High-efficiency Coding for Shaking Surveillance Videos Based on Global Motion Compensation

Lin Ding[1,3], Yonghong Tian[1,3*],
[1] National Engineering Laboratory for Video Technology,
School of EE&CS, Peking University, Beijing, China
[2] School of Information and Electronics,
Beijing Institute of Technology, Beijing, China

Hongfei Fan[1], Yaowei Wang[2*], Tiejun Huang[1,3]
[3] Cooperative Medianet Innovation Center, China
Email: ding.lin@pku.edu.cn, yhtian@pku.edu.cn,
hffan@pku.edu.cn, yaoweiwang@bit.edu.cn,
tjhuang@pku.edu.cn

***Abstract*—Due to the complex environment conditions, many surveillance videos are captured from cameras which are influenced by shaking more or less. This presents a significant challenge for background-modeling-based video coding since it is difficult to generate good background frames from such shaking videos. To solve this problem, this paper proposes a global motion compensation method using motion vectors (MV-GMC) for shaking surveillance video coding. In the proposed MV-GMC method, more accurate motion vectors (MVs) are extracted from HEVC encoder to estimate the global motion model in an efficient way, and we compensate each frame before background modeling. Then the compensated frames are used to model a good background frame for surveillance video coding. Compared with the optical-flow-based GMC (OPT-GMC) method which can be used to obtain more precise motion compensation, the proposed MV-GMC method has a comparable coding performance but a much lower computational complexity. Experiments on our surveillance video sequences show that the proposed MV-GMC method has significantly improved the coding performance by decreasing BD rate 49.83% over HM 12.0 on average while OPT-GMC can save 49.84% BD rate. The MV-GMC method also saves 92.71% background modeling time compared with the OPT-GMC method.**

Keywords—Surveillance Video coding, HEVC, Shaking, Background Modeling, GMC, Motion Vectors

I. INTRODUCTION

Video surveillance has a wide range of applications over the past years, and it is desirable for developing low-complexity and high-efficiency surveillance video coding methods. Our previous work [1] has introduced an efficient and practical coding scheme for surveillance videos captured by stationary cameras and the corresponding framework is shown in Fig.2. In [1], the background frame is periodically modeled and updated. The background frame is encoded with a low QP and foreground part is removed, which can be referenced by the following frames. Since background frame can be referenced more efficiently than regular Intra frames, coding performance will be increased with a well generated background frame.

However, the mentioned scheme has a poor performance for shaking surveillance videos because of the blurry background frame caused by shaking. In Fig. 1, (a) shows a part of one frame in a shaking surveillance sequence (EastGate) while (b) is the corresponding part which is modeled with Gaussian mixture model GMM [2] after 120 frames. As is shown in (b), the background frame is blurry and the textures are not well maintained. The blurry background will become a burden as a reference frame.

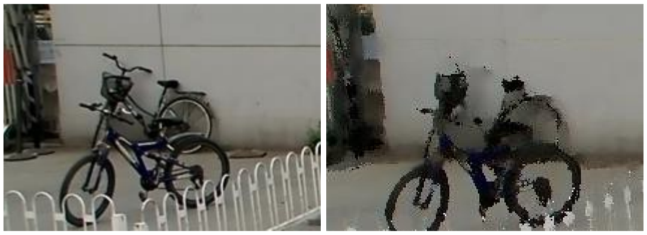
(a) (b)

Figure 1. Background frame modeled with GMM after 120 frames for a shaking surveillance sequence (EastGate). (a) a part of one frame in the sequence (b) corresponding part in the background frame

Many GMC algorithms have been proposed including pixel-based and vector-based approaches [3]. A pixel-based GMC method such as [4] has an enough precise result for most applications but also has high computational complexity. A vector-based GMC method such as [5] is less computational due to the availability of the block motion vectors in the bit stream. The vectors are essentially reused with the motivation to lower the computational complexity and avoid a repetition of motion estimation.

In order to solve the difficulty on background modeling for surveillance videos captured by shaking cameras, we proposed a vector-based method which is denoted as MV-GMC method based on HEVC in this paper. We applied MV-GMC in our existing surveillance video coding framework. Firstly, the first frame or a background frame is set to be a reference frame (denoted as ref-frame). We preserve the more accurate block motion vectors searched in TZ search according to SAD between each train set frame and the ref-frame with little extra computational cost. Secondly, we use our MV-GMC method to compensate the current frame to the ref-frame. Finally, a background modeling method is used to generate the background frame from the compensated frame and the background frame will be inserted into the bit-stream after modeling. The background frame is updated periodically as the ref-frame of the following frames.

* Yonghong Tian and Yaowei Wang are corresponding authors.

978-1-5090-2179-6/16 $31.00 © 2016 IEEE
DOI 10.1109/BigMM.2016.42

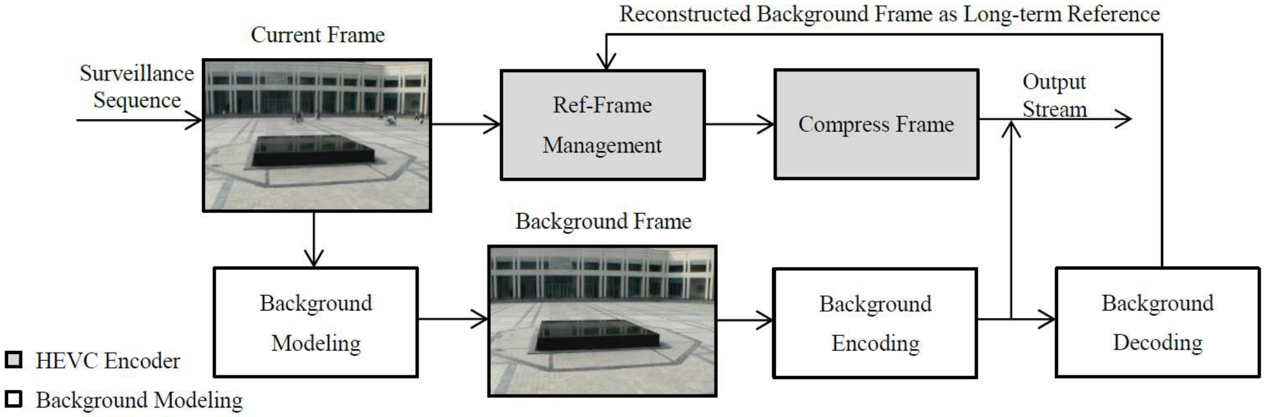

Figure 2. Framework of background-modeling-based surveillance video coding

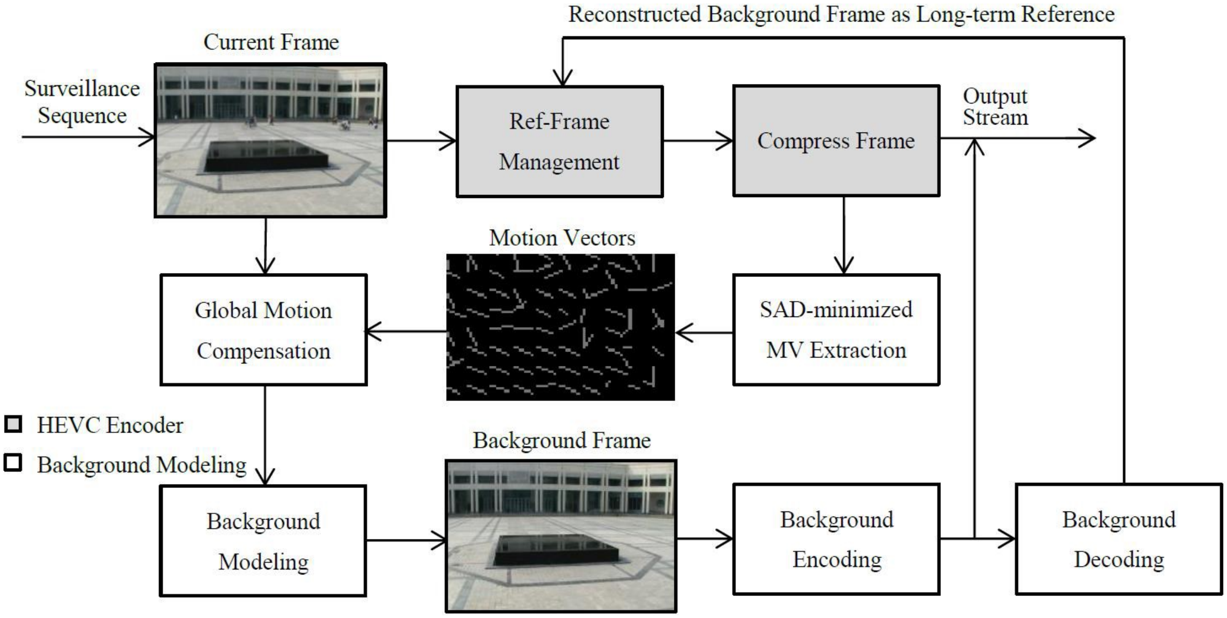

Figure 3. Framework of the proposed background-modeling-based shaking surveillance video coding

OPT-GMC is a classical pixel-based method which applies optical flow method to calculate global motion vectors between frames. Optical flow method is able to calculate precise global motion vector with high computational cost. Through our experiments conducted on our surveillance video sequences we show that our proposed MV-GMC has a similar performance with OPT-GMC but cost much lower time. The rest of the paper is organized as follows. Section II describes the proposed method. Section III presents the experimental results and the paper is concluded in section IV.

II. THE PROPOSED METHOD

A. The proposed surveillance video coding method

The outline of the proposed method is presented in Fig. 3. During the TZ search of each block, we preserve a SAD-minimized motion vector followed by refinement. Then we use RANdom SAmple Consensus (RANSAC) method [6] to filter out the SAD-minimized MVs and estimate the global motion, in which the global motion is expressed in an affine model. After that, the current frame is compensated to a reference frame and the compensated frame is used to model the background frame. The new background frame replaces the previous background frame and is referred by the following frames. Since all the train set frames are compensated to the same frame which is used to be a previous background frame, a latest background frame is generated without the influence caused by a shaking camera. We will give details of our method in the following two parts.

B. Distortion-Minimizing motion vector extracting

GMC method based on optical flow method (denoted as OPT-GMC) is a method applying optical flow to calculate global motion vectors between frames and then compensate these frames to generate a background frame with GMC algorithms. OPT-GMC has a precise result but a high computational complexity, and most of the time feature extraction is needed before motion estimation. In this part, a

MV-based GMC method is proposed which has a similar performance but a much lower computational cost than OPT-GMC.

In HEVC reference software, TZ search [7] is used in motion estimation and different Predict Unit (PU) sizes are used. We define the set of motion vectors searched by TZ search for one block as $\{TZSearch_{pu\_size}^{pos}\}$, where the superscript pos denotes the position of the block in one frame and the subscript pu_size ranging from 1 to 4 denotes the PU size 8×8, 16×16, 32×32 and 64×64 respectively. Other PU sizes such as 16×8, 8×16 are not considered for a temporary simplification and the reason will be discussed in the following part. The target function of motion estimation in HEVC is as follow,

$$MV_{COST}{}_{pu\_size}^{pos} = \min_{\{TZSearch_{pu\_size}^{pos}\}} Cost \tag{1}$$

where $Cost$ represents the Rate-Distortion (RD) cost of the corresponding MV, and $MV_{COST}{}_{pu\_size}^{pos}$ denotes the best MV searched by the encoder with best RD cost. Thus, the motion estimation problem can be formulated as minimize the RD cost with a set of MVs.

The definition of RD cost in motion estimation is as follows,

$$Cost_{IME} = SAD + \lambda \times Bits_{MV}$$
$$Cost_{FME} = SATD + \lambda \times Bits_{MV} \tag{2}$$

where $Cost_{IME}$ and $Cost_{FME}$ represent the RD cost for Integer Motion Estimation (IME) and Fraction Motion Estimation (FME) respectively. *SAD* represents the Sum of Absolute Difference between reference block and current block. *SATD* represents the Sum of Absolute Transformed Difference which is calculated by the sum of Hadamard transformed coefficients. $Bits_{MV}$ represents the numbers of bits of the block encoded by entropy coding. λ is a parameter to balance the *SAD* or *SATD* and bits in RD cost.

TABLE I. PERFORMANCE OF BD RATE AMONG DIFFERENT CASES

Sequence	MV_{COST}	MV_{SAD}				
	32×32	**All**	**64×64**	**32×32**	**16×16**	**8×8**
Library	11.16%	-24.90%	-61.92%	-61.40%	-61.68%	31.95%
Eastgate	7.16%	-10.41%	-18.16%	-18.11%	-9.25%	14.68%
Square	-45.16%	-53.59%	-64.01%	-63.72%	-63.48%	-8.13%
Building	-15.11%	-37.20%	-55.93%	-56.10%	-55.41%	18.64%

Eq.1 insures the encoder to choose a best MV for one block to have a better performance in RD-cost, which means the MV that have a smallest SAD value may not be chosen by the encoder because of a larger $Bits_{MV}$. However, the MV we need in background modeling is different from the one get from the encoder. A more precise MV is required without the consideration of $Bits_{MV}$ but only *SAD*. Therefore, the idealized condition of precise MV to solve our problem is as follows,

$$MV_{SAD}{}_{pu\_size}^{pos} = \min_{\{FullSearch_{pu\_size}^{pos}\}} SAD \tag{3}$$

where $\{FullSearch_{pu\_size}^{pos}\}$ denotes the set of MVs searched by full search. However, full search will definitely cost unbearable computing time. Therefore, we maintain a variable $MV_{TZ\_SAD}{}_{pu\_size}^{pos}$ in TZ search to store the MV which have the best SAD value, which is defined as,

$$MV_{TZ\_SAD}{}_{pu\_size}^{pos} = \min_{\{TZSearch_{pu\_size}^{pos}\}} SAD \tag{4}$$

After TZ search, an extra full search is used for refinement with an edge length 4 around the corresponding reference block. We denote $MV_{Ref\_SAD}{}_{pu\_size}^{pos}$ as the refined $MV_{TZ\_SAD}{}_{pu\_size}^{pos}$ which is defined as,

$$MV_{Ref\_SAD}{}_{pu\_size}^{pos} = \min_{\{RS(MV_{TZ\_SAD}{}_{pu\_size}^{pos})\}} SAD \tag{5}$$

where $RS\left(MV_{TZ\_SAD}{}_{pu\_size}^{pos}\right)$ denotes MVs surrounded $MV_{TZ\_SAD}{}_{pu\_size}^{pos}$ with an edge length 4 and $\left\{RS\left(MV_{TZ\_SAD}{}_{pu\_size}^{pos}\right)\right\}$ denotes the set of them.

We define $\{MV_{SAD}\}$ as the set of all the $\mathrm{MV}_{Ref\_\mathrm{SAD}}{}_{\mathrm{pu\_size}}^{\mathrm{pos}}$ with different position and PU sizes in one frame as,

$$\{MV_{SAD}\} = \begin{Bmatrix} \mathrm{MV}_{Ref_{\mathrm{SAD}}}{}_{\mathrm{pu\_size}}^{\mathrm{pos}} | 1 \le pos \le N \\ pos \in Z, p\mathrm{u\_size} \in \{1,2,3,4\} \end{Bmatrix} \tag{6}$$

where pu_size ranging from 1 to 4 denotes the PU size 8×8, 16×16, 32×32 and 64×64 respectively.

The global motion vector can be calculated by RANSAC with $\{MV_{SAD}\}$.

Since different PU sizes are searched in HEVC, these different PU sizes defined in Eq.6 are not all useful when estimating global motion. For example, $\mathrm{MV}_{Ref_{\mathrm{SAD}}}{}_{1}^{\mathrm{pos}}$ may leads to a poorer performance because of overmatching. What's more, the refinement described in Eq.5 for 8×8 PUs will cost more time. Therefore, we did some experiments focusing on the mentioned problem.

Six cases are tested and HEVC reference software HM 12.0 is employed as the anchor. The RD performance of Luma is shown in Table I. For the first case, $MV_{COST}{}_{pu\_size}^{pos}$ defined in Eq.1 is used as final block motion vector for GMC in framework given by Fig.3 (denoted as "MV_{COST}") with 32×32 PU size and 2N×2N prediction mode is used. For the rest five cases, $MV_{TZ\_SAD}{}_{pu\_size}^{pos}$ defined in Eq.4 is used (denoted as "MV_{SAD}"). In these five cases, "All" denotes that all the PU sizes are considered as defined in Eq.6, and "8×8", "16×16", "32×32", "64×64" denote only one PU size is counted in Eq.6. For example, "8×8" means only $\mathrm{MV}_{COST}{}_{1}^{\mathrm{pos}}$ is considered. As a result in Table I, "64×64", "32×32" and "16×16" obtain a much better performance than "8×8" and "All".

```
$\{MV_{SAD}\}$ = Ø; = 3;
for pos = 1; pos < N; pos++ do
    $SAD_{pu\_size}^{pos}$= -1,
                    //Calculate $MV_{SAD}{}_{pu\_size}^{pos}$
    if RefFrame & PU = 32×32 & PreMode = 2N×2N then
        for MV in $\{TZSearch_{pu\_size}^{pos}\}$ do
            calculate SAD;
            if < 0 or > SAD then
                $SAD_{pu\_size}^{pos}$= SAD;
                $MV_{SAD}{}_{pu\_size}^{pos}$= MV;
            end if
        end for
    end if
                    //Refine $MV_{Ref\_SAD}{}_{pu\_size}^{pos}$
    for x = -4; x < 4; x++ do
        for y = -4; y < 4; y++ do
            calculate SAD;
            if $SAD_{pu\_size}^{pos}$> SAD then
                $SAD_{pu\_size}^{pos}$= SAD;
                $MV_{Ref\_SAD}{}_{pu\_size}^{pos}$= $MV_{SAD}{}_{pu\_size}^{pos}$+(x, y);
            end if
        end for
    end for
    $\{MV_{SAD}\} = \{MV_{SAD}\} \cup MV_{Ref\_SAD}{}_{pu\_size}^{pos}$
end for
```

Algorithm 1: $\{MV_{SAD}\}$ searching method

According to our experimental results, we find that overmatching happens for smaller PU sizes. What's more, MVs of regular 32×32 PUs are enough to estimate global motion vector and take more MVs of other different PUs into consider will also be time consuming. Considering the analysis in Table I, we choose "32×32" as the final method which avoids overmatching and saves computational time. Therefore, Eq.6 is fixed as,

$$\{MV_{SAD}\} = \left\{ \begin{matrix} \mathrm{MV}_{Ref\mathrm{SAD}}{}_{\mathrm{pu\_size}}^{\mathrm{pos}} | 1 \leq pos \leq N \\ pos \in Z, \quad \mathrm{pu}_{\mathrm{size}} = 3 \end{matrix} \right\} \tag{7}$$

The whole $\{MV_{SAD}\}$ searching method is given in Algorithm 1. In Algorithm 1, "*PreMode*" represents the inter prediction mode in HEVC. "*RefFrame*" represents the current reference frame is the first frame or the background frame. "*PU*" represents the current PU size.

C. Motion vectors filtering and frame warping

Since the reference frame is not the original frame but the reconstructed frame, $\{MV_{SAD}\}$ estimated from the encoder are often imperfect. The result will also affected by noise or foreground objects. Therefore, RANSAC is employed to robustly compute the best estimate of the motion model while identifying motion vectors conforming to the background. During RANSAC, a set of six motion vectors is randomly selected. We use the six motion vectors to estimate a global motion model, and identify the motion vectors fitting to the motion model. After a number of iterations, a best motion model among the estimated global motion models is chosen according to the motion vectors fitting to it.

Since 2D models are more robust and faster estimating a linear transformation between consecutive frames, we take the 2D affine model in RANSAC with least-square solution to get the global motion. The affine motion model is defined by a mapping between coordinates ($x = [x, y]^T$ and $\hat{x} = [\hat{x}, \hat{y}]^T$) of corresponding pixels in a pair of frames, parameterized by a set of parameters,

$$\hat{x} = \mathbf{A}x + \boldsymbol{b} \tag{8}$$

where

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \qquad \boldsymbol{b} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \tag{9}$$

where **A** and ***b*** contain some basic motions of a camera including translation, rotation and scale. Those basic motions are sufficient for representing moving cameras in surveillance videos.

D. Gaussian mixture model

Then we warp the current frame onto the ref-frame plane according to the global motion model using bilinear interpolation. After that, the compensated frame is used to model the background frame using GMM method. Gaussian mixture model (GMM) [2] allows multimodal background models and objects blending to, or permanently leaving, the background. In [2], Stauffer and Grimson raised the case for a multi-valued background model corresponding to multiple background objects.

During background model estimation, the Gaussian distribution which has the most supporting evidence and the least variance is assumed to be most likely from background. The Gaussians are ordered by the value of ω / σ, which increases as a distribution gains more evidence (ω) or as the variance (σ) decreases. This ordering makes the most likely distributions remain on top and the less probable distributions move to the bottom and replaced finally. Then the first B distributions are chosen as the background model,

$$B = \underset{b}{argmin} \left(\sum_{k=1}^{b} \omega_k > T \right) \tag{10}$$

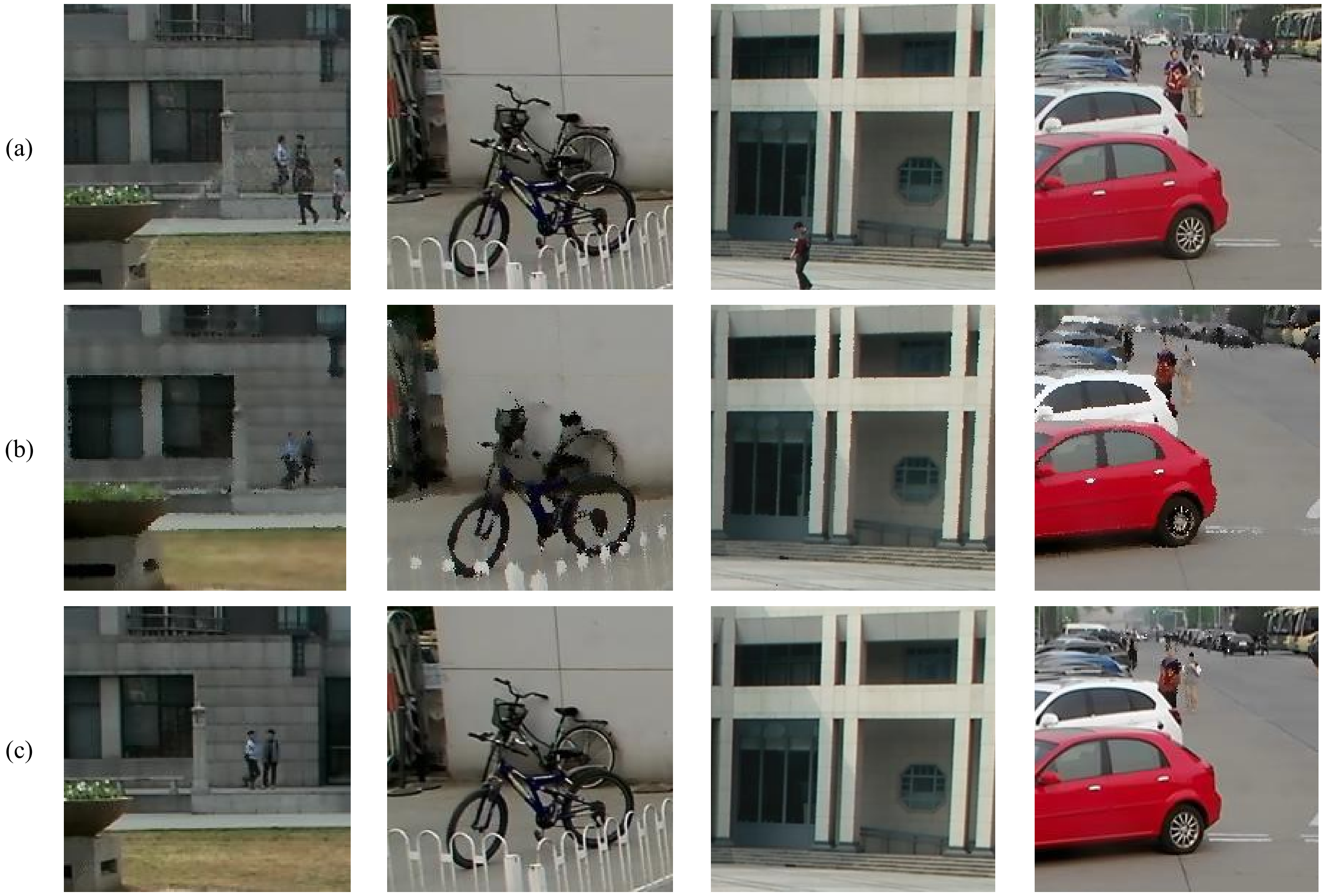

Figure 4. Comparison of background frames between GMM and MV-GMC (a) the original frames (b) the background frames by GMM (c) the background frames by MV-GMC

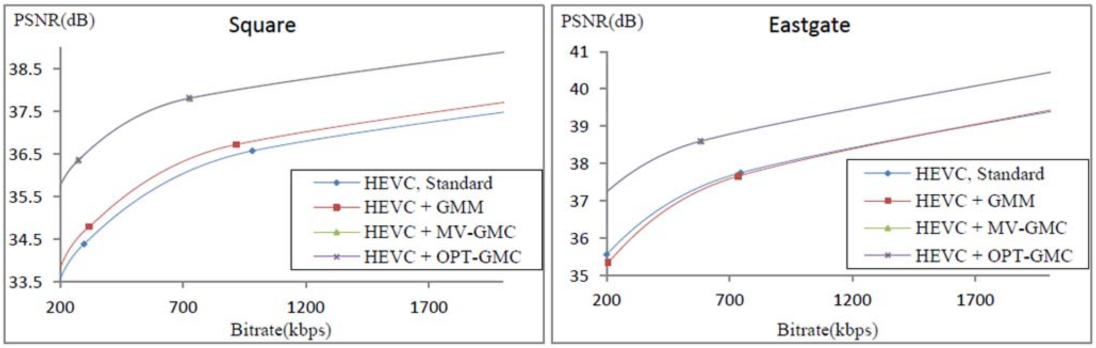

Figure 5. Coding performance of test sequences Square and Eastgate. The proposed MV-GMC (purple) is compared with OPT-GMC (green), GMM (red) and a standard HEVC encoder (blue).

where T is a measure of the minimum portion of the data that should be accounted for by the background. Usually, using the most probable distribution will save processing.

III. Experimental Result

A. Setup

The proposed scheme is integrated in HEVC reference software HM 12.0. Four different shaking surveillance sequences and two stationary sequences are evaluated with Lowdelay configuration. Table II gives the information about these sequences. 720 frames are tested and the beginning 120 frames are used for background modeling. The first frame is referenced by the beginning 120 frames. After that, the background frame is encoded with long term QP and inserted into the stream as the 121th frame which is referenced by the following 600 frames. Our BD rate is obtained when QP are 22, 27, 32, and 37 while long term QP are 12, 17, 22, and 27.

TABLE II. DESCRIPTION OF THE SEQUENCES

Sequence	Resolution	Camera Motion	Foreground Portion
Library	1280×720	Large	Low Percent
Eastgate	1280×720	Medium	High Percent
Square	1280×720	Small	Low Percent
Building	1280×720	Medium	Medium Percent
Campus	720×576	Stationary	Medium Percent
Crossroad	720×576	Stationary	High Percent

TABLE III. PERFORMANCE OF BD RATE AND TIME COST AMONG GMM, MV-GMM AND OPT-GMM

Motion	Sequence	GMM	OPT-GMC	MV-GMC	
		BD rate	BD rate	BD rate	TS BG
Stationary	Campus	-45.22%	-44.74%	-43.47%	/
	Crossroad	-34.76%	-32.82%	-34.52%	/
	Average	**-39.99%**	**-38.78%**	**-39.00%**	**/**
Shaking	Library	16.13%	-61.79%	-61.40%	92.41%
	Eastgate	10.29%	-17.80%	-18.11%	91.84%
	Square	-13.87%	-63.77%	-63.72%	91.32%
	Building	7.99%	-55.98%	-56.10%	93.55%
	Average	**5.14%**	**-49.84%**	**-49.83%**	**92.71%**

In OPT-GMC, we employ Lucas-Kanade (LK) [8] algorithm with a three-level pyramid as the optical flow method and Scale-Invariant Feature Transform (SIFT) as the feature extraction method. In RANSAC, the threshold t and the number of motion vectors selected to estimate motion model during every iteration n are set as t=0.5 and n=6 throughout our experiments. At the end of GMM, we choose the most probable distribution to generate our background frame.

HEVC reference software HM 12.0 is employed as the anchor. We implements three methods. The first one is using the framework given in Fig.2 (denoted as "GMM"). The other two methods are using the framework given in Fig.3 with OPT-GMC and MV-GMC methods respectively mentioned in part B of Section II (denoted as "OPT-GMC" and "MV-GMC").

B. Result

The background frames modeled after 120 frames by "GMM" and "MV-GMC" are compared in Fig 4. As shown in Fig 4, background frames of shaking sequences modeled by "MV-GMC" maintain the texture while those modeled by "GMM" are blurred. The better quality of background frames leads to a better performance in BD rate.

Fig.5 shows RD curves of our proposed "MV-GMC" based encoder for two test shaking sequences Square and Eastgate. The proposed "MV-GMC" has a similar performance with "OPT-GMC" but a significant improve than "GMM" and the anchor. Table III gives BD rate performance and time cost among "GMM", "OPT-GMC" and "MV-GMC". The "BD rate" column gives the BD rate performance of the three methods compared with the anchor. The "TS BG" column represents the time saving of "MV-GMC" compared with "OPT-GMC" in background modeling time which is defined as follow,

$$TS\ BG = \frac{BG\ time\ of\ OPT\_GMC - BG\ time\ of\ MV\_GMC}{BG\ time\ of\ OPT\_GMC} \tag{11}$$

where BG time is the background modeling time including the GMC part and the GMM part. BG time is measured on a Core i7 processor and no parallel optimizations are applied.

As shown in Table III, for stationary sequences, "GMM" performs well with saving 39.99% BD rate in average compared with the anchor. "OPT-GMC" and "MV-GMC" save 38.78% and 39.00% BD rate in average respectively. Due to the effect of foreground motions, the global motion calculated in "MV-GMC" is not zero which cause a little side effect.

For shaking sequences, "GMM" increases 5.14% BD rate in average, while "OPT-GMC" and "MV-GMC" save 49.84% and 49.83% BD rate in average respectively. "MV-GMC" saves 92.71% background modeling time in average of that using "OPT-GMC". In sequences with higher percent foreground portion such as Eastgate and Building, "MV-GMC" performs a little better than "OPT-GMC". The main reason is that foreground is more likely to be extracted by SIFT and the MVs of foreground will be outliers for GMC in "OPT-GMC". As a result, "MV-GMC" has a similar performance in BD rate with "OPT-GMC" and a much better performance than "GMM". Meanwhile, "MV-GMC" has a much lower computational complexity than "OPT-GMC".

IV. CONCLUSIONS

In this paper, we propose a fast and efficient MV-GMC method for background-modeling-based shaking surveillance video coding. MV-GMC reuses the TZ search in HEVC and extracts more accurate motion vectors with little extra computational cost. The results indicate that the MV-GMC based encoder is robust against moving objects and camera motions, and it has a similar coding performance but a much lower computational cost than the OPT-GMC based encoder.

ACKNOWLEDGMENT

This work is partially supported by the National Basic Research Program of China under grant 2015CB351806, the National Natural Science Foundation of China under contract No. 61425025, No. 61390515, No. 61421062 and No. 61471042, and Shenzhen Peacock Plan.

REFERENCES

[1] X. Zhang, L. Liang, Q. Huang, Y. Liu, T. Huang, and W. Gao, "An efficient coding scheme for surveillance videos captured by stationary cameras," in P Proc. Visual Commun. Image Process, VCIP, Huang Shan, An Hui, China, vol.7744, pp.77442A-1-10, Jul. 2010.

[2] C. Stauffer and W. E. L. Grimson, “Adaptive backgroundmixture models for real-time tracking,” in IEEE Conf. on Computer Vision and Pattern Recognition (1999).

[3] M. Haller, A. Krutz, and T. Sikora, “Evaluation of pixe- land motion vector-based global motion estimation for camera motion characterization,” in Proc. WIAMIS, London, UK, pp. 49-52, May 2009.

[4] J.-M. Odobez and P.Bouthemy,”Robust multiresolution estimation of parametric motion models,” J. Visual Commun. Image Represent., vol. 6, No. 4, pp.348-365, 1995

[5] Y. Su, M.-T. Sun, and V. Hsu, “Global motion estimation from coarsely sampled motion vector field and the applications,” IEEE Trans. Circuits Syst. Video Technol., vol. 15, no. 2, pp.232-242, 2005

[6] M. Fischler and R. Bolles, “Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography”, Comm.ACM, vol.24, no.6, pp.381-395, Jun. 1981.

[7] I.K. Kim, K. McCann, K. Sugimoto, B. Bross and W.-J. Han, “High Efficiency Video Coding (HEVC) Test Model 10 (HM10) Encoder Description,” Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, 12th Meeting, Geneva, CH, 14-23, Jan. 2013.

[8] B. D. Lucas and T. Kanade, “An iterative image registration technique with an application to stereo vision,” Proceedings of the 1981 DARPA Imaging Understanding Workshop, pp. 121-130, 1981.