Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Long-term prediction for hierarchical-B-picture-based coding of video with repeated shots*

Xu-guang ZUO, Lu YU†‡

Zhejiang Provincial Key Laboratory of Information Processing, Communication and Networking (IPCAN), Institute of Information and Communication Engineering, Zhejiang University, Hangzhou 310027, China

†E-mail: yul@zju.edu.cn

Received Sept. 13, 2016; Revision accepted Dec. 2, 2016; Crosschecked Mar. 15, 2018

Abstract: The latest video coding standard High Efficiency Video Coding (HEVC) can achieve much higher coding efficiency than previous video coding standards. Particularly, by exploiting the hierarchical B-picture prediction structure, temporal redundancy among neighbor frames is eliminated remarkably well. In practice, videos available to consumers usually contain many repeated shots, such as TV series, movies, and talk shows. According to our observations, when these videos are encoded by HEVC with the hierarchical B-picture structure, the temporal correlation in each shot is well exploited. However, the long-term correlation between repeated shots has not been used. We propose a long-term prediction (LTP) scheme to use the long-term temporal correlation between correlated shots in a video. The long-term reference (LTR) frames of a source video are chosen by clustering similar shots and extracting the representative frames, and a modified hierarchical B-picture coding structure based on an LTR frame is introduced to support long-term temporal prediction. An adaptive quantization method is further designed for LTR frames to improve the overall video coding efficiency. Experimental results show that up to 22.86% coding gain can be achieved using the new coding scheme.

Key words: High Efficiency Video Coding (HEVC); Long-term temporal correlation; Long-term prediction; Hierarchical B-picture structure

https://doi.org/10.1631/FITEE.1601552 **CLC number:** TN919.8

1 Introduction

In recent years, the number of TV series and movies has had an explosion in growth, which presents an enormous challenge for video coding technology. Generally, a TV episode or a movie is composed of a series of meaningful story units. At the lower level, each story unit is a chain of shots that communicate a unified action with a common locale and time (Vendrig and Worring, 2002). In a story unit, it is common that some of the shots appear alternately. The long-term temporal correlation between repeated shots can be exploited to improve video coding efficiency.

High Efficiency Video Coding (HEVC) (Sullivan et al., 2012) is the latest video coding standard developed by the Joint Collaborative Team on Video Coding (JCT-VC). By adopting various high-efficiency video coding tools, it achieves an improvement in coding efficiency of about 50% relative to the prior standard, H. 264/AVC (Wiegand et al., 2003). Since the standardization of HEVC, great efforts have been made to optimize the HEVC encoder for real applications. A number of fast algorithms have been proposed to accelerate the video coding process, such as fast coding unit size decision (Lee et al., 2015), fast mode decision (Pan et al., 2014;

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (No. 61371162)

ORCID: Xu-guang ZUO, http://orcid.org/0000-0002-2868-434X; Lu YU, http://orcid.org/0000-0002-0550-7754

Hu and Yang, 2015), and fast motion estimation (Zuo and Yu, 2015; Pan et al., 2016a, 2016b). Meanwhile, some other researchers (Gao et al., 2016; Li et al., 2016) proposed to optimize encoder parameters according to video content and achieve higher coding efficiency. These studies inspired us to explore a way to make HEVC contribute more to specific videos.

When employing the hierarchical B-picture prediction structure (Schwarz et al., 2007; Pan et al., 2016c, 2016d; Rosewarne et al., 2016) in video coding, the temporal correlation between adjacent frames can be easily removed. The hierarchical B-picture prediction structure shows higher performance than other coding structures such as 'IPPP…' and 'IBBP…' because it uses hierarchical reference frame selection and hierarchical quantization. In the random access common test condition of HEVC (Bossen, 2013), the hierarchical B-picture coding structure is applied to each group of pictures (GOP) with a duration of eight frames (Fig. 1).

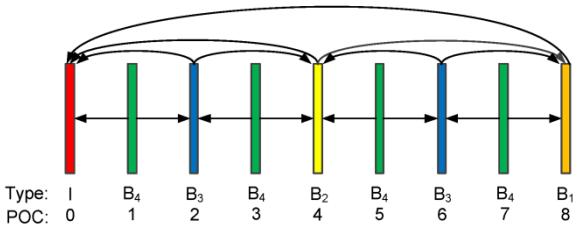

Fig. 1 Referencing of the hierarchical B-picture structure where eight frames are coded as a group of pictures (GOP)

The frames in a GOP are all B frames and are distinguished into four levels, which are marked by different colors. The quantization parameter (QP) is increased by one, from one hierarchy level to the next. The arrows indicate predictive coding with the arrow head pointing towards the reference frame. The reference frames of a frame are either the intra-frame or the referenced B frames from lower levels. As a result, the short-term temporal correlation between neighbor frames can be removed. References to color refer to the online version of this figure

However, if a scene change occurs within a GOP (referred to as SCGOP hereafter), the coding efficiency of frames in the new shot will degrade since they cannot achieve efficient inter prediction from the previous shot. A smart encoder may encode the scene change frame as an intra-frame and then start a new GOP. This adaptive GOP (AGOP) structure can provide good reference for the following GOPs in the new shot (Alfonso et al., 2006; Lenka et al., 2018). However, it takes too many bits to encode the scene change frames. Thus, if the new shot is a repeated shot and we can find a way to make it reference previous similar shots, the coding efficiency of the SCGOP can be further improved.

Long-term reference (LTR) is supported in HEVC to use long-term temporal correlation in videos. HEVC allows up to 32 candidate LTR frames for a sequence. It does not specify the method for choosing LTR frames, which gives the encoder flexibility in choosing the LTR frames according to the applications. Previous work has always focused on selecting LTR frames for a long shot. For example, a couple of studies (Tiwari and Cosman, 2008; Liu et al., 2010) explored the optimal LTR frame updating interval in dual frame applications, where only two reference frames, a short-term reference (STR) frame and an LTR frame, are used. The LTR frame is updated when the content change in the shot accumulates to a certain degree to ensure the prediction quality. However, these methods are not suitable for the hierarchical B-picture structure since the importance of the LTR frame declines with the use of multiple STR frames. Another class of studies tried to use a background frame as the LTR frame for videos captured with stable cameras. The background frame was generated by classical background modeling algorithms such as mean shift (Zhang et al., 2010), the running average method (Zhang et al., 2012, 2014), and the Gaussian mixture model (Paul et al., 2011, 2014). It can significantly improve the efficiency of predicting the exposed background regions. However, a large portion of shots in TV series and movies are not captured by stable cameras. It is infeasible to generate a ground truth background frame for LTR. For a story unit from films or TV shows, there is much long-term temporal correlation between the repeated shots. To our knowledge, no work has been done to use the correlation between repeated shots in video coding.

In each shot, as the temporal correlation has been well exploited by the hierarchical B-picture structure, it is observed that long-term prediction (LTP) cannot achieve much coding gain. However, the coding efficiency of SCGOPs is still not high. In this study, we propose an LTP scheme for the coding of SCGOPs in

videos containing repeated shots. With the proposed LTP scheme, the long-term correlation between repeated shots can be exploited and the video coding efficiency can be improved significantly. The main contributions of this study are as follows: (1) A modified hierarchical B-picture coding structure based on an LTR frame is designed for SCGOP to enable long-term temporal prediction; (2) A clustering-based LTR frame selection algorithm is introduced to maximize the correlation between LTR frames and encoding video; (3) A quality adaptive coding method is proposed for the LTR frames to improve the overall video coding efficiency.

2 Related discussion on video coding with hierarchical B-picture structure

In this section, the coding of videos with repeated shots using the hierarchical B-picture coding structure is analyzed to explain the motivation for the proposed work. We will use four test sequences: Bigbang, Cards, Emperor, and Girls. The test sequences are extracted from four TV series' episodes and they each represent a story unit that is constituted by a group of shots. The details of the sequences can be found in Section 4. In each sequence, there are some shots appearing more than once. Fig. 2 depicts the shots of sequence Emperor, which is from *Empresses in the Palace* (Zheng, 2012), as an example.

2.1 Coding of SCGOP

It is known that the hierarchical B-picture structure can make full use of the temporal correlation among neighboring frames. However, in SCGOP there is hardly any correlation between frames in the new shot and the old shot. Fig. 3a illustrates an SCGOP (GOP *I*) in which the scene change occurs at frame *F*. Frames *F* and *F*+4 still reference the frame from the previous shot. Specifically, for frame *F*+4, the only reference frame is from the old shot and shares little similarity with it, which will lead to poor prediction efficiency. Most regions in frame *F*+4 are coded with intra mode. The same phenomenon can be observed in the last frame of all SCGOPs. We encode the sequences with the HEVC test model HM12.1 under the random access common test condition. Only the first frame is coded as an intra-frame. Table 1 shows the percentages of intra-coded regions in the last frame of SCGOP under four QPs {22, 27, 32, 37}. On average, an area of about 95.3% of the last frame of SCGOP is intra- coded, which indicates the low coding efficiency of the frame.

The aforementioned AGOP method solves the above problem by directly inserting an intra-frame at the scene change position. As in the example shown in Fig. 3b, frame *F* is coded as an intra-frame. The

Table 1 Percentages of intra-coded regions in the last frame of SCGOP under four quantization parameters

Sequence	Percentage of intra-coded regions				Average
	22	27	32	37	
Bigbang	99.5%	99.2%	98.7%	97.9%	98.8%
Cards	94.7%	94.0%	93.0%	91.7%	93.4%
Emperor	95.0%	94.4%	93.1%	91.9%	93.6%
Girls	95.7%	95.5%	95.4%	95.0%	95.4%
Average	96.2%	95.8%	95.1%	94.1%	**95.3%**

Fig. 2 The first frame of each shot (i.e., scene change frame) in Emperor

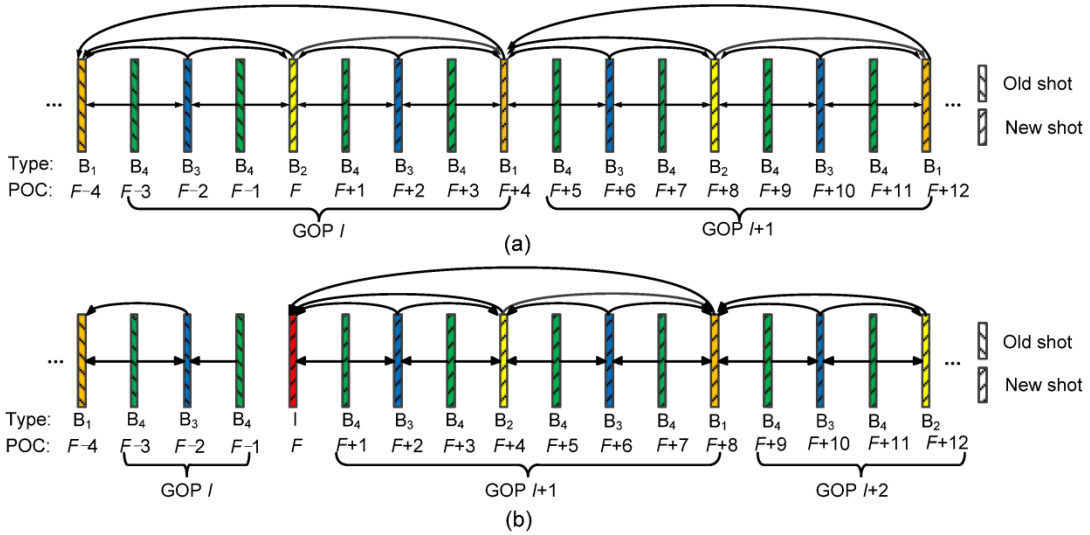

Fig. 3 The coding structures of SCGOP (GOP *I*) (a) and AGOP (b) (References to color refer to the online version of this figure)

intra-frame can refresh video quality and provide better reference for following GOPs in the new shot. In this way, the coding efficiency of the new shot can be improved. However, an intra-frame requires several times the number of bits compared with an inter-frame of the same quality. Thus, we can conclude that although the AGOP method can improve the video coding efficiency, the coding cost of the scene change frames is still too high. Observing that some shots appear alternately within a story unit, we therefore explore the use of LTR to assist the coding of repeated shots in the rest of this study. For example, it can be seen from Fig. 2 that shots 2, 4, 6, 8, 10, 12, 14, 20, 24, and 26 in Emperor share very high similarity. By choosing a frame from one shot as a reference, the coding efficiency of SCGOPs of the other shots can be improved by LTP.

2.2 Long-term extension of the AGOP method

For the AGOP method, an LTP scheme can be implemented using the scene change frame as the LTR frame for the following frames in the same shot. This long-term extension can exploit the long-term temporal correlation in each shot. We use BD-rate (Bjontegaard, 2001) to evaluate the performance of the AGOP LTP scheme, and the results are shown in Table 2. Compared to AGOP, the AGOP LTP scheme can achieve a 0.9% coding gain on average. Overall, the coding gain is very low and can be neglected. It gives indirect evidence that the temporal correlation in a shot has been well exploited by the hierarchical B-picture structure. It does not count for much to further exploit the AGOP LTP scheme. As a result, we pay our attention to exploiting only the long-term correlation between repeated shots to improve the coding efficiency of SCGOPs.

Table 2 The coding performance of AGOP LTP compared to AGOP

Sequence	Y BD-rate	U BD-rate	V BD-rate
Bigbang	−1.0%	−1.4%	−1.3%
Cards	−0.4%	−0.6%	0.0%
Emperor	−1.1%	−0.9%	−1.0%
Girls	−1.3%	−1.3%	−1.6%
Average	−0.9%	−1.0%	−1.0%

3 Proposed long-term prediction scheme

Motivated by the above analysis, we propose an LTP scheme for videos with repeated shots. In this scheme, a pre-analysis of encoding videos is required to derive the LTR frames, which are then referenced during video coding. The LTR frames are used only for the coding of the SCGOPs. We therefore introduce an LTR-frame-based coding structure for the SCGOP. The LTR frames are selected by clustering similar shots and extracting the center frames. The LTR frames are also adaptively quantized with a smaller

QP according to their importance, to improve the inter prediction efficiency.

Note that the shots in the encoding videos need to be recognized for LTR frame selection. Also, the positions of the SCGOPs should be provided to the encoder. So, we use scene change detection to count the number of shots (scene change frames) and to determine the positions of the SCGOPs. The scene change detection algorithm is based on luma histogram comparison (Wang and Weng, 2000; Youm and Kim, 2003). First, each frame l of the video sequence is represented by its luma histogram $H(l)$, which can be written as $H(l)=\{h_l^0,h_l^1,...,h_l^{255}\}$, where h_l^p ($0\le p<256$) represents the number of luma components with value p. Then we define the histogram change of each frame l as the absolute histogram difference of frame l and frame $l-1$ and compute it as

$$\mathrm{HC}(l)=\left|H(l)-H(l-1)\right|=\sum_{p=0}^{255}\left|h_l^p-h_{l-1}^p\right|. \quad (1)$$

Finally, we compare the histogram change of each frame l with the average change of the previous frames to judge whether it is a scene change fame, which can be expressed as

$$\begin{cases}\mathrm{HC}(l)>\dfrac{\alpha}{l-m-1}\displaystyle\sum_{i=m+1}^{l-1}\mathrm{HC}(i), & \text{frame } l \text{ is a scene change frame,}\\ \text{else,} & \text{frame } l \text{ is not a scene change frame,}\end{cases} \quad (2)$$

where m is the frame number of the last scene change frame, and α is an empirical constant set as 4. This scene change detection method is simple but effective. It can accurately detect all the shot boundaries in the test sequence. After all shots are determined, our proposed LTP scheme can be applied. In the following, the details of the proposed scheme will be introduced.

3.1 Coding structure of SCGOP based on the LTR frame

In SCGOP, frames in the new shot cannot achieve efficient prediction from the old shot. The AGOP method inserts an intra-frame at the scene change position and starts a new GOP. The intra-frame can refresh the quality and be used as the reference frame for the following GOPs, either directly or indirectly. Following the spirit of the AGOP method, we also start a new GOP at the scene change position and keep the new GOP independent of the previous shot. The difference is that we use an LTR frame for the reference of the new GOP so that no intra-frame is inserted. As a result, the coding structure in Fig. 3a is changed to that shown in Fig. 4. The LTR frame comes from another shot that shares high similarity with the new shot. Frames F to $F+7$ form the new GOP, which is an SCGOP since it contains the scene change frame (frame F). With the optimized coding structure, the bits used to insert an intra-frame are saved. Also, SCGOP can obtain efficient inter prediction from the LTR frame to improve the coding efficiency. Besides, the hierarchical B-picture structure is applied to the new SCGOP to make full use of the short-term temporal correlation.

3.2 Clustering-based LTR frame selection

We choose LTR frames from scene change frames since they share very high similarity with the SCGOPs. The LTR frames should be coded with finer quality to provide more efficient reference, which will also consume many coding bits. So, our strategy is to choose only one LTR frame for a series of correlated shots and encode it with high quality. Using the scene change frame as the representative of each shot, if similar shots can be clustered together, the center frame can serve as a good reference frame of the others. We employ the K-means algorithm to cluster

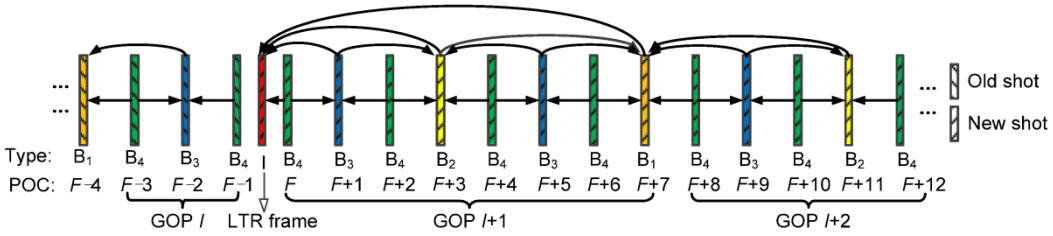

Fig. 4 The coding structure of SCGOP (GOP I+1) with referencing a long-term reference (LTR) frame (References to color refer to the online version of this figure)

the scene change frames, which are signaled as L_n (n=1, 2, …, N), because of its simplicity and efficiency (Hartigan and Wong, 1979; Ngo et al., 2001). Given the clustering number K and the corresponding initial clustering centers μ_k (k=1, 2, …, K), the K-means algorithm is implemented as follows:

Step 1: Classify each frame L_n ($1\le n\le N$) to class $\hat{k}$ with the minimum distance, as

$$\hat{k}=\arg_{1\le k\le K}\min\left(D(\mu_k,L_n)\right), \tag{3}$$

where $D(\mu_k, L_n)$ is the distance between L_n and μ_k.

Step 2: Update the clustering centers. For class k, the center is updated as the frame that has the smallest sum of distances between other frames in class k, which is expressed as

$$\mu_k=L_{i^*}^k=\arg_{1\le i\le n_k}\min\left(\sum_{j=1}^{n_k}D(L_i^k,L_j^k)\right), \tag{4}$$

where n_k is the number of frames belonging to class k, while L_i^k and L_j^k are the i^{th} and j^{th} frames in class k respectively. Repeat steps 1 and 2 until clustering centers μ_k (k=1, 2, …, K) do not change any more.

In Eq. (3), $D(\mu_k, L_n)$ is measured as the sum of absolute residues of L_n with referencing μ_k and calculated as follows. We first partition L_n into $M\times M$ blocks B_t (t=1, 2, …, T), where M=16, and T is the total number of blocks. Then motion estimation is implemented for each block B_t to find its matching prediction block $B_{t,p}$ in μ_k. The fast motion estimation method TZsearch (Tang et al., 2010) is implemented to limit the computational complexity. Finally, $D(\mu_k, L_n)$ is calculated as

$$D(\mu_k,L_n)=\sum_{t=1}^{T}\sum_{x,y=1}^{M}|B_t(x,y)-B_{t,p}(x,y)|, \tag{5}$$

where $B_t(x, y)$ and $B_{t,p}(x, y)$ are pixels at position (x, y) of B_t and $B_{t,p}$, respectively.

The clustering number K needs to be specified explicitly for the K-means algorithm. However, the number of LTR frames that should be chosen is not known in advance. To find the optimal clustering number, clustering options with K varying from 1 to N are all traversed. We calculate the clustering cost of each option, and the minimum cost corresponds to the optimal number of clusters. Besides, the choice of initial clustering centers is essential to the final clustering results and should be paid special attention to. If the initial centers belong to the same cluster, the final results may probably fall into local optima. Therefore, the initial centers should be far apart from each other in distance. To guarantee this, the initial clustering centers of K-cluster clustering should be selected by inheriting the centers of (K−1)-cluster clustering. Using the final clustering centers of (K−1)-cluster clustering μ_1, μ_2, …, μ_{K-1} as the initial clustering centers for K-cluster clustering, the last initial center μ_K is determined as follows:

1. Calculate the maximum distance between the center frame and other frames in the same cluster, which is defined as the intra-cluster distance, of all the K−1 clusters. For cluster k, $1\le k\le K-1$, the frame farthest from μ_k and the corresponding intra-cluster distance are respectively derived as

$$L_{j^*}^k=\arg_{1\le j\le n_k}\max\left(D(\mu_k,L_j^k)\right), \tag{6}$$

$$d_{j^*}^k=D(\mu_k,L_{j^*}^k). \tag{7}$$

2. Compare the intra-cluster distances of the K−1 clusters, and choose the frame corresponding to the largest intra-cluster distance as the new initial clustering center μ_K, which is written as

$$\mu_K=L_{j^*}^{k^*}=\arg_{1\le k\le K-1}\max\left(d_{j^*}^k\right). \tag{8}$$

To improve the overall coding efficiency, we set the goal of selecting the LTR frames as minimizing the amount of coding information. Thus, we define the cost of each clustering option as the amount of information of all clusters, which is measured by inter/intra prediction residues. For K-cluster clustering, the clustering cost is computed as

$$\text{Cost}(K)=\sum_{k=1}^{K}C_k, \tag{9}$$

$$C_k=D_{\text{intra}}(\mu_k)+\sum_{i=1}^{n_k}D(\mu_k,L_i^k), \tag{10}$$

where C_k is the cost of the k^{th} cluster, $D_{\text{intra}}(\mu_k)$ is the sum of absolute residues of μ_k coded with all intra modes. To calculate $D_{\text{intra}}(\mu_k)$, μ_k is also divided into $M\times M$ (M=16) blocks B_t, t=1, 2, …, T. Then intra

prediction is executed in each block to derive the optimal intra prediction mode and corresponding prediction residues. Note that we do not encode the blocks, and the original pixels of neighboring blocks are used for prediction. In this way, the computation complexity for deriving intra prediction residues can be reduced. Finally, $D_{\text{intra}}(\mu_k)$ can be written as

$$D_{\text{intra}}(\mu_k)=\sum_{t=1}^{T}\text{SAD}(B_t,\text{Mode}_t), \tag{11}$$

where SAD(B_t, Mode$_t$) is the sum of absolute residues of block B_t predicted with the optimal intra prediction mode Mode$_t$. By integrating Eqs. (5), (10), and (11) into Eq. (9), the clustering cost Cost(K) can be derived. Traversing 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. For example, the clustering curves of Emperor are shown in Fig. 5. The optimal clustering number K_{opt} is six. Classifying the scene change frames into six clusters, the corresponding clustering results are shown in Table 3. Finally, the center frame of each cluster is chosen as the LTR frame. For clusters including only one frame, the center frames are not extracted since they have little correlation with other frames. In the example, frames L_1, L_8, L_{17}, L_{21}, L_{16}, and L_{28} are the center frames of the six clusters. As frame L_{28} is the only frame in cluster 6, frames L_1, L_8, L_{17}, L_{21}, and L_{16} are selected as the LTR frames.

3.3 LTR frame QP calculation

Hierarchical quantization is employed by each GOP in the hierarchical B-picture coding structure. Supposing the QP of the lowest level in a GOP is QP$_L$, we use ΔQP to denote the QP difference between QP$_L$ and the QP of the LTR frame. Then the QP of the LTR frame can be expressed as QP$_L$−ΔQP. Increasing ΔQP will improve the coding quality of the LTR frame, which can provide a better reference for the SCGOPs. As a result, the bits produced by the SCGOPs can be reduced. However, with a larger ΔQP, there will also be a larger number of coding bits produced by encoding the LTR frame. To improve the overall coding efficiency, it should be ensured that the additional bits of the LTR frame with a smaller QP should be fewer than the bits saved on the SCGOPs that reference the LTR frame. In other words, if many coding bits of the SCGOPs can be saved by referencing the LTR frame, we should employ a relatively large ΔQP, and vice versa. The more SCGOPs by which an LTR frame is referenced, the more bits it can save. So, we make the values of ΔQP adaptive to the number of times the LTR frame is referenced.

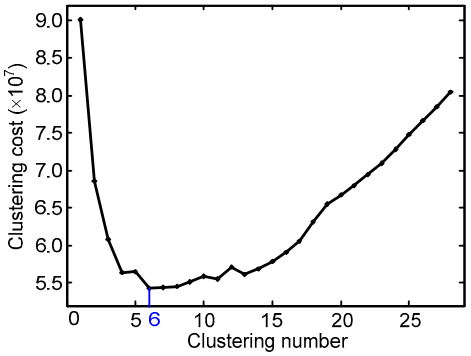

Fig. 5 The curve of clustering cost relative to the clustering number of Emperor

Table 3 The clustering results of Emperor

Cluster	Frame(s)
Class 1	L_1^*, L_3
Class 2	L_2, L_4, L_6, L_8^*, L_{10}, L_{12}, L_{14}, L_{20}, L_{24}, L_{26}
Class 3	L_5, L_7, L_{11}, L_{13}, L_{15}, L_{17}^*, L_{19}, L_{23}, L_{25}, L_{27}
Class 4	L_9, L_{21}^*
Class 5	L_{16}^*, L_{18}, L_{22}
Class 6	L_{28}^*

Frames marked with '*' are center frames

In the test sequences used in Section 2, there are four shots, which are from Bigbang, Cards, and Emperor, repeated 10 times or more. We use these four shots to experimentally find the optimal ΔQP values for the LTR frames. For each shot, the LTR frame is selected according to the method introduced in Section 3.2 first. Then we encode n SCGOPs by referencing the LTR frame each time with ΔQP varying from 0 to 10. Using a ΔQP value of zero as the anchor, the rate-distortion performance of the n SCGOPs with all ΔQP values can be evaluated. We can draw a curve of the average coding performance relative to ΔQP of n SCGOPs from the four shots. Traversing the value of n from 2 to 10, the curves of n with all values can be drawn (Fig. 6). For each n value, there is an optimal ΔQP with the highest coding performance. It can be seen that, with the increase of n, the optimal ΔQP

increases. Thus, the relationship of the optimal ΔQP relative to n can be derived (Fig. 7). Note that n has another meaning, which is the number of times the LTR frame is referenced by the SCGOPs. For each LTR frame, the value of n is set as the number of frames in the cluster to which it belongs. Finally, we set the value of ΔQP for each LTR frame according to the number of times it is referenced, expressed as

$$\Delta\mathrm{QP}=\begin{cases}1, & n=2,\\ 3, & n=3,\ 4,\\ 4, & n=5,\ 6,\\ 5, & n=7,\ 8,\\ 6, & n\geq 9.\end{cases} \tag{12}$$

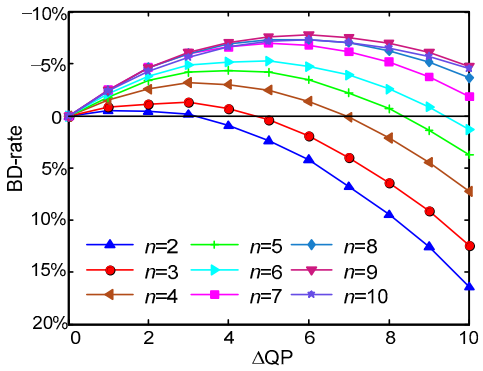

Fig. 6 The curves of coding performance (BD-rate) relative to ΔQP of different n values

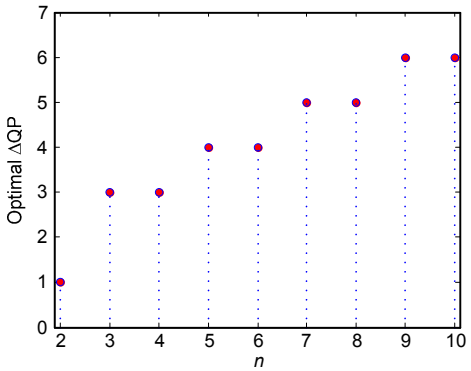

Fig. 7 The optimal ΔQP relative to different n values

In our experiment, the maximum value of n is set as 10, as the shots in a story unit are rarely repeated more than 10 times according to our observation. For simplicity, we set the value of ΔQP as six when the LTR frame is referenced more than 10 times.

4 Experimental results

4.1 Experimental setup

We employ eight sequences containing repeated shots to verify the performance of the proposed scheme. The sequences are extracted from the TV series: *The Big Bang Theory* (Cendrowski, 2013), *House of Cards* (Dahl, 2015), *Empresses in the Palace* (Zheng, 2012), *2 Broke Girls* (Scardino, 2015), *Once Upon a Time* (Tirone, 2015), *Sherlock* (McCarthy, 2014), and *Game of Thrones* (Nutter, 2012). Each sequence is a story unit constituted by a group of shots. A brief introduction of the sequences is shown in Table 4. The lengths of the test sequences vary from tens of seconds to more than one hundred seconds. There are at least 20 shots in each sequence and even up to 69 in *Sherlock*. Some of the shots appear more than once.

Table 4 Description of test sequences*

Sequence	Length (s)	Number of shots
Bigbang	85	27
Cards	141	21
Emperor	94	28
Girls	83	20
Time	82	33
QueenMother	81	32
Sherlock	182	69
Throne	173	46

* For all these sequences, the frame rate is 25 Hz, and the resolution is 640×360

The proposed LTP scheme is implemented on HM12.1 to evaluate its performance. HEVC using the hierarchical B-picture coding structure with a fixed GOP (FGOP) size is used as the anchor. Besides, the aforementioned AGOP method and its long-term extension AGOP LTP are tested for comparison. All schemes are tested under the RA common test condition (Table 5). Note that the intra period is set as the length of the test sequence for the anchor. When no scene change occurs, the insertion of intra-frames will degrade the coding performance. Thus, only the first frame is coded as an intra-frame. For the AGOP method and its long-term extension, the scene change frames (the first frame included) are coded as intra-frames. For the proposed LTP scheme, only the LTR frames are coded as intra-frames. Each sequence is encoded with four QP values of 22, 27, 32, and 37. All

the experiments are conducted on an Intel® Xeon® CPU E5-26900@2.90 GHz with 190 GB RAM memory.

Table 5 The random-access configurations of HM

Parameter	Value	Parameter	Value
Profile	Main	AMP	Enabled
Frame structure	Hierarchical B	Hadmard ME	Enabled
		SAO	Enabled
GOP size	8	RDOQ	Enabled
ME range	64	RDOQTS	Enabled
Fast search	Enabled	Rate control	Disabled

In the anchor, as only the first frame of a story unit is coded as the intra-frame, a video (e.g., a movie or TV show) can be randomly accessed at the level of the story unit. This is acceptable in general applications like stored videos (e.g., DVD and BD) and video-on-demand streaming. For the AGOP method, random access at the level of the shot can be supported since the first frames of each shot are coded as intra-frames. In the proposed scheme, as the LTR frames are intra-coded, each shot can also be randomly accessed with its LTR frame available. As a result, the AGOP method and the proposed scheme can support finer random access than the anchor.

4.2 LTR frame selection results

By traversing all the clustering options, the optimal clustering numbers of each sequence are derived. Fig. 8 shows the curves of clustering cost relative to the clustering number of all sequences (except for sequence Emperor, the curve of which has been shown in Fig. 5). For each sequence, there is a minimum in the curve, corresponding to the optimal clustering number. The clustering numbers of each sequence are shown in Table 6. Given the clustering number, the scene change frames of each sequence are classified into a smaller number of clusters. Since only the center frames of those clusters containing more than one frame are chosen as LTR frames, a further smaller number of LTR frames are extracted, as also shown in Table 6.

4.3 Performance of the proposed LTP scheme

BD-rate and BD-PSNR are used to evaluate the coding performance of the proposed scheme. Note that the coding performance is evaluated for the whole sequence. The performances of AGOP, AGOP LTP, and the proposed LTP scheme are summarized in Table 7. The AGOP method can improve the coding efficiency by 0.10 dB on average when compared with the HEVC anchor. This is why it is currently

Table 6 The numbers of scene changes, clusters, and long-term reference (LTR) frames

Sequence	Number of scene changes	Number of clusters	Number of LTR frames
Bigbang	27	6	3
Cards	21	6	3
Emperor	28	6	5
Girls	20	10	6
Time	33	13	5
QueenMother	32	14	7
Sherlock	69	19	11
Throne	46	11	8

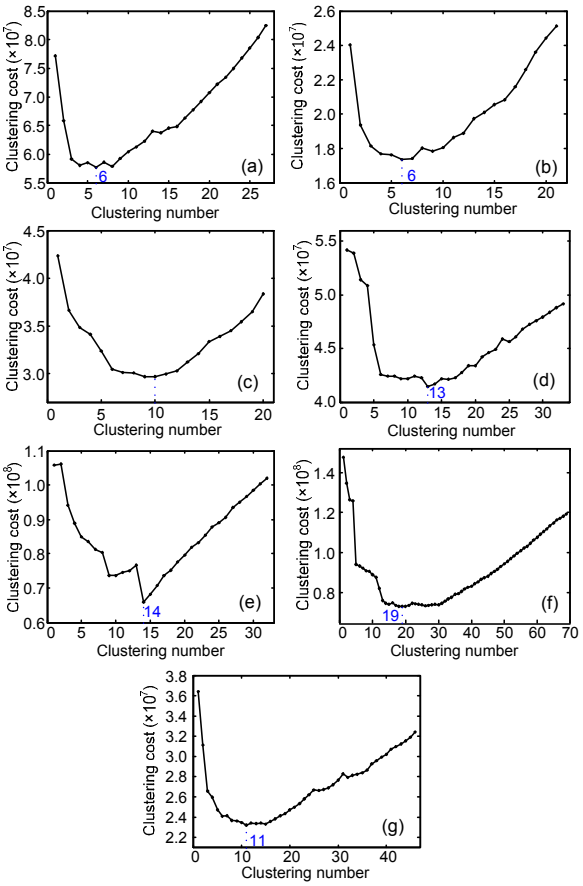

Fig. 8 The curves of clustering cost relative to the clustering number of all test sequences (except for Emperor)
(a) Bigbang; (b) Cards; (c) Girls; (d) Time; (e) QueenMother; (f) Sherlock; (g) Throne

Table 7 The coding performance of AGOP, AGOP LTP, and the proposed LTP scheme

Sequence	BD-PSNR (dB)			BD-rate		
	AGOP	AGOP LTP	Proposed LTP	AGOP	AGOP LTP	Proposed LTP
Bigbang	0.05	0.10	0.48	–1.00%	–1.95%	–9.14%
Cards	0.18	0.20	0.58	–3.57%	–3.95%	–11.43%
Emperor	0.00	0.05	0.75	–0.03%	–1.13%	–14.71%
Girls	0.21	0.26	0.62	–4.42%	–5.60%	–12.75%
Time	0.00	0.02	0.23	–0.05%	–0.37%	–5.29%
QueenMother	0.01	0.03	1.14	–0.30%	–0.72%	–21.84%
Sherlock	0.15	0.18	1.23	–2.99%	–3.62%	–22.86%
Throne	0.19	0.20	0.66	–4.54%	–4.67%	–15.09%
Average	0.10	0.13	0.71	–2.11%	–2.75%	–14.14%

widely employed in video coding. However, because of the high efficiency of the hierarchical B-picture structure, the AGOP LTP scheme further improves the coding efficiency by 0.03 dB. In contrast, the proposed LTP scheme can exploit long-term temporal correlation between repeated shots. Compared to HEVC FGOP, it can achieve an improvement in coding efficiency of 0.71 dB, which is equivalent to 14.14% bit savings. The proposed scheme significantly outperforms the AGOP scheme and its long-term extension.

Although we use the same four QP values to encode each sequence, the bitrate and PSNR ranges vary in different sequences because of their different content. Fig. 9 compares the rate-distortion curves of the test sequences over a wide range of bitrates. The curves of the AGOP scheme and its LTR extension are hardly distinguishable because of their similar performances. However, the curve of the proposed LTP scheme is obviously above that of the AGOP LTP scheme, which reveals the remarkable performance of the proposed scheme. The bitrates of Cars and Throne locate in a relatively narrow range of 0–80 kb/s because they were both shots in the evening, so the images contain a large portion of dark regions and lack high-frequency components. Although they are coded with a small QP (22) and the PSNR is up to about 48 dB, the bitrate is still less than 100 kb/s. In contrast, the other sequences, especially Bigbang, Emperor, and QueenMother, are full of color and texture. It takes more bits to encode these sequences, and the bitrate can be up to hundreds of kb/s. Overall, the proposed LTP scheme is applicable to different kinds of sequences and shows advantage over a wide range of bitrates.

4.4 Computational complexity

Fig. 10 shows the computational time results of the FGOP anchor, AGOP scheme, AGOP LTP scheme, and our LTP scheme. Compared with the FGOP anchor, the AGOP method can save 6.4% of encoding time. This is because the scene change frames are encoded as intra-frames and consume less encoding time in the AGOP scheme. Also, the number of reference frames in the first GOP in each shot is reduced.

Compared with AGOP, its long-term extension AGOP LTP requires 24% more computational time. Because an extra LTR frame is exploited for inter prediction, this increases the complexity of motion estimation. In the proposed LTP scheme, as the LTR frames are used only for the reference of the SCGOPs, there is only 2.2% complexity increase for video coding. The proposed LTP scheme also needs extra computation for clustering to select the LTR frames. As only dozens of scene change frames are contained in a story unit, the computation time cost for clustering is limited. According to our test, the selection of LTR frames increases the computational time by only 1.2%. Overall, an additional 3.4% of computational complexity is imposed by the proposed scheme compared to the AGOP method. Considering the high performance of the proposed scheme, the complexity increase is acceptable. Besides, compared with the AGOP LTP scheme, the proposed LTP scheme consumes much less computational time but achieves much higher coding efficiency.

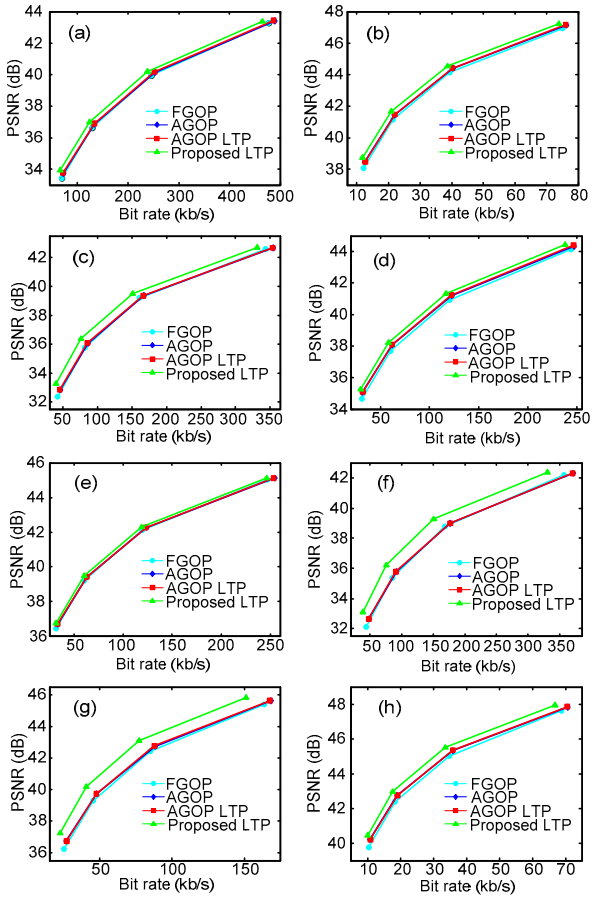

Fig. 9 The rate-distortion curves of all test sequences
(a) Bigbang; (b) Cards; (c) Emperor; (d) Girls; (e) Time; (f) QueenMother; (g) Sherlock; (h) Throne

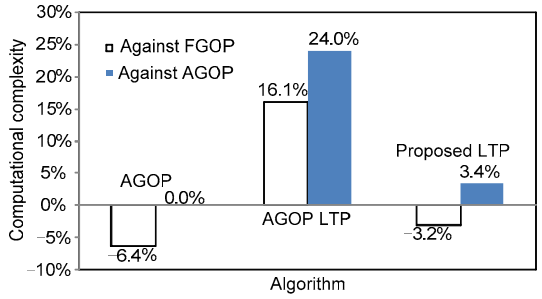

Fig. 10 The computational complexity of AGOP, AGOP LTP, and the proposed LTP

5 Conclusions

An LTP scheme has been proposed for the coding of videos with repeated shots using a hierarchical B-picture coding structure. The proposed scheme can use long-term temporal correlations between repeated shots. In the scheme, LTR frames are chosen by clustering similar shots in a video and encoded with adaptive quality according to their importance. They are referenced only by the SCGOPs to improve the video coding efficiency while controlling the computational complexity. The overall experimental results show that the proposed scheme improves the coding efficiency from 5% to 21% compared with the relevant AGOP LTP scheme, over a wide range of bitrates and for a large number of test sequences. Meanwhile, compared with AGOP LTP, the proposed scheme involves less computational complexity.

Since the proposed LTP scheme is based on a pre-analysis of a video sequence, it is not appropriate for real-time applications. However, it is worthwhile for stored video applications and video-on-demand streaming, where videos are encoded in advance. In real-time applications, the selection of LTR frames can depend only on those shots that have already been encoded and each shot can reference only the LTR frame from the preceding shots. Because of these specific characteristics, we will design a new LTP scheme for real-time applications in the future.

References

Alfonso D, Biffi B, Pezzoni L, 2006. Adaptive GOP size control in H.264/AVC encoding based on scene change detection. Proc 7th Nordic Signal Processing Symp, p.86-89. https://doi.org/10.1109/NORSIG.2006.275283

Bjontegaard G, 2001. Calculation of average PSNR differences between RD curves. Document VCEG-M33. Austin, TX, USA.

Bossen F, 2013. Common HM test conditions and software reference configurations. Document JCT-VC L1100. Geneva, Switzerland.

Cendrowski M, 2013. The Hofstadter Insufficiency. The Big Bang Theory. DVD. Season 7. Episode 1. CBS.

Dahl J, 2015. Chapter 33. House of Cards. DVD. Season 3, Episode 7. Netflix.

Gao YB, Zhu C, Li S, 2016. Hierarchical temporal dependent rate-distortion optimization for low-delay coding. Proc IEEE Int Symp on Circuits and Systems, p.570-573. https://doi.org/10.1109/ISCAS.2016.7527304

Hartigan JA, Wong MA, 1979. Algorithm AS 136: a *K*-means clustering algorithm. *J R Stat Soc*, 28(1):100-108. https://doi.org/10.2307/2346830

Hu N, Yang EH, 2015. Fast mode selection for HEVC intra-frame coding with entropy coding refinement based on a transparent composite model. *IEEE Trans Circ Syst Video Technol*, 25(9):1521-1532. https://doi.org/10.1109/TCSVT.2015.2395772

Lee J, Kim S, Lim K, et al., 2015. A fast CU size decision algorithm for HEVC. *IEEE Trans Circ Syst Video Technol*, 25(3):411-421. https://doi.org/10.1109/TCSVT.2014.2339612

Lenka K, Jaroslav P, Michal M, 2018. Adaptive group of pictures structure based on the positions of video cuts. Proc World Academy of Science, Engineering and Technology, p.377-380.

Li S, Zhu C, Gao YB, et al., 2016. Lagrangian multiplier adaptation for rate-distortion optimization with inter-frame dependency. *IEEE Trans Circ Syst Video Technol*, 26(1): 117-129. https://doi.org/10.1109/TCSVT.2015.2450131

Liu D, Zhao DB, Ji XY, et al., 2010. Dual frame motion compensation with optimal long-term reference frame selection and bit allocation. *IEEE Trans Circ Syst Video Technol*, 20(3):325-339. https://doi.org/10.1109/TCSVT.2009.2031442

McCarthy C, 2014. The Sign of Three. Sherlock. DVD. Season 3, Episode 2. BBC.

Ngo CW, Pong TC, Zhang HJ, 2001. On clustering and retrieval of video shots. Proc 9th ACM Int Conf on Multimedia, p.51-60. https://doi.org/10.1145/500141.500151

Nutter D, 2012. A Man Without Honor. Game of Thrones. DVD. Season 2, Episode 7. HBO.

Pan ZQ, Kwong S, Sun MT, et al., 2014. Early MERGE mode decision based on motion estimation and hierarchical depth correlation for HEVC. *IEEE Trans Broadcast*, 60(2):405-412. https://doi.org/10.1109/TBC.2014.2321682

Pan ZQ, Zhang Y, Lei JJ, et al., 2016a. Early DIRECT mode decision based on all-zero block and rate distortion cost for multiview video coding. *IET Image Process*, 10(1): 9-15. https://doi.org/10.1049/iet-ipr.2014.1018

Pan ZQ, Zhang Y, Kwong S, 2016b. Fast mode decision based on texture–depth correlation and motion prediction for multiview depth video coding. *J Real-Time Image Process*, 11(1):27-36. https://doi.org/10.1007/s11554-013-0328-3

Pan ZQ, Lei JJ, Zhang Y, et al., 2016c. Fast motion estimation based on content property for low-complexity H.265/HEVC encoder. *IEEE Trans Broadcast*, 62(3):675-684. https://doi.org/10.1109/TBC.2016.2580920

Pan ZQ, Jin P, Lei JJ, et al., 2016d. Fast reference frame selection based on content similarity for low complexity HEVC encoder. *J Vis Commun Image Represent*, 40:516-524. https://doi.org/10.1016/j.jvcir.2016.07.018

Paul M, Lin WS, Lau CT, et al., 2011. Explore and model better I-frames for video coding. *IEEE Trans Circ Syst Video Technol*, 21(9):1242-1254. https://doi.org/10.1109/TCSVT.2011.2138750

Paul M, Lin WS, Lau CT, et al., 2014. A long-term reference frame for hierarchical B-picture-based video coding. *IEEE Trans Circ Syst Video Technol*, 24(10):1729-1742. https://doi.org/10.1109/TCSVT.2014.2302555

Rosewarne C, Bross B, Naccari M, et al., 2016. High Efficiency Video Coding (HEVC) Test Model 16 (HM 16). Document JCTVC-X1002. Geneva, Switzerland.

Scardino D, 2015. And the Show and Don't Tell. 2 Broke Girls. DVD. Season 5, Episode 17. CBS.

Schwarz H, Marpe D, Wiegand T, 2007. Overview of the scalable video coding extension of the H.264/AVC standard. *IEEE Trans Circ Syst Video Technol*, 17(9):1103-1120. https://doi.org/10.1109/TCSVT.2007.905532

Sullivan GJ, Ohm JR, Han WJ, et al., 2012. Overview of the High Efficiency Video Coding (HEVC) standard. *IEEE Trans Circ Syst Video Technol*, 22(12):1649-1668. https://doi.org/10.1109/TCSVT.2012.2221191

Tang XL, Dai SK, Cai CH, 2010. An analysis of TZSearch algorithm in JMVC. Proc IEEE Int Conf on Green Circuits and Systems, p.516-520. https://doi.org/10.1109/ICGCS.2010.5543008

Tirone R, 2015. The Price. Once Upon a Time. DVD. Season 5, Episode 2. ABC.

Tiwari M, Cosman PC, 2008. Selection of long-term reference frames in dual-frame video coding using simulated annealing. *IEEE Signal Process Lett*, 15:249-252. https://doi.org/10.1109/LSP.2007.914928

Vendrig J, Worring M, 2002. Systematic evaluation of logical story unit segmentation. *IEEE Trans Multim*, 4(4):492-499. https://doi.org/10.1109/TMM.2002.802021

Wang XY, Weng ZK, 2000. Scene abrupt change detection. Proc IEEE Conf on Electrical and Computer Engineering, p.880-883. https://doi.org/10.1109/CCECE.2000.849592

Wiegand T, Sullivan GJ, Bjontegaard G, et al., 2003. Overview of the H.264/AVC video coding standard. *IEEE Trans Circ Syst Video Technol*, 13(7):560-576. https://doi.org/10.1109/TCSVT.2003.815165

Youm S, Kim W, 2003. Dynamic threshold method for scene change detection. Proc IEEE Int Conf on Multimedia and Expo, p.337-340. https://doi.org/10.1109/ICME.2003.1221622

Zhang XG, Liang LH, Huang H, et al., 2010. An efficient coding scheme for surveillance videos captured by stationary cameras. Proc SPIE Visual Communications and Image Processing, p.1-10. https://doi.org/10.1117/12.863522

Zhang XG, Tian YH, Huang TJ, et al., 2012. Low-complexity and high-efficiency background modeling for surveillance video coding. Proc IEEE Visual Communications and Image Processing, p.769-784. https://doi.org/10.1109/VCIP.2012.6410796

Zhang XG, Huang TJ, Tian YH, et al., 2014. Background-modeling-based adaptive prediction for surveillance video coding. *IEEE Trans Image Process*, 23(2): 769-784. https://doi.org/10.1109/TIP.2013.2294549

Zheng XL, 2012. Empresses in the Palace. DVD. Beijing Television Arts Centre, Beijing, China (in Chinese).

Zuo XG, Yu L, 2015. A novel interpolation-free scheme for fractional pixel motion estimation. Proc Picture Coding Symp, p.80-84. https://doi.org/10.1109/PCS.2015.7170051