


# Background Modeling and Referencing for Moving Cameras-Captured Surveillance Video Coding in HEVC

Gang Wang , Bo Li, Yongfei Zhang , *Member, IEEE*, and Jinhui Yang

**Abstract**—Surveillance video coding is crucial for improving compression efficiency in intelligent video surveillance systems and applications. Plenty of work has been done, which can be roughly divided into two categories: the former mainly focuses on low-complexity background modeling to obtain the clear background, while the latter focuses on an appropriate coding strategy to generate the high-quality background reference picture for effective background prediction. However, almost all existing works focus only on stationary camera scenes, while moving cameras-captured surveillance video coding is left untouched and is still an open problem. In this paper, a background modeling and referencing scheme for moving cameras-captured surveillance video coding in high-efficiency video coding (HEVC) is proposed. First, this paper proposes a low-complexity motion background modeling algorithm for surveillance video coding using the running average based on a global-motion-compensation method. To obtain the global motion vector, we propose a global motion detection method based on character blocks by establishing a low-rank singular value decomposition model for clustering and estimating motion vectors of background character blocks in the cameras movement circumstance. Second, we propose a background referencing coding strategy, in which the motion background coding tree units (MBCTUs) would be selected by anchoring the input video frame on the modeling background frame and coded with the optimized quantization parameter. Then, the reconstructed MBCTU will be used to update the previous coding tree unit in the global compensation location of the background reference picture. Extensive experimental results show that the proposed scheme can achieve significant bit savings of up to 26.6% and, on average, 6.7% with similar subjective quality and negligible encoding complexity, compared to HM12.0. Besides, the proposed scheme consistently outperforms two state-of-the-art surveillance video coding schemes with remarkable bitrate savings.

**Index Terms**—Surveillance video coding, HEVC, rate-distortion optimization, global motion estimation, motion background modeling, background reference.

## I. INTRODUCTION

IN RECENT years, intelligent video surveillance systems and applications have achieved rapid development and become the important urban infrastructure. This lead to enormous amounts of surveillance video data, which brings great challenges for storage and transmission.

The video compression capability has been improved persistently in every generation video coding standard, and the latest high efficiency video coding (HEVC) can compress video about twice as much as its predecessor H.264/ (AVC) [1]. However, the growth rate of surveillance videos is much higher than the video compression rate that the traditional video compression standard can achieve [2]–[4].

Fortunately, beyond the traditional temporal, spatial, information entropy, visual and structure redundancies, surveillance videos own special data redundancy, i.e., the background redundancy, which is one kind of temporal knowledge redundancy. The video compression performance can be much improved by efficiently removal of the enormous background redundancy [2]. In common videos, temporal redundancy is the major part of redundancies need to be removed in the task of video compression. Inter prediction is a powerful video compress technique which can effectively reduce temporal redundancy. Multi-reference prediction has been used for inter prediction since the H.264/ AVC standard [5]. The multiple references method usually adopts the reconstructed picture coded before as the short-term reference picture. However, this might cause the problem that some coded regions can hardly find matched regions in short-term reference pictures. In scene videos (e.g., surveillance or conference videos), contents of videos exit a lot of background redundancy which might keep unchanged in a long period [6], [7]. Some background regions, called exposed background regions, are unable to find matched regions in short-term reference pictures because of foreground occlusions or temporary disappearance out of the camera. This would decrease the coding efficiency since the inter prediction ineffectively work in the coding process [8].

To take advantage of the special characteristics of surveillance videos, some related methods are proposed to improve

Manuscript received September 23, 2017; revised February 17, 2018; accepted April 3, 2018. Date of publication April 26, 2018; date of current version October 15, 2018. This work was supported in part by the National Key R&D Program of China under Grant 2016YFC0801001, in part by the National Natural Science Foundation of China Key Project under Grant 61632001, and in part by the National Natural Science Foundation of China under Grant 61772054. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Enrico Magli. (*Corresponding author: Yongfei Zhang.*)

G. Wang and J. Yang are with the Beijing Key Laboratory of Digital Media, School of Computer Science and Engineering, Beihang University, Beijing 100191, China (e-mail: smile588@sina.com; jinhuiy@foxmail.com).

B. Li and Y. Zhang are with the Beijing Key Laboratory of Digital Media, School of Computer Science and Engineering and the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing 100191, China (e-mail: boli@buaa.edu.cn; yfzhang@buaa.edu.cn).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TMM.2018.2829163

the compression efficiency for surveillance video coding. As far as the background reference selection methods for surveillance video coding are concerned, surveillance video coding methods can be coarsely classified into two categories: key frame based long-term reference [9], [10] and background-modeling-based frame as the long-term reference [11]–[13]. The key frame based long-term reference utilizes the key frame as the long-term inter prediction to improve prediction efficiency. Background-modeling-based reference for video coding employs the background modeling method to generate the background frame as the reference picture. By this way, the codec will obtain the clearer background reference to match the background block.

Although previous works for surveillance video coding have shown promising results, specifically in static scene videos captured from static camera [12], [13], more and more videos captured by moving cameras in real-world, such as pan-tilt-zoom (PTZ) camera, hand-held camera, the aerial vehicle camera, etc. Since the previous conventional scheme assume that the surveillance cameras are stationary, there exist some intrinsic limitations as follows and thus might lead to severely degraded video coding performance, as will be shown in the performance comparisons in Section VI.

- i) The previous surveillance video coding schemes adopt the static background modeling methods for static cameras. Obviously, the previous methods are designed for static cameras without global motion estimation, which are unsuitable for moving cameras with the motion background. Since the static background modeling methods only model pixels of training pictures at the same temporal position in these schemes, the codec is unable to obtain the clearer background modeling picture in the moving cameras case, which leads to ineffectively matching the background block.
- ii) In previous background reference picture generation works, the background reference picture is generated and updated at the collocated location in the coding strategy. Since the reconstructed selected background CTUs update statically the collocated location in the long-term reference picture, the codec is unable to generate the high-quality background reference picture in moving cameras case.

Thus, in this paper, we address the problem of efficient coding of moving cameras-captured surveillance videos, which also works well for stationary cameras-captured surveillance videos. Furthermore, statistical results [14] show horizontal direction and vertical direction motion of cameras in the fixed platform are the most common camera motions in most surveillance circumstance (with some large view complicated motions can be decomposed as horizontal and vertical movements), this paper mainly focuses on surveillance video coding with horizontal and vertical movements. However, the video captured by handled cameras and airborne sensor have less repetitive background pixels temporally. Therefore, if the background reference picture of the proposed method still be adopted, inter prediction will hardly be improved. Besides, the transmission of the background reference picture will be wasteful. And the camera movement

of zoom-in/-out case is out of scope of this paper and will be addressed as our future work.

More specifically, this paper proposes a background modeling and referencing scheme for moving cameras-captured surveillance video coding. First, we propose a low-complexity motion background modeling algorithm for surveillance video coding using the running average based on global-motion-compensation method. To obtain the global motion vector, we propose a global motion detection method based on character blocks by establishing a low rank singular value decomposition (LR-SVD) model for clustering and estimating motion vectors of background character blocks in the cameras movement circumstance. Second, a background referencing coding strategy, in which the motion background CTUs (MBCTUs) would be selected by anchoring the input video frame on the modeling background by the global-motion-compensation method. The selected MBCTUs would be coded with the optimized quantization parameter (QP). Then the reconstructed MBCTU will be used to update the previous CTU in the global compensation location of the background reference picture.

The rest of this paper is organized as follows. In Section II, we give a brief review of related works. Section III provides a system overview of our proposed background modeling and referencing surveillance video coding scheme. In Section IV, our motion background modeling using the global motion compensation strategy is proposed. In Section V, the motion background referencing coding strategy is designed. In Section VI, experimental results are shown. Finally, the conclusion is drawn in Section VII.

## II. RELATED WORKS

Two key challenges of surveillance video coding are to obtain the clear and high-quality background reference picture, respectively. First, the clear background reference region would provide a better background prediction performance. The background modeling approach is the key technique for building the clear background modeling frame. The exposed background regions would be generated by the background modeling algorithm for subsequent coded pictures. Second, the high-quality background reference picture with smaller QP would provide more reference pixels for inter prediction. This would also improve the background region prediction performance. Therefore, this section firstly introduces static and motion background modeling methods in existing works for video analytics fields, and then reviews prior background reference generation algorithms in video coding.

### A. Background Modeling for Static Cameras and Moving Cameras

In the last decade, literatures have seen a considerable progress in the modeling of background [15], [16]. Some sophisticated background modeling algorithms are proposed [17]–[23], which can be roughly classified into two categories. The former mainly focus on the stationary cameras. The running average approach is the basic background modeling. This approach calculates average values by weight values which

requires lower-cost complexity. Wren *et al.* presented the pixel-wise Gaussian model which can substantially solve the problem for arbitrarily complex but single-person, fixed-camera situations [17]. Stauffer and Grimson proposed the well-known Gaussian mixture model (GMM) [18]. They modeled the variation in background appearance using an underlying mixture of Gaussians. Barnich presented a universal method for background modeling (ViBe) [19]. Their algorithm classified a new pixel value with respect to its immediate neighborhood in the chosen color space by modeling a set of background pixels samples. Although ViBe method can extract moving object quickly, it might lead to the “Ghost” phenomenon in the background modeling frame. However, all these background modeling approaches assume that surveillance video cameras are stationary, which significantly limits the applications of background modeling algorithms, where more and more surveillance videos are captured by moving cameras, such as pan-tilt-zoom (PTZ) camera, hand-held camera, the aerial vehicle camera, etc. Thus the issue of the motion background modeling from these moving cameras has been paid increasing attention in recent years.

The moving camera brings new change for background modeling [24]–[30]. In [31], the proposed method classified each image pixel into planar background, parallax, or motion regions by sequentially applying 2D planar homographies, the epipolar constraint, and a novel geometric constraint. Sheikh *et al.* [32] proposed a rank constraint trajectory pruning method. They estimated a compact trajectory basis from trajectories of salient features and the background was subtracted by removing trajectories that lay within the space spanned by the basis. Then, an optimal pixel-wise foreground/background labeling was obtained using a probabilistic graphical model. Wu *et al.* [33] proposed a moving detection method with a freely moving camera via background subtraction. This method segmented the foreground motion and background motion by performing reduced singular value decomposition.

Similarly, since the previous surveillance video coding schemes adopt the static background modeling methods for static cameras, the motion background modeling algorithm should be considered in the moving cameras case.

### B. Conventional Surveillance Video Coding Schemes

Since more and more videos captured in specific scenes are characterized by temporally redundant background, some works for surveillance video coding schemes are proposed recently [34]–[40]. Paul *et al.* [11] utilized GMM modeled pixels of many pictures at the same position to generate a background picture as the long-term reference. Since Gaussian mixture modeling brings more float computation, this would increase coding time cost. To decrease the GMM modeling complexity, Zhang *et al.* [12] proposed generating a background picture by simply averaging many pictures pixel by pixel for scene video coding. In Zhang’s work, the whole background picture was coded into the stream as a special I-picture with smaller quantization parameter. This may cost a large number of bits to transmit the background picture within a short time, which probably results in a traffic burst and thus packet losses and severe video quality

degradation. To address the problem of traffic burst for the whole background picture, Chen *et al.* [13] proposed a block-composed background reference video coding scheme. In Chen’s scheme, they split the whole background picture into background coding tree unit blocks to achieve smooth bitrate output. However, these methods are designed for and thus only suitable for videos captured by stationary cameras thus stable background. Since the assumption of stationary camera results in modeling pixels of training pictures at the same temporal position, this may cause a problem of background quality which would decrease the coding prediction performance, as will be shown in the performance comparisons in Section VI. Besides, the background reference picture coding strategy should be adjusted in moving cameras circumstance. This is because the reconstructed selected background CTUs update statically the collocated location in the long-term reference picture, which will affect the quality of background reference picture.

### III. THEORETICAL ANALYSIS AND ARCHITECTURE OF THE PROPOSED SCHEME

This section will introduce the theoretical analysis and architecture of the proposed scheme. The rate-distortion (R-D) performance analysis is conducted in Section III-A, and the overall architecture of our proposed scheme is introduced in Section III-B.

#### A. R-D Performance Analysis

According to the rate-distortion optimization (RDO) theory [41], the R-D problem can in general be formulated as:

$$\min J = D + \lambda R \text{ s.t. } R \leq R_T \quad (1)$$

where  $J$  denotes R-D cost,  $D$  denotes distortion,  $R$  denotes rate,  $R_T$  denotes target rate,  $\lambda$  denotes Lagrange multiplier.

In the scene video, an input video frame can be separated into two parts, which can be expressed by

$$I = B + F \quad (2)$$

where  $I$  presents the input video frame,  $B$  denotes the background part,  $F$  denotes the foreground part.

Thus, the Lagrangian cost function can be rewritten as

$$\begin{aligned} \min J &= (D_B + D_F) + \lambda(R_B + R_F) \\ &= \underbrace{(D_B + \lambda R_B)}_{RDCost_{BG}} + \underbrace{(D_F + \lambda R_F)}_{RDCost_{FG}} \text{ s.t. } (R_B + R_F) \leq R_T \end{aligned} \quad (3)$$

where  $D_B$  and  $R_B$  denote the distortion and the bitrate of the background area, while  $D_F$  and  $R_F$  denote the distortion and the bitrate of the foreground area rate. In (3), the item in the first bracket of the second line represents the R-D cost of the background area, while the second represent the R-D cost of the foreground area.

In this paper, we aim to improve the prediction performance of background regions since there usually exists more background regions in the scene video. It means that the content of the first bracket is the main optimal goal. Therefore, the R-D cost



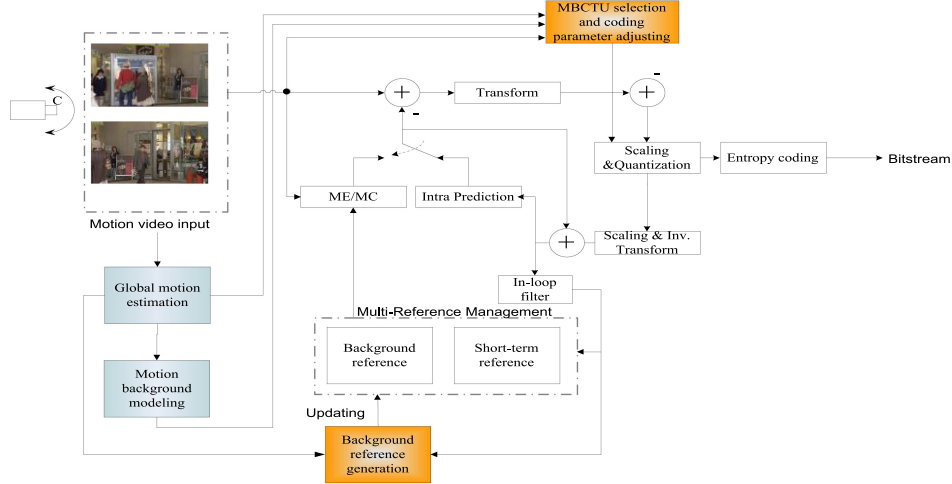


Fig. 1. Framework of our proposed surveillance video coding scheme.

function can be described by

$$\min J_{BG} = D_B + \lambda R_B$$

$$= \sum_{i=1}^n (C_i - REF_i(QP^*) + \varepsilon) + \lambda R_B \quad (4)$$

where  $n$  is the total number of background CTUs,  $C_i$  denotes the pixel value of the  $i$ -th background CTU in the current original frame,  $REF_i(QP^*)$  denotes the corresponding prediction CTU pixel value of the reference frame with  $QP^*$  derived in Section V-A,  $\varepsilon$  denotes the prediction error of the background pixel.

From the R-D cost function in (4), it should be noticed that when the  $REF_i(QP^*)$  approximates  $C_i$ ,  $J$  will be optimal. For this reason, building the clear and high-quality background reference will obtain more background reference pixels for improving the prediction efficiency of background regions.

### B. Overall Architecture of the Proposed Scheme

In order to improve background region prediction performance, our proposed scheme incorporates motion background modeling for building background modeling frame and designs a motion background reference picture coding strategy for moving cameras circumstance.

Fig. 1 shows the overall architecture of our proposed background modeling and referencing (BMR) method for moving camera-captured surveillance video coding scheme. Our proposed scheme includes two parts as highlighted in Fig. 1.

The first one is the background modeling for the clearer background modeling frame, as shown in the blue part of Fig. 1. In Fig. 1, two modules, including global motion estimation and motion background modeling, are employed to build the background modeling frame. The ‘global motion estimation’ module utilizes SVD-based global motion estimation model to obtain the global moving camera motion. The ‘motion background modeling’ module builds the background modeling frame in the moving camera circumstance. More details will be elaborated in Section IV.

The second one includes two modules, ‘MBCTU selection and coding parameters adjusting’ module and ‘background reference generation’ module, as shown in the yellow part of Fig. 1. In ‘MBCTU selection and coding parameters adjusting’ module, every encoding CTU would be compared with the corresponding block in the global compensation location of the background modeling frame. In the quantization procedure,  $QP^*$  values of selected MBCTUs are adjusted to generate high-quality CTUs with smaller QP. In ‘background reference generation’ module, the clear and high-quality MBCTUs will be used to update the initial background reference picture as a long-term reference by the global-motion-compensation updating strategy. Please refer to Section V for details.

## IV. PROPOSED MOTION BACKGROUND MODELING

In this section, the global motion estimation model firstly is established by LR-SVD decomposition and adaptively clustering the background motion vectors of character blocks in the safe boundary. Then based on the global motion estimation model, the motion background modeling is proposed by using global motion compensation.

### A. SVD-Based Global Motion Estimation Model

Inspired by the advanced motion vector prediction (AMVP) in HEVC and the safe boundary concept [42], [43], we propose a global motion estimation model based on character blocks in a bounding box.

For illustrating our global motion estimation modeling scheme, we take account of a representative indoor scene video, *BasketballPass* in HEVC, as the analytical example. Fig. 2(a) shows the positions of nine character blocks in a video frame. The gap is usually the width of  $8 \times 4$  pixels between four boundaries in practical film making [43]. Fig. 2(b) and (c) show the foreground segment and the optical flow field, respectively. Fig. 2(b) illustrates that there are more background regions in the scene video. Fig. 2(c) verifies motion vectors of background are consistent strongly in the scene video.

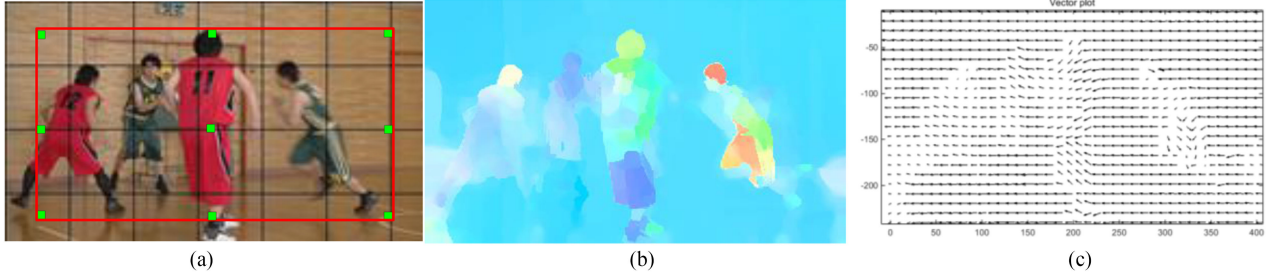


Fig. 2. Motion characteristic analysis for scene video. (a) Character blocks selection in the 81st frame of *BasketballPass* sequence; (b) foreground and background segment; (c) optical flow field.

For mathematical analysis, we define  $\mathbf{L}_k = \{\mathbf{L}_{k,i}, i = 1, 2, \dots, N\}$  which represents  $N$  character blocks in the  $k$ -th frame. In our modeling method, we use  $N = 9$  [43]. The positions of the  $N$  character blocks are processed in a raster scan order.  $\mathbf{L}_{k,1}$  represents the upper left corner location, and  $\mathbf{L}_{k,9}$  represents the bottom right corner location in the  $k$ -th frame. We define  $\mathbf{mv}_{k,i}$  as the motion vector of character block  $\mathbf{L}_{k,i}$  at the  $k$ -th frame,  $i = 1, 2, \dots, N$ .

$\mathbf{mv}_{k,i}$  need be scaled by its corresponding temporal reference distance because of multi-reference prediction strategy in HEVC. Thus, we define  $\mathbf{mv}_{k,i}^*$  as the scaled  $\mathbf{mv}_{k,i}$ , which can be expressed as

$$\mathbf{mv}_{k,i}^* = (mv_{k,i}^H/d, mv_{k,i}^V/d) \quad (5)$$

where  $d$  is the temporal reference distance between the reference frame and the current frame,  $mv_{k,i}^H$  represents the horizontal vector of  $\mathbf{mv}_{k,i}$ , and  $mv_{k,i}^V$  represents the vertical vector of  $\mathbf{mv}_{k,i}$ .

Since background motion vectors exhibit strong spatio-temporal low rank characteristic as shown in Fig. 2(c), we adopt a  $T$ -sized temporal sliding window to obtain a matrix  $\mathbf{G}$  of motion vectors for SVD analysis. The SVD method is used to estimate the background motion component from the character blocks in this paper. The insight underlying the use of SVD is that it is an efficient and unique decomposition method, which can partition the correlated data into a set of independent components and order the dimensions along which component exhibits the most variation.

The dimension of the matrix  $\mathbf{G}$  is  $2 \times T \times i$ , which can be given by

$$\mathbf{G} = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_i] \\ = \begin{bmatrix} mv_{k-T+1,1}^{*H} & \cdots & mv_{k-T+1,i}^{*H} \\ mv_{k-T+1,1}^{*V} & \cdots & mv_{k-T+1,i}^{*V} \\ \vdots & \ddots & \vdots \\ mv_{k,1}^{*H} & \cdots & mv_{k,i}^{*H} \\ mv_{k,1}^{*V} & \cdots & mv_{k,i}^{*V} \end{bmatrix} \quad i \in [1, N] \quad (6)$$

The motion vectors of character blocks in the video can be measured by the matrix  $\mathbf{G}$ . When the rank of measuring the matrix  $\mathbf{G}$  is lower, the variance of motion vectors is smaller which is consistent with background motion vectors in Fig. 2(c).

Thus, SVD is used to measure the matrix  $\mathbf{G}$  for estimating its background component by restricting  $\Sigma$  to the first singular value and setting other singular values to zero. Here,  $\mathbf{U}$  and  $\mathbf{V}$  are orthogonal matrices. The reconstruction matrix can be expressed as

$$\hat{\mathbf{G}} = \mathbf{U}\Sigma\mathbf{V}^T \quad (7)$$

Since the inconsistency of motion in foreground character block is higher compared with that in background character block, as shown in Fig. 2(c), through eliminating the foreground motion component, the inconsistency can be represented by the residual motion component, which is defined as  $\mathbf{G} - \hat{\mathbf{G}}$ . It is easy to classify the motion vectors of character blocks by an adaptive threshold. We compute the squared value of each column of the residual motion component and find out those larger than a threshold. In this paper, the mean squared error is simply computed as the threshold, which can be expressed by

$$Th_k = \frac{1}{N} \sum_{i=1}^N \|\mathbf{X}_i - \hat{\mathbf{X}}_i\|^2 \quad (8)$$

where  $N$  is the number of the character block,  $\mathbf{X}_i$  is the column vector of the matrix  $\mathbf{G}$ ,  $\hat{\mathbf{X}}_i$  is the column vector of the matrix  $\hat{\mathbf{G}}$ .

Based on the classified background motion vectors by using the SVD technique, the horizontal and the vertical vector of  $\mathbf{MV}_{k,global}$  can be defined by

$$MV_{k,global}^H = \frac{1}{N} \sum_{mv^* \in \Omega_{BG}} P_{k,i} mv_{k,i}^{*H} \quad (9)$$

$$MV_{k,global}^V = \frac{1}{N} \sum_{mv^* \in \Omega_{BG}} P_{k,i} mv_{k,i}^{*V} \quad (10)$$

where  $P_{k,i}$  denotes the probability of every motion vector in the background set  $\Omega_{BG}$ . Usually,  $P_{k,i}$  value can use the mean value by the size of  $\Omega_{BG}$ .

Therefore, the global motion vector can be expressed by

$$\mathbf{MV}_{k,global} = (MV_{k,global}^H, MV_{k,global}^V) \quad (11)$$

So far, our proposed SVD-based global estimation modeling algorithm can be summarized as in ‘Algorithm 1’, in term of pseudo code.

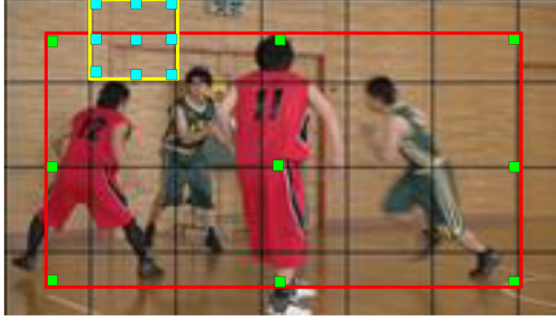


Fig. 3. CTU-level and frame-level character blocks selection for potential MBCTUs selection.

---

**Algorithm 1:** Global Motion Estimation.

---

- 1: Input:  $\mathbf{mv}_{k,i}, T$
  - 2: Initialization:  $i = 1, \Omega_{BG} = \emptyset, \Omega_{FG} = \emptyset$
  - 3: Calculate  $Th_k$  in (8)
  - 4: If  $i \leq N$  then
  - 5: Loop:
  - 6: Scaled normalized the input  $\mathbf{mv}_{k,i}$  by Calculate  $\mathbf{mv}_{k,i}^*$
  - 7: If  $\|\mathbf{mv}_{k,i}^*\| < Th_k$
  - 8:  $\Omega_{BG} = \mathbf{mv}_{k,i}^* \cup \Omega_{BG}$
  - 9: Else
  - 10:  $\Omega_{FG} = \mathbf{mv}_{k,i}^* \cup \Omega_{FG}$
  - 11: Endif
  - 12: End loop
  - 13: Endif
  - 14: Calculate  $\mathbf{MV}_{k,global}$  in (11)
  - 15: Output:  $\mathbf{MV}_{k,global}$
- 

### B. Background Modeling Using Global Motion Compensation

Conventional background modeling approaches for scene coding, such as Zhang's [12] and Chen's [13] works, are unsuitable in the moving scene case, since the assumption of stationary camera results in modeling pixels of training pictures at the same temporal position. Moreover, conventional background modeling methods usually bring heavy complexity with high accuracy float computation (e.g., GMM model).

Taking advantage of global motion information by combining frame-level global motion and CTU-level local motion, our algorithm can firstly segment motion regions in moving camera cases to reduce the increasing impact from foreground pixels in background modeling. Then different background modeling strategies for different motion regions will be used to build the background modeling frame by utilizing global motion information.

To segment motion regions in moving camera cases, a combining frame-level global motion and CTU-level local motion region segmentation method is introduced. Fig. 3 shows our background CTU selection method with combining frame-level global motion and CTU-level local motion. According to the motion vectors correlation of the character blocks in CTU-level

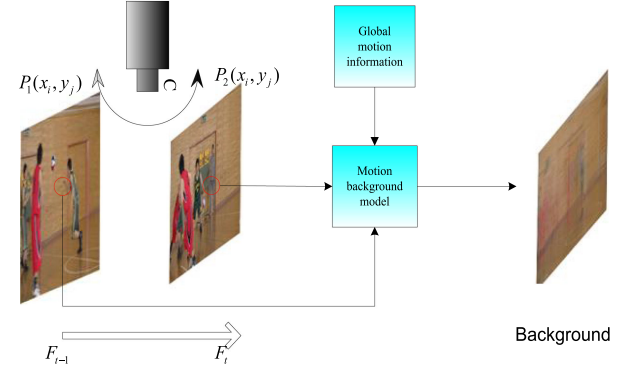


Fig. 4. The flowchart of our proposed motion background model using global motion compensation.

and frame-level as shown in Fig. 3, an adaptive threshold method is proposed to segment the background frame into two classes of motion regions, including background motion (BGM) regions and foreground motion (FGM) regions, inspired from our prior work in the local motion estimation region classification [44]. When the CTU is MBCTU, every motion vector of the CTU-level character block will be the same with the global motion vector. So, the adaptive threshold can be given by

$$\delta_k = \min_{\mathbf{MV}_k^* \in \Omega_{BG}} \|\mathbf{MV}_k^* - \mathbf{MV}_k'\|_F \quad \text{s.t.} \quad \|\Omega_{BG}\| \leq N \quad (12)$$

where  $\|\mathbf{MV}_k^* - \mathbf{MV}_k'\|_F$  is the Frobenius norm for the difference of motion vectors between frame-level global motion vectors and CTU-level local motion vectors.  $\|\Omega_{BG}\|$  denotes the number of background motion vectors obtained by our proposed SVD-based global motion estimation model.  $\mathbf{MV}_k^*$  is the motion vector set of the frame-level background character block located at the  $k$ -th frame.  $\mathbf{MV}_k'$  is the motion vector set of the CTU-level character block location corresponding the frame-level background character block located at the  $k$ -th frame. We use  $th1$  as the max similarity distance threshold. When  $\delta_k \leq th1$ , the CTU will be selected as the potential background CTU. In the modeling process, the original pixel in the potential background CTU will update the background pixel in the global motion compensation location. Since the velocity of camera is usually 6–8 pixels per degree in most surveillance circumstances, we set  $th1 = 16$  in our experiment.

Note that regions segmentation can coarsely identify potential background pixels which is different from previous background modeling methods in video coding. These regions can reduce the “blurry effect” brought from foreground pixels in the background modeling frame. Getting rid of the foreground pixels will make the background better and clearer in the background modeling process. Furthermore, regions segmentation would select different background modeling strategies for corresponding regions in the moving cameras circumstance.

Our background modeling process by utilizing global motion information and above the region segmentation will build the background modeling frame. Fig. 4 shows our proposed

modeling process. In Fig. 4, every pixel of the input video frame should find its compensation position in the background modeling frame by global motion information. Then utilizing the characteristics in different motion regions, we design different background modeling strategies for every input pixel. The mathematical representation of our motion background model is expressed by

$$B_k(x, y) = f(I_k, \mathbf{MV}_{k, global}) \quad (13)$$

where  $\mathbf{MV}_{k, global} = (mv_{k, global}^H, mv_{k, global}^V)$  is the global motion vector from our proposed SVD-based global motion estimation model, the function of  $f$  represents the motion background modeling function.

Since the running average method is lower-complexity compared with GMM method, this method is suitable for real-time surveillance video coding. Moreover, since the conventional running average method builds the modeling pixels of training pictures at the same temporal location, it leads to the blurry effect in the background modeling frame. Therefore, our method introduces the global motion vector in the background modeling. Making use of the global motion vector, our method builds the modeling pixels of training pictures at the global motion compensation temporal location. So far, combining the running average method based on the static background modeling [8], (13) can be rewritten by

$$B_k(x, y) = \begin{cases} I_0 & k = 0 \\ \alpha \times B_{k-1}(x, y) + (1 - \alpha) \times \bar{I}_s(x, y) & k > 0 \end{cases} \quad (14)$$

where  $\alpha = W_s / (W_s + L_s^2)$  represents a updating factor calculated by the weight value  $W_s$  and the length  $L_s$  of the background training segment  $S$ ,  $I_0$  denotes the first original frame as the initial background,  $B_{k-1}(x, y)$  denotes the previous background frame in the location  $(x, y)$ ,  $\bar{I}_s(x, y)$  denotes the mean pixel intensity of the  $s$ -th training segment in the location  $(x, y)$ , which can be expressed in (15). In (15), shown at the bottom of the page,  $d(x, y) = |I_k(x - \sum_{m=0}^k mv_{m, global}^H, y - \sum_{m=0}^k mv_{m, global}^V) - \bar{I}_s(x, y)|$  denotes the difference between the pixel intensity in the global compensation position and the mean pixel intensity in the current position, BGM denotes the background global motion region which is corresponding with the collocated MBCTU in the background modeling frame, FGM denotes the foreground global motion region which is inconstant with the global motion. In our experiment, we set  $th2 = 14$  which is the same with [8]. Our background modeling algorithm is detailed described in 'Algorithm 2'.

## V. MOTION BACKGROUND-BASED CODING STRATEGY

The main idea of surveillance video coding is that the codec can obtain the clear and high-quality background reference pic-

### Algorithm 2: Motion Background Modeling.

---

```

1: Input:  $\mathbf{P} = \{I_i = f_i(x, y) | i = 0, \dots, k\}$ , where
    $f_i(x, y)$  is every input original frame.
2: Initialization:  $s = 0, k = 0, L_s = 0, W_s = 0, \alpha = 0$ 
3: For each CTU in  $\mathbf{P}$ 
4:   If the input frame is the first frame
5:      $B_k = I_0$ , and continue
6:   Else
7:     Estimate Minkowski distance  $\delta_k$  using (12)
8:   Endif
9:   If  $\delta_k \leq thr1$  then
10:    Loop:
11:      Choose each pixel of a collocated background CTU
        in the background modeling frame
12:      If  $d(x, y) \leq thr2$  then
13:         $\bar{I}_s(x, y) = (\bar{I}_{s-1}(x, y) \times L_s + L_s/2) / (L_s + 1)$ 
14:         $L_s = L_s + 1$ , update  $\alpha$ 
15:      Else
16:         $W_s = W_s + L_s^2, s = s + 1, L_s = 0$ , update  $\alpha$ 
17:         $\bar{I}_s(x, y) = I_k(x - \sum_{m=0}^k mv_{m, global}^H, y - \sum_{m=0}^k mv_{m, global}^V)$ 
18:      Endif
19:      Calculate  $B_k(x, y)$  using (14)
20:    End loop
21:  Else
22:     $B_k(x, y) = B_{k-1}(x, y)$ 
23:  Endif
24: End
25: Output:  $B_k$ 

```

---

ture. In this paper, Section IV describe our proposed motion background modeling scheme to obtain a better background modeling frame in the moving cameras circumstance. Thus, this section will introduce our proposed coding strategy for generating clear and high-quality background reference picture in the moving scene.

### A. QP Parameter Estimation and Optimization for MBCTUs

In order to obtain high-quality background MBCTUs, the quantization approach need be adjusted in the coding process. Since HEVC adopts RDO theory to select the optimal coding mode in video coding process, all coding parameters adjusting should consider the relationship between bit-rate and distortion. In the R-lambda model, where  $\lambda$  is a key factor which can adjust coding parameters to achieve the best R-D cost performance. Hence, our QP adjusting method adjusts  $\lambda$  to calculate the new QP for MBCTUs.

$$\bar{I}_s(x, y) = \begin{cases} (\bar{I}_{s-1}(x, y) \times L_s + L_s/2) / (L_s + 1) & d(x, y) \leq th2 \& \& BGM \\ I_k(x - \sum_{m=0}^k mv_{m, global}^H, y - \sum_{m=0}^k mv_{m, global}^V) & d(x, y) > th2 \& \& BGM \\ B_{k-1}(x, y) & FGM \end{cases} \quad (15)$$



In Zhang's [12] proposed coding scheme, background QP values are calculated from the hierarchical prediction structure QP setting. This quantization method usually exists the precision finite problem for empirically selecting constant QP values. In order to solve the traffic burst problem, Chen *et al.* [13] proposed the CTU-level background quantization approach, in which the optimal QP and lambda values are deduced for background CTUs in a group of picture (GoP) to constrain the bitstream. However, this method is proposed based on the assumption that the QP of background CTUs empirically adopts constant QP, i.e.,  $QP_{intra} - 10$ , where  $QP_{intra}$  is the QP of the IDR frame in the GoP. Note that the optimal QP can be further induced by (4), which can be represented as

$$QP_{MBCTU} = \{QP^* | \arg \min J_{BG} = D_B + \lambda R_B\}$$

$$= \left\{ QP^* | \arg \min \left( \sum_{i=1}^n (C_i - REF_i(QP^*) + \varepsilon) + \lambda R_B \right) \right\} \quad (16)$$

Hence, finding the optimal MBCTU QP for inter prediction is worth for improving coding performance in the R-lambda model.

According to the R-lambda model [45], the liner relationship between the QP and lambda value can be expressed as

$$QP_{opt} = \alpha \times \ln(\lambda) + \beta \quad (17)$$

where  $\lambda$  is Lagrange multiplier,  $\alpha$  and  $\beta$  are set to 4.2005 and 13.7122, respectively. Note that newly developed CTU-level rate control algorithms [46], [47] can also be used.

From the R-lambda model, the lambda value of each CTU can be given by

$$\lambda_i = - \frac{\partial D_i}{\partial R_i} \quad (18)$$

where  $D_i$  is the distortion of each CTU,  $R_i$  is bit-rate of each CTU.

Moreover the bit-rate of each CTU is nearly independent [13], the derivative can be represented as

$$\frac{\partial (\sum_{i=1}^n (C_i - REF_i(QP^*) + \varepsilon) + \lambda R_B)}{\partial R_B} = 0 \quad (19)$$

Combine (18) and (19), the lambda value of MBCTU can be given by

$$n \times \lambda_i = \lambda \quad (20)$$

where  $\lambda$  is a global Lagrange multiplier,  $n$  is the number of MBCTU,  $\lambda_i$  is lambda value of each MBCTU.

With (17) and (20), the optimal QP for each MBCTU can be represented as

$$QP_{MBCTU} = \alpha \times \ln(\lambda/n) + \beta \quad (21)$$

where  $\lambda$  is the common lambda value of the current sequence.

### B. Background Reference Picture Generation and Updating

In order to obtain the clear and high-quality background reference picture, the background reference picture will be generated and updated in the encoding process. The background reference

picture generation and updating strategy in the HEVC encoding process are performed as follows: (i) The first I frame or the first instantaneous decoder refresh (IDR) frame will be the initial background reference picture as the long-term reference picture in the reference list buffer; (ii) The MBCTUs of the subsequent input frame would be selected by anchoring the input video frame on the background modeling frame; (iii) The selected MBCTUs would be coded with the optimal QP calculated by (21) in the quantization process; (iv) The reconstructed MBCTU will update the previous CTU in the global compensation location of the background reference picture. Similarly, the decoder side would generate and update also the background reference picture using the same generation and updating strategy with the encoder side.

By introducing the modified coding strategy in HEVC encoding process, the clear and high-quality background reference picture will be obtain in the moving cameras circumstance. Since the high-quality background reference picture is as the long-term reference picture, background regions will achieve better prediction performance by finding the optimal match region in the background reference picture and thus improve the overall video coding performance as will be verified in the next section.

## VI. EXPERIMENTAL RESULTS AND DISCUSSIONS

This section will evaluate performance of our proposed BMR scheme, with comparisons to HEVC and two state-of-the-art scene video coding algorithms. The experimental setting is introduced in Section VI-A. The experimental results and analysis are given in Section VI-B.

### A. Experimental Settings

In order to fairly evaluate the performance of the proposed BMR scheme, our scheme is incorporated into the HEVC test model HM 12.0 reference software for Zhang's scheme [12] and Chen's scheme [13] were both implemented in HM 12.0 [48]. Since Low-delay B main Profile is adopted in Zhang's background modeling hierarchical prediction structure optimization (BHO) scheme specified in [12] and Chen's block-composed background reference (BCBR) scheme specified in [13], our experimental test condition is set the same encoding configuration for fair comparison. Therefore the BMR scheme will be compared with the original HM 12.0, Zhang's BHO scheme of HM 12.0 and the recent Chen's BCBR scheme of HM 12.0 under the same encoding configuration. Since four references are used in [12], [13], including three short-term references and one background long-term reference, we set the same reference configuration to be fair. Both objective and subjective coding performance are evaluated in term of Bjøntegard Distortion-Rate (BD-rate) and the Multiple Scale-Structural Similarly Index (Ms-SSIM) measure, respectively.

Since our proposed scheme aims at improving the coding performance of moving cameras-captured surveillance videos, eight video sequences captured by moving camera are selected from three typical video datasets, namely HEVC Dataset, SJTU UHD Dataset and Surveillance Videos Dataset. The camera motion characteristics of these sequences are represented in the following and also summarized in Table I.



TABLE I  
THE REPRESENTATIVE TEST SEQUENCE FROM MOVING CAMERAS

Video Source	Video Resolution	Video Name	Frame Rate	Camera Motion Characteristic
HEVC	832×480	BQMall	60	Left panning
	416×240	BasketballPass	50	Left-right panning
	1920×1080	BasketballDrive	50	Left-right panning
	1920×1080	BQTerrace	60	Top-right panning
SJTU UHD	3840×2160	Library	30	Left panning
	3840×2160	TallBuilding	30	Bottom-right panning
Surveillance Video	720×576	Scene1	25	Periodic panning
	1920×1080	Scene2	25	Periodic panning

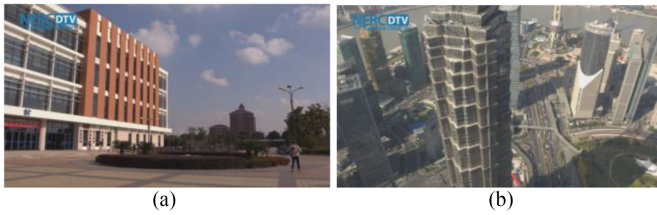


Fig. 5. SJTU UHD videos captured by moving cameras [49]. (a) Library; (b) TallBuilding.



Fig. 6. Surveillance videos captured by PTZ monitor cameras. (a) Scene1; (b) Scene2.

**HEVC Dataset:** Four HEVC test sequences captured by moving cameras are selected, including three indoor scene videos (*BQMall*, *BasketballPass*, *BasketballDrive*) and an outdoor scene video (*BQTerrace*). The camera motion characteristic includes left panning (*BQMall*), left-right panning (*BasketballPass* and *BasketballDrive*) as well as top-right panning (*BQTerrace*).

**SJTU UHD Dataset:** Due to the increasing demand for and emerging wide applications of high definition video content services, two  $3840 \times 2160$  (4 K) Ultra High Definition (UHD) videos are selected from SJTU UHD Dataset [49] to verify our proposed algorithm in the moving scene. The camera motion characteristic of *Library* is right panning. And the camera motion characteristic of *TallBuilding* is bottom-right panning.

**Surveillance Videos Dataset:** Considering scene video coding is usually employed in the surveillance application, we select two surveillance videos with  $720 \times 576$  (720 p) and  $1920 \times 1080$  (1080 p) which are mainly video resolutions in the current surveillance video application. Moreover, these two surveillance videos are captured by practical PTZ cameras with the periodic moving camera characteristic.

## B. Performance Evaluation and Comparisons

**1) Motion Background Reference Generation:** The Generated motion background references of the Sence2 sequence by three methods are shown in Figs. 7(b), 8(a) and 8(b). As mentioned above, the Sence2 sequence is a surveillance video captured by a PTZ monitor camera. The PTZ camera periodically monitors the road scene, which is shown in Fig. 7(a). Thus, the background is unchanged periodically which can be modeled

effectively by our approach. It can be seen in Fig. 7(b) that our method generates a clearer background reference in the motion background circumstance. In other word, our method will obtain more gain in background prediction processing to improve encoding efficiency.

In order to demonstrate the performance of the proposed motion background modeling algorithm, we extract the background reference of Zhang's and Chen's method in the same 60-th frame with our method, respectively. In Fig. 8(a), there is the whole blurry effect in the background reference picture of Zhang's method. This is due to the fact that Zhang's method adopts the running average background modeling in the same pixel position of every frame. Therefore, this method can hardly calculate the average background pixel value for the same pixel in the temporal motion direction. The result of Chen's method is shown in Fig. 8(b). It is noticed that Chen's method generates a clearer background reference than Zhang's method. However, there are two problems leading to obtaining poor background reference. On the one hand, some foreground objects (e.g., the back car in the red circle) are unable to be eliminated effectively and quickly which will decrease the prediction efficiency for the exposed background regions. This is because this method uses the GMM method as the anchor background model which would establish the background model at the same temporal position. On the other hand, the background CTU updates in the error position of the background reference (e.g., the gray block in the red circle). This is due to the fact that this method is unable to find the corresponding motion compensation position in the motion background case. Since the proposed global motion



Fig. 7. The generated background references of the Sence2 sequence at QP 27 with proposed method.(a)Original frame; (b) our method.



Fig. 8. The generated background references of the Sence2 sequence at QP 27 with other methods. (a) The generated background reference using the first same 60 frames with method in Zhang's; (b) the generated background reference using the first same 60 frames with method in Chen's.

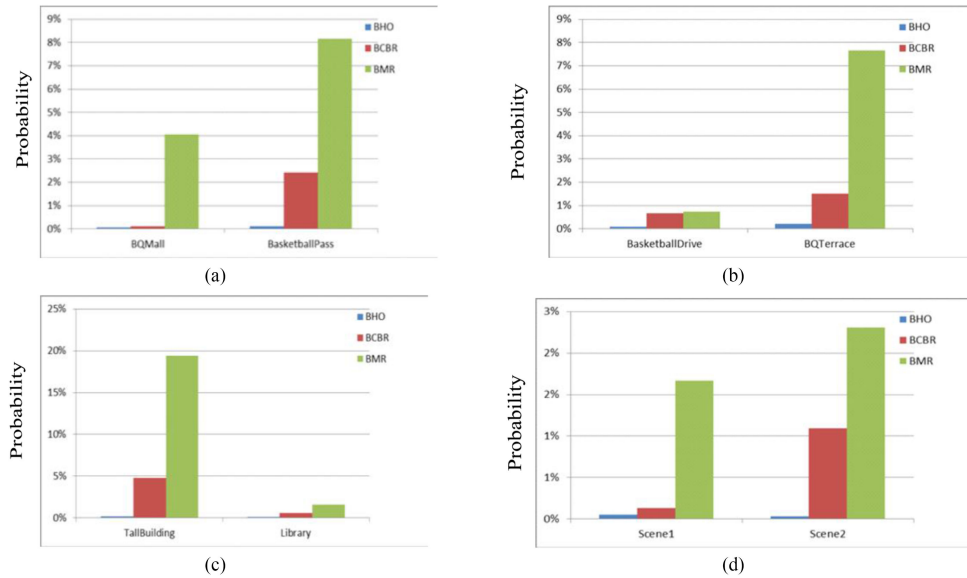


Fig. 9. The background referencing probability distribution. (a) The probability of *BQMall* and *BasketballPass*; (b) the probability of *BasketballDrive* and *BQTerrace*; (c) the probability of *Tallbuilding* and *Library*; (d) the probability of *Scene1* and *Scene2*.

estimation method is considered in the background modeling, our algorithm yields a better background reference in the motion circumstance.

2) *Background Referencing Probability Distribution*: To demonstrate further the problem of the prediction performance

restrained by the quality of background reference, background reference probability distribution will be analyzed in Fig. 9. Note that our proposed algorithm achieves high probability for background-referencing in moving cameras circumstance, compared to the BHO method and the BCBR method. This is

TABLE II  
OVERALL EXPERIMENTAL RESULTS ANCHOR ON HM12.0

Class	Sequences	BD Rate				Coding time
		Y psnr	U psnr	V psnr	Y Ms-SSIM	$\Delta$ time
HEVC	BQMall	-6.1%	-8.7%	-9.2%	0.19%	5.97%
	BasketballPass	-2.0%	-3.8%	-1.0%	0.39%	8.67%
	BasketballDrive	-0.1%	-1.2%	-0.4%	0.01%	2.89%
	BQTerrace	-7.0%	-41.3%	-51.7%	0.15%	0.12%
SJTU UHD	Sjtu TallBuilding	-26.6%	-59.1%	-57.6%	0.47%	7.45%
	Sjtu Library	-1.9%	-4.3%	-3.4%	0.01%	6.77%
Surveillance Video	Scene1	-3.4%	-8.7%	-13.6%	0.02%	2.67%
	Scene2	-6.7%	-16.6%	-19.1%	0.16%	4.87%
Overall		-6.7%	-18.0%	-19.5%	0.17%	4.91%

Y psnr: Y component peak signal to noise ratio, U psnr: U component peak signal to noise ratio, V psnr: V component peak signal to noise ratio

TABLE III  
OVERALL EXPERIMENTAL RESULTS ANCHOR ON THE BHO METHOD  
(BD-RATE SAVING)

Class	Sequences	BD Rate			
		Y psnr	U psnr	V psnr	Y Ms-SSIM
HEVC	BQMall	-32.0%	-33.7%	-33.0%	0.28%
	BasketballPass	-13.6%	-15.1%	-12.3%	0.41%
	Basketballdrive	-14.4%	-16.6%	-15.2%	0.21%
	BQTerrace	-41.5%	-64.6%	-71.8%	0.22%
SJTU UHD	Sjtu Tallbuilding	-43.0%	-87.2%	-85.1%	0.48%
	Sjtu Library	-9.1%	-14.3%	-13.9%	0.03%
Surveillance Video	Scene1	-20.6%	-16.9%	-21.1%	0.39%
	Scene2	-24.8%	-29.0%	-29.9%	0.20%
Overall		-24.9%	-34.7%	-35.3%	0.3%

TABLE IV  
OVERALL EXPERIMENTAL RESULTS ANCHOR ON THE BCBP METHOD  
(BD-RATE SAVING)

Class	Sequences	BD Rate			
		Y psnr	U psnr	V psnr	Y Ms-SSIM
HEVC	BQMall	-5.9%	-8.8%	-8.7%	0.14%
	BasketballPass	-2.1%	-3.4%	-1.6%	0.22%
	Basketballdrive	-0.3%	-1.1%	-0.6%	-0.05%
	BQTerrace	-5.9%	-40.5%	-51.3%	0.14%
SJTU UHD	Sjtu Tallbuilding	-23.0%	-67.2%	-65.1%	0.42%
	Sjtu Library	-2.5%	-4.2%	-3.8%	-0.11%
Surveillance Video	Scene1	-0.9%	-4.8%	-9.2%	0.23%
	Scene2	-2.4%	-11.9%	-13.8%	0.16%
Overall		-5.4%	-17.7%	-19.3%	0.2%

due to the fact that our method proposes a motion background modeling and global motion compensation update strategy to generate the better background reference as the long-term reference. Thus, the high-quality background reference will be referenced with high probability to improve prediction performance. It should be noticed that *BQTerrace* and *Tallbuiding* sequences obtain higher probability for the long-term background reference. It is verified that the background reference picture generated by the proposed scheme is more effectively used for inter prediction, compared to other sequences.

3) *The Overall Bit-Saving Results:* Table II illustrates the overall results of BD-rate and coding time for the proposed BMR method anchored on HM 12.0. It can be noticed in Table II that the average BD-rate reduction is 6.7% for Y component, 18% for V component and 19.5% for U component, compared to the original HM 12.0 with similar Ms-SSIM values. It is because that the proposed scheme improves the motion background reference quality in the motion camera circumstance. It means that the background prediction performance will be increased

obviously, which also is verified in Fig. 9. The reconstructed quality is measured by Ms-SSIM values of Y component which can mostly represent the subjective quality. The similar Ms-SSIM value means the video viewer can rarely perceive visible distortion among reconstructed frames with two encoding approaches. For the encoding time, the proposed scheme increases 4.91%, compared to the original HM 12.0. It is because that the proposed scheme adopts a SVD-based global motion estimation method and motion background modeling which bring the extra computational complexity.

Especially, it is noticed that *BQTerrace* and *Tallbuiding* sequences obtains higher gain on R-D performance than other test sequences. This is due to the fact that these two test sequences with more regular texture regions will effectively improve the prediction performance by the clear and high-quality background reference in moving cameras cases.



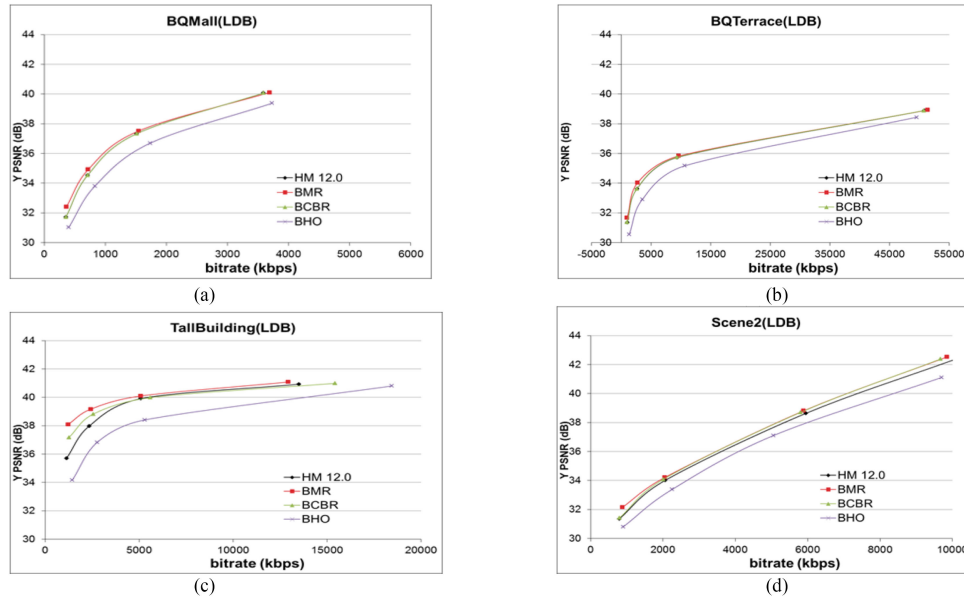


Fig. 10. R-D curves for some test sequences. (a) BQMall; (b) BQTerrace; (c) Tallbuilding; (d) Scene2.

4) *Comparisons With the State-of-the-Art Methods on R-D Performance:* Two state-of-art scene coding methods, called the BHO method [12] and the BCBR method [13], are selected as representative algorithms for comparison. The BHO method proposed a background modeling based hierarchical prediction structure optimization method for surveillance and conference videos. Since the BHO method considered the background modeling for static pictures, this leads to poor background prediction. The other scene coding scheme, the BCBR method, proposed a block-composed background reference generation algorithm to smooth the traffic with smaller QP background. However, this method still employs the conventional static background model, which is a single GMM model. As we known, the GMM model need more floor computation which is unsuitable for the real-time application and hardware codec design. The coding performance of the BHO method and the BCBR method vs. proposed method are shown on Y, U, V components in Tables III and IV, respectively.

Compared with the BHO method, the proposed method achieves a bit-savings of 24.9%, 34.7%, 35.3% on average for Y, U, and V components, respectively. Moreover, the proposed method outperforms also the BCBR method with the average bitrate reduction of 5.4%, 17.7%, 19.3% for Y, U, and V components, respectively. This is due to the fact that our method utilizes a global motion estimation method and motion-background modeling to establish the better background reference picture. Especially, exposed background regions [e.g., the back car in Fig. 8(a) and (b)] will achieve better prediction performance by background reference. Besides, our proposed method adopts optimal QP for MBCTU to predict the background regions effectively by the R-D theory and the R-lambda model. The R-D curves of some test sequences are shown in Fig. 10. From the result of R-D curves, the encoding performance of the proposed method obtains the higher gain at low bitrates. This is because that the high-quality reference with small QP

will provide more reference pixel values for the current coding picture.

## VII. CONCLUSION

In this paper, we have addressed the problems of high efficiency surveillance video coding in moving cameras circumstance and proposed a background modeling and referencing surveillance video coding scheme using a motion vector SVD-based global motion estimation model. First, this paper proposed a low-complexity motion background modeling algorithm for surveillance video coding using the running average based on global-motion-compensation method. To estimate the global motion vector, we proposed a global motion detection method based on character blocks. Second, a background referencing coding strategy is proposed to generate and update the clear and high-quality background reference in the moving cameras case. Extensive experimental results show that the proposed scheme implemented in HEVC can yields significant bit saving of up to 26.6% and on average 6.7% with similar subjective quality and negligible encoding complexity, compared to HM 12.0. Besides, the proposed scheme consistently outperforms two state-of-art surveillance video coding schemes with remarkable reduced bitrate overhead. In future work, the BMR method should investigate further the zoom-in and -out case and scene change to effectively use the background picture for more general surveillance video applications.

## ACKNOWLEDGMENT

The authors are grateful to Dr. Xianguo Zhang and Dr. Fangdong Chen for providing their video codecs for comparisons. They would also like to thank SJTU Media Lab from Shanghai Jiao Tong University for providing the UHD dataset. Thanks to the anonymous reviewers and the editor.



## REFERENCES

- [1] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Sep. 2012.
- [2] W. Gao, Y. Tian, T. Huang, S. Ma, and X. Zhang, "The IEEE 1857 standard: Empowering smart video surveillance systems," *IEEE Intell. Syst.*, vol. 29, no. 5, pp. 30–39, Sep./Oct. 2014.
- [3] J. Xiao, R. Hu, L. Liao, Y. Chen, and Z. Xiong, "Knowledge-based coding of objects for multisource surveillance video data," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1691–1706, Sep. 2016.
- [4] Z. F. Shao, J. J. Cai, and Z. Y. Wang, "Smart monitoring cameras driven intelligent processing to big surveillance video data," *IEEE Trans. Big Data*, vol. 4, no. 1, pp. 105–116, Mar. 2018.
- [5] D. Liu, D. Zhao, X. Ji, and W. Gao, "Dual frame motion compensation with optimal long-term reference frame selection and bit allocation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 3, pp. 325–339, Mar. 2010.
- [6] L. Zhao, Y. H. Tian, and T. J. Huang, "Background-foreground division based search for motion estimation in surveillance video coding," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2014, pp. 1–6.
- [7] Y. Liu and D. A. Pados, "Compressed-sensed-domain L1-PCA video surveillance," *IEEE Trans. Multimedia*, vol. 18, no. 3, pp. 351–363, Mar. 2016.
- [8] X. Zhang, T. Huang, Y. Tian, and W. Gao, "Background-modeling-based adaptive prediction for surveillance video coding," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 769–784, Feb. 2014.
- [9] T. Wiegand, X. Zhang, and B. Girod, "Long-term memory motion compensated prediction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 1, pp. 70–84, Feb. 1999.
- [10] M. Tiwari and P. C. Cosman, "Selection of long-term reference frames in dual-frame video coding using simulated annealing," *IEEE Signal Process. Lett.*, vol. 15, pp. 249–252, 2008.
- [11] M. Paul, W. Lin, C. Lau, and B. Lee, "A long-term reference frame for hierarchical b-picture-based video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 10, pp. 1729–1742, Oct. 2014.
- [12] X. Zhang, Y. Tian, T. Huang, S. Dong, and W. Gao, "Optimizing the hierarchical prediction and coding in HEVC for surveillance and conference videos with background modeling," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4511–4526, Oct. 2014.
- [13] F. D. Chen, H. Q. Li, L. D. Liu, and F. Wu, "Block-composed background reference for high efficiency video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 12, pp. 2639–2651, Dec. 2017.
- [14] K. Xue, G. Ogunmakin, Y. Liu, P. A. Vela, and Y. Wang, "PTZ camera-based adaptive panoramic and multi-layered background model," in *Proc. 18th IEEE Int. Conf. Image Process.*, Sep. 2011, pp. 2949–2952.
- [15] A. Sobral and A. Vacavant, "A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos," *Comput. Vision Image Understand.*, vol. 122, pp. 4–21, May 2014.
- [16] T. Bouwmans, "Traditional and recent approaches in background modeling for foreground detection: An overview," *Comput. Sci. Rev.*, vols. 11–12, pp. 31–66, May 2014.
- [17] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 780–785, Jul. 1997.
- [18] C. Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 747–757, Aug. 2000.
- [19] O. Barnich and M. V. Droogenbroeck, "ViBe: A universal background subtraction algorithm for video sequences," *IEEE Trans. Image Process.*, vol. 21, no. 5, pp. 2582–2591, Dec. 2011.
- [20] P. St-Charles, G. Bilodeau, and R. Bergevin, "SuBSENSE: A universal change detection method with local adaptive sensitivity," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 359–373, Dec. 2014.
- [21] P. St-Charles, G. Bilodeau, and R. Bergevin, "Universal background subtraction using word consensus models," *IEEE Trans. Image Process.*, vol. 25, no. 10, pp. 4768–4781, Oct. 2016.
- [22] S. Javed, A. Mahmood, T. Bouwmans, and S. Jung, "Spatiotemporal low-rank modeling for complex scene background initialization," *IEEE Trans. Circuits Syst. Video Technol.*, to be published, doi: 10.1109/TCSVT.2016.2632302.
- [23] T. Bouwmans, A. Sobral, S. Javed, S. Jung, and E. Zahzah, "Decomposition into low-rank plus additive matrices for background/foreground separation: A review for a comparative evaluation with a large-scale dataset," *Comput. Sci. Rev.*, vol. 23, pp. 1–71, Feb. 2017.
- [24] Y. Sheikh and M. Shah, "Bayesian modeling of dynamic scenes for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 11, pp. 1778–1792, Nov. 2005.
- [25] G. Zhang *et al.*, "Moving object extraction with a hand-held camera," in *Proc. IEEE 11th Int. Conf. Comput. Vision*, Oct. 2007, pp. 1–8.
- [26] Y. M. Chen and I. V. Bajic, "A joint approach to global motion estimation and motion segmentation from a coarsely sampled motion vector field," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 9, pp. 1316–1328, Apr. 2011.
- [27] D. Zamaliev and A. Yilmaz, "Background subtraction for the moving camera: A geometric approach," *Comput. Vision Image Understand.*, vol. 127, pp. 73–85, Oct. 2014.
- [28] H. Bhakar, K. Dwivedi, D. Dogra, M. Al-Mualla, and L. Mihaylova, "Autonomous detection and tracking under illumination changes, occlusions and moving camera," *Signal Process.*, vol. 117, pp. 343–354, Jun. 2015.
- [29] A. Ferone and L. Maddalena, "Neural background subtraction for pan-tilt-zoom cameras," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 44, no. 5, pp. 571–579, Sep. 2013.
- [30] G. Chau and P. Rodriguez, "Panning and jitter invariant incremental principal component pursuit for video background modeling," in *Proc. IEEE Int. Conf. Comput. Vision Workshops*, Oct. 2017, pp. 1844–1852.
- [31] C. Yuan, G. Medioni, J. Kang, and I. Cohen, "Detecting motion regions in the presence of a strong parallax from a moving camera by multiview geometric constraints," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 6, pp. 1709–1724, Jun. 2011.
- [32] Y. Sheikh, O. Javed, and T. Kanade, "Background subtraction for freely moving cameras," in *Proc. IEEE 12th Int. Conf. Comput. Vision*, Sep./Oct. 2009, pp. 1219–1225.
- [33] Y. Y. Wu, X. H. He, and T. Q. Nguyen, "Moving object detection with a freely moving camera via background motion subtraction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 2, pp. 236–248, Feb. 2017.
- [34] M. Geng, X. Zhang, Y. Tian, L. Liang, and T. Huang, "A fast and performance-maintained transcoding method based on background modeling for surveillance video," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2012, pp. 61–67.
- [35] S. Chakraborty, M. Paul, M. Murshed, and M. Ali, "An efficient video coding technique using a novel non-parametric background model," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, Jul. 2014, pp. 1–7.
- [36] B. Dey and M. Kundu, "Efficient foreground extraction from HEVC compressed video for application to real-time analysis of surveillance big data," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3574–3585, Jun. 2015.
- [37] Z. Huang, R. Hu, and Z. Wang, "Background subtraction with video coding," *IEEE Signal Process. Lett.*, vol. 20, no. 11, pp. 1058–1061, Nov. 2013.
- [38] C. Chen, J. Cai, W. Lin, and G. Shi, "Surveillance video coding via low-rank and sparse decomposition," in *Proc. ACM Int. Conf. Multimedia*, Oct./Nov. 2012, pp. 713–716.
- [39] C. Chen, J. Cai, W. Lin, and G. Shi, "Incremental low-rank and sparse decomposition for compressing videos captured by fixed cameras," *J. Vis. Commun. Image Represent.*, vol. 26, pp. 338–348, Jan. 2015.
- [40] L. Zhao, X. Zhang, Y. Tian, R. Wang, and T. Huang, "A background proportion adaptive Lagrange multiplier selection method for surveillance video on HEVC," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2013, pp. 1–6.
- [41] G. J. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Process. Mag.*, vol. 15, no. 6, pp. 74–90, Nov. 1998.
- [42] K. Wang and J. A. Kangas, "Character location in scene images from digital camera," *Pattern Recogn.*, vol. 36, no. 10, pp. 2287–2299, Oct. 2003.
- [43] R. Carman and J. I. Greenberg, *An Editor's Guide to Adobe Premiere Pro*. Berkeley, CA, USA: Peachpit, 2012.
- [44] R. Fan, Y. F. Zhang, and B. Li, "Motion Classification-based fast motion estimation for high efficiency video coding," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 893–907, Jun. 2017.
- [45] B. Li, H. Li, L. Li, and J. Zhang, "Domain rate control algorithm for high efficiency video coding," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 3841–3854, Sep. 2014.
- [46] M. L. Zhou, Y. F. Zhang, B. Li, and X. P. Lin, "Complexity correlation-based CTU-level rate control with direction selection for HEVC," *ACM Trans. Multimedia Comput.*, vol. 13, no. 4, Aug. 2017, Art. no. 53.
- [47] M. Wang, K. N. Ngan, and H. Li, "Low-delay rate control for consistent quality using distortion-based Lagrange multiplier," *IEEE Trans. Image Process.*, vol. 25, no. 7, pp. 2943–2955, Jul. 2016.
- [48] "HM reference software 12.0." [Online]. Available: [http://hevc.hhi.fraunhofer.de/svn/svn\\_HEVCSoftware](http://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware)
- [49] L. Song, X. Tang, W. Zhang, X. Yang, and P. Xia, "The SJTU 4K video sequence dataset," in *Proc. 5th Int. Workshop Qual. Multimedia Experience*, Jul. 2013, pp. 34–35.



**Gang Wang** received the B.S. degree in electrical engineering from China West Normal University, Sichuan, China, in 2009, and the M.S. degree in computer science from Zhejiang Sci-Tech University, Hangzhou, China, in 2012. He is currently working toward the Ph.D. degree at Beihang University, Beijing, China. His current research interests include multimedia compression, image/video processing, machine learning, and hardware video codec design.



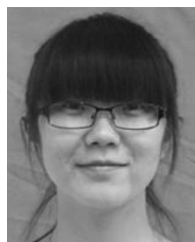
**Bo Li** received the B.S. degree in computer science from Chongqing University, Chongqing, China, in 1986, the M.S. degree in computer science from Xi'an Jiaotong University, Xi'an, China, in 1989, and the Ph.D. degree in computer science from Beihang University, Beijing, China, in 1993. In 1993, he joined the School of Computer Science and Engineering, Beihang University, where he has been a Full Professor since 1997. In 2002, he visited the University of Washington, Seattle, WA, USA, as a Senior Visiting Scholar for one year. He is currently the Director of

the Digital Media Laboratory, School of Computer Science and Engineering, and the Director of the Professional Committee of Multimedia Technology of the China Computer Federation. He has authored or co-authored more than 100 conference proceedings and journal papers in diverse research fields, including digital video and image compression, video analysis and understanding, remote sensing image fusion, and embedded digital image processors.



USA, from 2007 to 2009. His current research interests include image/video processing, compression and network transmission, and video surveillance.

**Yongfei Zhang** (M'12) received the B.S. degree in electrical engineering and the Ph.D. degree in pattern recognition and intelligent systems from Beihang University, Beijing, China, in 2005 and 2011, respectively. He is currently an Associate Professor with the Beijing Key Laboratory of Digital Media, School of Computer Science and Engineering, and the State Key Laboratory of Virtual Reality Technology and Systems, Beihang University. He was a Visiting Research Scholar with the Video Processing and Networking Lab, University of Missouri, Columbia, MO,



**Jinhui Yang** received the B.S. degree from Northeastern University at Qinhuangdao, Qinhuangdao, China, in 2015. She is currently working toward the M.S. degree in computer science engineering at Beihang University, Beijing, China. Her current research interest focuses on video coding.