of computing distances between frames is $O(N^3)$. Thus, the increase of complexity is much higher than T_2 and T_3. Overall, compared with the complexity of video coding, T_4, T_1, T_2, and T_3 are rather small and can be neglected.

Table 6 The complexity distribution of each process in SLBVC with RAI 32.

Sequence	RAP frame clustering (%)	Library frame encoding (%)	MSLF selection (%)	Video coding (%)
Bigbang	0.50	0.15	0.002	99.35
Cards	0.33	0.11	0.001	99.56
Sherlock	0.55	0.13	0.002	99.32
Earthman	1.14	0.10	0.002	98.76
Girls	0.98	0.16	0.002	98.86
Throne	1.27	0.12	0.002	98.61
Average	0.79	0.13	0.002	99.08

Table 7 The complexity distribution of each process in SLBVC with RAI 152.

Sequence	RAP frame clustering (%)	Library frame encoding (%)	MSLF selection (%)	Video coding (%)
Bigbang	0.03	0.05	0.001	99.91
Cards	0.01	0.04	0.001	99.94
Sherlock	0.04	0.06	0.001	99.89
Earthman	0.04	0.04	0.001	99.92
Girls	0.04	0.05	0.001	99.91
Throne	0.08	0.05	0.001	99.87
Average	0.04	0.05	0.001	99.91

Table 8 The video coding time of LTR and SLBVC compare to HEVC.

	RAI 152		RAI 32	
Sequence	LTR (%)	SLBVC (%)	LTR (%)	SLBVC (%)
Bigbang	125.9	131.0	125.8	133.7
Cards	125.3	130.0	125.1	130.1
Sherlock	127.0	133.7	126.8	133.5
Earthman	125.6	132.8	125.6	133.7
Girls	126.0	132.6	125.7	130.9
Throne	126.7	129.1	123.9	127.5
Average	126.1	131.5	125.5	131.6

Use the video coding time of HEVC as anchor, the video coding time of the LTR scheme and the proposed scheme is shown in Table 8. It can be seen that about 24% to 34% extra complexity is brought into video coding by the LTR scheme and the proposed scheme. It is because an additional LTR frame is employed by each frame in both schemes, thus the ME complexity is increased. Also, the video coding complexity of SLBVC is slightly higher than the LTR scheme by 6% in average as the RAP frames are coded as inter-frames. Considering the high performance of the SLBVC scheme, the complexity increase is acceptable.

5 Conclusion

In this paper, an SLBVC scheme is presented to improve the coding efficiency of videos with repeated scenes. The proposed scheme is capable of exploiting long-term temporal correlation between RASs which belong to similar scenes. It can be applied to stored video application (e.g., DVD, BD, etc.) and VOD streaming. Compared to state-of-the-art method, the scheme can achieve 8.1% and 7.2% coding performance improvement on the whole sequence with RAI 32 and 152, respectively. Although the performance may degrade when only part of sequence is transmitted in streaming, the proposed scheme still shows advantage over a large range of transmitted length.

A couple of extensions of our current work can be further explored in the future. For example, more accurate clustering criteria can be investigated for the clustering algorithm to build a more efficient scene library. Also, the coding quality of library frames can be adaptively optimized according to the times they are referenced to improve overall coding efficiency. We will work on these issues to further improve the efficiency of the SLBVC scheme.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61371162). The authors declare that there are no conflicts of interest regarding the publication of this paper.

References

1. "Information technology—generic coding of moving pictures and associated audio information: video," ISO/IEC 138 18-2 and ITU-T Rec. H.262, Geneva, Switzerland (1995).
2. K. Rijkse, "H. 263: video coding for low-bit-rate communication," *IEEE Commun. Mag.* **34**(12), 42–45 (1996).
3. K. Asai et al. , "Core experiments of video coding with block-partitioning and adaptive selection of two frame memories (STFM / LTFM)," MPEG 96/M0654, ISO/IEC JTC1/SC29/WG11, Geneva, Switzerland (1996).
4. T. Wiegand et al., "Long-term memory motion-compensated prediction," *IEEE Trans. Circuits Syst. Video Technol.* **9**(1), 70–84 (1999).
5. T. Wiegand et al., "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.* **13**(7), 560–576 (2003).
6. G. Sullivan et al., "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.* **22**(12), 1649–1668 (2012).
7. J. Bankoski et al., "Technical overview of VP8, an open source video codec for the web," in *IEEE Int. Conf. Multimedia and Expo*, Barcelona, Spain, pp. 1–6 (2011).
8. S. Ma et al., "AVS2-making video coding smarter [standards in a nutshell]," *IEEE Signal Process. Mag.* **32**(2), 172–183 (2015).
9. M. Paul et al., "Explore and model better I-frames for video coding," *IEEE Trans. Circuits Syst. Video Technol.* **21**(9), 1242–1254 (2011).
10. X. Zhang et al., "Low-complexity and high-efficiency background modeling for surveillance video coding," in *IEEE Visual Communications and Image Processing (VCIP 2012)*, San Diego, California, pp. 1–6 (2012).
11. X. Zhang et al., "Background-modeling-based adaptive prediction for surveillance video coding," *IEEE Trans. Image Process.* **23**(2), 769–784 (2014).
12. X. Zhang et al., "Optimizing the hierarchical prediction and coding in HEVC for surveillance and conference videos with background modeling," *IEEE Trans. Image Process.* **23**(10), 4511–4526 (2014).
13. M. Tiwari et al., "Selection of long-term reference frames in dual-frame video coding using simulated annealing," *IEEE Signal Process. Lett.* **15**(1), 249–252 (2008).
14. D. Liu et al., "Dual frame motion compensation with optimal long-term reference frame selection and bit allocation," *IEEE Trans. Circuits Syst. Video Technol.* **20**(3), 325–339 (2010).
15. M. Paul et al., "A long term reference frame for hierarchical B-picture based video coding," *IEEE Trans. Circuits Syst. Video Technol.* **24**(10), 1729–1742 (2014).
16. X. Zuo et al., "Library based coding for videos with repeated scenes," in *Picture Coding Symp.*, pp. 100–104 (2015).
17. R. Schenkman, "The Man from Earth," 2007, https://www.youtube.com/watch?v=8ZMmo8TFaQk (July 2016).
18. R. M. Gray, "A new class of lower bounds to information rates of stationary sources via conditional rate-distortion functions," *IEEE Trans. Inform. Theory* **19**(4), 480–489 (1973).
19. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed., p. 581, John Wiley & Sons, Inc., Hoboken, New Jersey (2006).
20. J. A. Hartigan et al., "Algorithm AS 136: a K-means clustering algorithm," *J. R. Stat. Soc.* **28**(1), 100–108 (1979).
21. B. Günsel et al., "Temporal video segmentation using unsupervised clustering and semantic object tracking," *J. Electron. Imaging* **7**(3), 592–604 (1998).
22. M. Mignotte, "Segmentation by fusion of histogram-based-means clusters in different color spaces," *IEEE Trans. Image Process.* **17**(5), 780–787 (2008).
23. D. Mistry et al., "Image similarity based on joint entropy (joint histogram)," in *Int. Conf. on Advances in Engineering and Technology* (2013).
24. J. Fan et al., "Adaptive motion-compensated video coding scheme towards content-based bit rate allocation," *J. Electron. Imaging* **9**(4), 521–533 (2000).
25. G. Bjontegaard, "Calculation of average PSNR differences between R-D curves," *Document VCEG-M33, 13th ITU-T SG16/Q6 VCEG meeting*, Austin, Texas (2001).
26. F. Bossen, "Common test conditions and software reference configurations," in *Document JCTVC-L1100, 12th JCT-VC meeting*, Geneva, Switzerland (2013).
27. H. Zhang et al., "Automatic partitioning of full-motion video," *Multimedia Syst.* **1**(1), 10–28 (1993).
28. F. Bossen et al., "HM reference software 12.1," 2013, https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/tags/HM-12.1 (January 2016).
29. C. Costa et al., "Analyzing client interactivity in streaming media," in *Proc. of the 13th Int. Conf. on World Wide Web*, New York, pp. 534–543 (2004).
30. J. Choi et al., "A survey of user behavior in VoD service and bandwidth-saving multicast streaming schemes," *IEEE Commun. Surv. Tutorials* **14**(1), 156–169 (2012).

Xuguang Zuo received his BS degree in information and communication engineering from Zhejiang University, Hangzhou, China, in 2010, where he is currently pursuing a PhD. His research interests include video processing, coding, and streaming.

Lu Yu is currently a professor at the Institute of Information and Communication Engineering, Zhejiang University. She received her BS degree in radio engineering and her PhD in communication and electronic systems from Zhejiang University, Hangzhou, China, in 1991 and 1996, respectively. Her research area includes video coding, multimedia communication, and relative ASIC design.

Hualong Yu is pursuing his PhD in the College of Information Science and Electronic Engineering at Zhejiang University, Hangzhou, China. He received his BS degree in Information and Communication Engineering from Zhejiang University in 2013. Currently, he is a member of Multimedia and Communication Laboratory at Zhejiang University and his research interests include video coding, image processing, and media streaming.

Jue Mao received her BS degree in the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China, in 2014. She has been awarded excellent undergraduate in Zhejiang University, 2014. She is currently a PhD candidate in Zhejiang University, supervised by Professor Lu Yu. Her research interests include video compression, image, and video processing.

Yin Zhao is with Huawei Technologies Co., Ltd., Hangzhou, China. He received his BS and PhD degrees in information engineering from Zhejiang University, Hangzhou, China, in 2008 and 2013, respectively. He was a visiting student at Nanyang Technological University, Singapore. His research interests include 3-D video processing, video quality assessment, and video coding.