MULTISOURCE SURVEILLANCE VIDEO CODING BY EXPLOITING 3D AND 2D KNOLWEDGE

Yu Chen[1,4], *Ruimin Hu*[2], *Jing Xiao*[1,4], *Zhongyuan Wang*[3]

1. National Engineering Center of Multimedia and Software, School of Computer, Wuhan University
2. Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University
3. Collaborative Innovation Center of Geospatial Technology, Wuhan, China
4. Suzhou Institute of Wuhan University, Jiangsu, China

ABSTRACT

The rapidly increasing surveillance video data has challenged the existing video coding standards. Even though knowledge based video coding scheme proposed for moving objects so far has achieved high efficiency, it does not take full advantages of local information and highly relies on the accuracy of pose parameter of the objects, thus leading to large prediction residuals. In this paper, a novel surveillance video coding utilizing 3D and 2D knowledge is proposed. On the one hand, we generate a knowledge based reference frame from 3D models of the objects and incorporate it into the block based coding framework to remove global redundancy while improve the robustness to pose errors. On the other hand, 2D knowledge in the form of visual appearances of the objects in the previously encoded frames is employed to rectify the knowledge based reference frame for local redundancy removal. Experimental results demonstrate the effectiveness of our proposed method against HEVC and the knowledge based coding method.

Index Terms— surveillance video coding, block based coding, 3D and 2D knowledge, redundancy removal.

1. INTRODUCTION

At present, video surveillance system has become one of the most important urban infrastructures owing to its wide range of applications in traffic and public security. With the high definition trends of surveillance cameras, large amount of video data is generated every day. Hence, there is an urgent demand for high-efficiency video coding methods.

Different from generic videos, surveillance data is usually decomposed into dynamic foreground and static background and then processed separately to take full advantages of their characteristics. Moreover, a previous study [1] reported that 99% of the bit cost on average is used for moving objects when coding surveillance videos, in which 77% is used for moving vehicles. There are more opportunities to further reduce the bits spent on foreground than those spent on background. Therefore, this study mainly focuses on coding foreground objects, especially the vehicles.

Fig. 1 Drawbacks of the knowledge based coding

Among the existing works, object based coding can date back to model based methods [2-4], which model the objects-of-interest then encode the model parameters and the remaining contents. Various kinds of information, e.g. color [5], motion [6], shape [7], mesh [8], have been utilized to model foreground objects. However, these proposed object models are either too simple to maintain the quality of reconstructed objects or too complicated thus causing too much overhead, few of which achieved success. In order to balance the representation accuracy and model complexity, [9] resorted to the object or block based motion compensation. The performance is greatly improved since the relations of objects between adjacent frames are exploited. Nonetheless, the above methods model motions of object, which actually lie in 3D space in the real world, by translation on 2D image plane, thus having difficulty in dealing with rotation or scaling variations. In addition, most of the redundancies within single video can be effectively removed after extensively studied in the past decades and there is little room for improvement when coding each video individually.

As for jointly coding multiple videos, most of the research efforts are devoted to multiview video coding [10-12], which focuses on eliminating redundancies between adjacent views. Multiview video coding can be generally classified into two categories. The first one utilizes the coded frames from different views as the reference for the frames from the current view [13]. The other one takes advantage of the depth information to synthesize the virtual view point [14]. In this way, only few views and some extra depth maps need to be encoded. One important characteristic of the multiview videos is that the cameras are well arranged around the target scene or object to meet the requirement of enough overlap

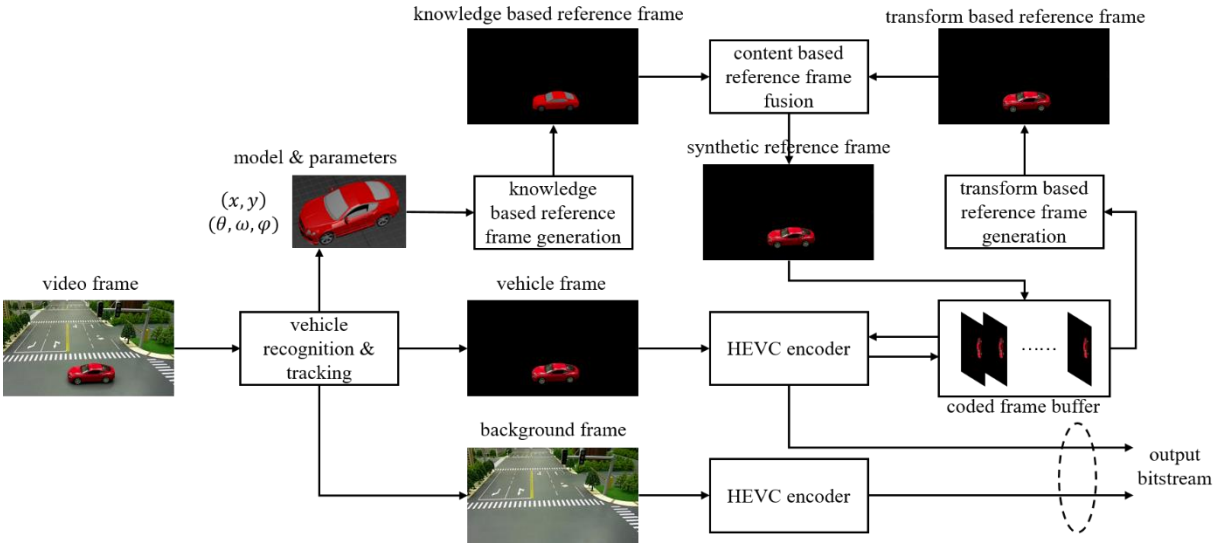

Fig. 2 Overall framework of the proposed scheme

between adjacent views. However, the regions, which different surveillance cameras cover, seldom overlap with each other. Hence, multiview video coding methods provide a heuristic insight to remove redundancies among multiple videos yet cannot be directly applied to surveillance videos.

Recently, multisource surveillance video coding [15] emerges, which takes all the surveillance videos within a large region into account and jointly encodes them. It exploits the fact that the appearances of a certain object in different surveillance videos are highly related. Each appearance of the object can be roughly considered as the perspective projection of its 3D model with corresponding pose parameters under the ideal circumstances. Therefore, 3D models are projected to the 2D image plane to generate an object based prediction in [15]. Despite the high efficiency multisource surveillance video coding achieves, there are still two major drawbacks presented in Fig. 1. First, this method is not flexible and heavily relies on the accuracy of the pose parameters due to its object based nature. Second, this method only employs the global knowledge to generate the prediction, ignoring the local information. Hence, the prediction is usually far away from its actual appearance.

To address the abovementioned problems, we propose a multisource surveillance video coding method by exploiting 3D and 2D knowledge. On the one hand, we generate a knowledge based reference frame utilizing 3D models and introduce it into the block based coding framework, where motion estimation can compensate the shift caused by pose errors to some extent. On the other hand, we employ the 2D knowledge, which is in the form of visual appearances of the vehicle in the encoded frames, and adaptively fuse it with the knowledge based reference frame to exploit both the global and local information. In this way, the proposed coding framework is not only more robust to pose errors, but also effective in removing the global and local redundancies in multisource surveillance videos.

The reminder of this paper is organized as follows. Details of our proposed method are given in section 2. Experimental results and analysis are presented in section 3, and the conclusion is drawn in Section 4.

2. PROPOSED METHOD

2.1. Overall framework

The overall framework of the proposed method is shown in Fig. 2. First, vehicle model recognition and tracking are adopted to extract the vehicles from the video frames, during which the pose parameters can also be obtained. Then, we search for the 3D model of the vehicle and project it to the image plane with the corresponding pose parameters to generate the knowledge based reference frame. Similarly, a transform based reference frame is generated from the encoded frames with the help of 3D transform. After that, we divide each frame into several regions according to the video content. Based on the division, different strategies are adopted to fuse the knowledge based reference frame and the transform based reference frame. Finally, we add the synthetic reference frame into the reference picture list and utilize it for inter-frame prediction. As for the background frames, we simply send them to the HEVC encoder.

2.2. Knowledge based reference frame generation

For the $i\ th$ vehicle in the $t\ th$ frame, we first recognize its vehicle model Mod_i. During the vehicle recognition process, each possible model is iteratively projected to the image plane to match the appearance of vehicle in the frame. Hence, we can obtain the corresponding pose $(\theta_i^t, \omega_i^t, \varphi_i^t)$ and position (x_i^t, y_i^t) when the optimal match is found. Given the 3D model and parameters as well, the synthetic vehicle appearance SV_i^t can be expressed as

$$SV_i^t = K \cdot [R|T] \cdot Mod_i, \tag{1}$$

where K is the camera intrinsic matrix which is calibrated in advance, R presents the rotation matrix obtained from the pose parameters by Rodrigues Transform and T denotes the column vector form of the position parameters.

Note that the synthetic vehicle appearance SV_i^t is only a part of the knowledge based reference frame and the rest part of the frame is filled with zeros.

2.3. Transform based reference frame generation

In order to introduce local information into the synthetic reference frame, the previously encoded frames are employed. Since the pose and position of the vehicle vary in different frames, we need to transform the pose of the vehicle in the coded frame to that in the knowledge based reference frame. According to the pose and the position parameters, each pixel of vehicle in the coded frame is first back projected to the 3D model and then projected to the knowledge based reference frame. This process can be formulated as

$$V_i^t = P_i^t \cdot P_i^{t-1^{-1}} \cdot V_i^{t-1}, \tag{2}$$

where V_i^t and V_i^{t-1} denote the vehicle region in the current frame and the coded frame respectively. P_i^t is the projection matrix that equals to the extrinsic matrix $[R|T]$ in Eq. (1).

As the same as knowledge based reference frame, we fill the regions excluding the V_i^t with zeros to generate the transform based reference frame.

2.4. Content based reference frame fusion

After obtaining the knowledge based reference frame and the transform based reference frame, we adaptively fuse them based on the video content to generate the synthetic reference frame containing both global and local information. The motivation of this idea comes from the following facts. The vehicle in the knowledge based reference frame differs a lot in color from the current frame while its structure, edges, contours are clear, as can be seen in Fig. 1(b). The vehicle in the transform based reference frame looks similar to the current frame but the edges and contours are missing or blurred due to the encoding artifacts. It is natural to divide video content into flat regions that mostly present the color of the vehicle and structural regions that include edges or contours, then design fusion strategies for them respectively.

To classify the video content, the structure tensor is adopted, which can be expressed as

$$M = \begin{bmatrix} g_x^2 & g_x g_y \\ g_x g_y & g_y^2 \end{bmatrix}, \tag{3}$$

where M is the structure tensor, g_x and g_y denote the gradient magnitude of the transform based reference frame in the horizontal direction and vertical direction respectively. The flat regions and the structural regions can be detected according to the following criterion:

$$\begin{cases} pix(x,y) \in r(f) & if\ tr(M) = 0 \\ pix(x,y) \in r(s) & otherwise \end{cases}. \tag{4}$$

In Eq. (4), $pix(x,y)$ denotes the pixels whose coordinate is (x,y). $r(f)$ and $r(s)$ are the flat regions and structural regions. $tr(M)$ presents the trace of the matrix M.

For the flat regions, the knowledge based reference frame is far away from the current frame. Hence, we only utilize the information from the transform based reference frame. For the structural regions, the knowledge based reference frame and the transform based reference frame are combined to enhance the edges and contours. Considering that the fusion coefficients need to be transmitted for the decoding purpose and complicated model will result in too much overhead, we employ the simple linear model for the reference frame fusion, which can be expressed as

$$Ref_S(x,y) = \begin{cases} \alpha_1 \cdot Ref_K(x,y) + \beta_1 \cdot Ref_T(x,y) + b_1 & if\ (x,y) \in r(s) \\ \alpha_2 \cdot Ref_T(x,y) + b_2 & if\ (x,y) \in r(f) \end{cases}. \tag{5}$$

In Eq. (5), Ref_S, Ref_K and Ref_T denote the final synthetic reference frame, knowledge based reference frame and transform based reference frame respectively. α, β, b are the fusion coefficients that can be obtained by solving the following optimization problem:

$$\begin{cases} \underset{\alpha_1,\beta_1,b_1}{\operatorname{argmin}} \|VF - (\alpha_1 \cdot Ref_K + \beta_1 \cdot Ref_T + b_1)\|_2 \\ \underset{\alpha_2,b_2}{\operatorname{argmin}} \|VF - (\alpha_2 \cdot Ref_T + b_2)\|_2 \end{cases}, \tag{6}$$

where VF is the current frame that only contains the vehicle.

3. EXPERIMENTS

3.1. Experimental setup

To evaluate the performance of the proposed multisource surveillance video coding method, we employ the HEVC, which is implemented in HM 16.20 and configured with Low Delay P Main Profile [16], as one of the comparison method. In addition, the state-of-the-art multisource surveillance video coding method, knowledge based coding (KBC), proposed in [15] is also chosen for performance evaluation.

For the convenience of controlling the experimental environment, we establish a dataset that contains two groups of simulated video clips. Each group contains nine clips that taken by different cameras with the resolution of 1080p while each video clip contains 300 frames. They capture the same red vehicle running in the simulated urban environment under different conditions: (1) under stable illumination; (2) under varying illumination. Some samples of the simulated video clips are given in Fig. 3.

3.2. Performance on vehicle regions with accurate pose parameters

First, we conduct experiments on vehicle frames excluding the background under ideal conditions where the pose of vehicle in each frame is manually calibrated. To present the results intuitively, we calculate the average BD-Rate and BD-PSNR [17] of nine clips for each group.

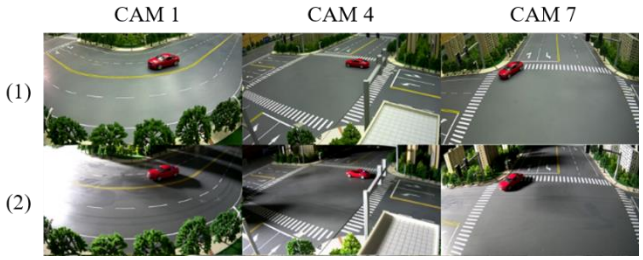

Fig. 3 Samples of the test sequences

Table 1. Performance with accurate pose parameters

	BD-rate (%)		BD-PSNR (dB)	
	vs. HM	vs. KBC	vs. HM	vs. KBC
group (1)	-43.17	-6.17	2.01	0.34
group (2)	-38.49	-15.74	1.65	0.68
average	-40.83	-10.96	1.83	0.51

Table 1 shows the BD-rate and BD-PSNR of the proposed method over HEVC and KBC. 40.83%/10.96% bitrate savings and 1.83/0.51dB PSNR gains are achieved compared with HEVC and HM respectively. It can be observed that both the proposed method and KBC remarkably outperform the HEVC due to the 3D knowledge utilization. Moreover, we can see that the proposed method has the close performance on group (1) but improves a lot on group (2) compared with KBC. The main reason is that the proposed method exploits the local information from the encoded frames to compensate the appearance difference caused by environmental factors, thus achieving better performance.

3.3. Performance on vehicle regions with inaccurate pose parameters

In this experiment, pose estimation method proposed in [18] is adopted to calculate the pose parameters. As the author reported, the estimation error is around 10 degrees.

Table 2. Performance with inaccurate pose parameters

	BD-rate (%)		BD-PSNR (dB)	
	vs. HM	vs. KBC	vs. HM	vs. KBC
group (1)	-35.21	-12.77	1.56	0.43
group (2)	-31.44	-21.23	1.38	1.01
average	-33.32	-17.01	1.47	0.72

As can be seen in Table 2, the KBC suffers a significant performance drop when using the inaccurate pose parameters. As for the proposed method, the bitrate savings and PSNR gains slightly decrease. Compared with KBC, the proposed method is more robust to the pose errors owing to the flexibility that block based framework provides.

3.4. Usage percentage of synthetic reference frame

In addition to the test of coding efficiency, we also count the number of prediction units (PUs) predicted via each reference frame. The usage percentage p_n is defined as the percentage of PUs using the $n\ th$ reference frame. In this experiment, we take all the PUs in the two groups of video clips into account.

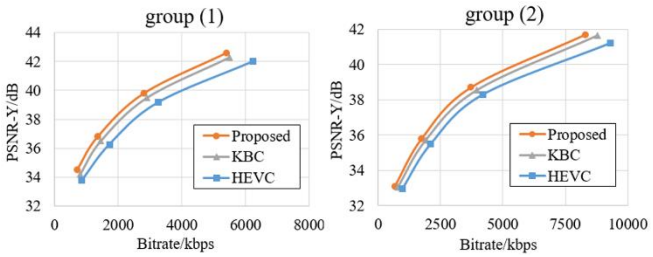

Fig. 4 RD curves of the performance on the whole videos

The reference frame numbered with 0 is the proposed synthetic reference frame, the 1-4 th reference frames are selected according to the configuration profile of HEVC.

Table 3. Usage percentage of each reference frame

n	0	1	2	3	4
p_n (%)	51.08	33.46	10.93	3.34	1.19

From Table 3 we can see that the usage percentage of the proposed synthetic reference frame is higher than that of the others, which demonstrates the effectiveness of the proposed method.

3.5. Performance on the whole video clips

In this experiment, we utilize the framework introduced in section 2.1 to encode the whole video clips. As described before, we calculate the average BD-Rate and BD-PSNR of nine clips for each group. The RD curves are shown in Fig. 4.

Averagely, the proposed hybrid prediction based coding framework achieves 23.35% bitrate savings and 1.17dB PSNR gains over HEVC while 10.67% and 0.44dB over KBC.

4. CONCLUSION

In this paper, we proposed a multisource surveillance video coding method by exploiting 3D and 2D knowledge. We first generate a knowledge based reference frame by projecting the 3D model to the image plane with the pose parameters. In the meantime, a transform based reference frame is constructed based on the 2D encoded frames. Then, we adaptively fuse the knowledge based and the transform based reference frames according to the video contents to exploit both the 3D and 2D knowledge. Finally, the synthetic reference frame is added in the reference picture list for inter-frame prediction in order to remove the global and the local redundancies. Experimental results show that the proposed method outperforms the existing works.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (U1736206), the National Key R&D Program of China (2018YFB1201602), the Natural Science Foundation of Jiangsu Province (BK 20180234), the National Natural Science Foundation of China (61671336, 91738302) and the Hubei Province Technological Innovation Major Project (2017AAA123).

REFERENCES

[1] Xiao, J., Chen, Y., Liao, L., Hu, J., & Hu, R. (2015, April). Global coding of multi-source surveillance video data. In *Data Compression Conference (DCC), 2015* (pp. 33-42). IEEE.

[2] Hakeem, A., Shafique, K., & Shah, M. (2005, November). An object-based video coding framework for video sequences obtained from static cameras. In *Proceedings of the 13th annual ACM international conference on Multimedia* (pp. 608-617). ACM.

[3] Mofaddel, M. A., & Abd-Elhafiez, W. M. (2011, November). Object-based hybrid image and video coding scheme. In *Computer Engineering & Systems (ICCES), 2011 International Conference on* (pp. 245-251). IEEE.

[4] Tsai, T. H., & Lin, C. Y. (2012). Exploring Contextual Redundancy in Improving Object-Based Video Coding for Video Sensor Networks Surveillance. *IEEE Trans. Multimedia*, *14*(3-2), 669-682.

[5] Le Gall, D. (1991). MPEG: A video compression standard for multimedia applications. *Communications of the ACM*, *34*(4), 46-58.

[6] Babu, R. V., & Makur, A. (2006, December). Object-based surveillance video compression using foreground motion compensation. In *Control, Automation, Robotics and Vision, 2006. ICARCV'06. 9th International Conference on* (pp. 1-6). IEEE.

[7] Ng, K. T., Wu, Q., Chan, S. C., & Shum, H. Y. (2010). Object-based coding for plenoptic videos. *IEEE Transactions on Circuits and Systems for Video Technology*, *20*(4), 548-562.

[8] Aizawa, K., Choi, C. S., Harashima, H., & Huang, T. S. (1993). Human facial motion analysis and synthesis with application to model-based coding. In *Motion Analysis and Image Sequence Processing* (pp. 317-348). Springer, Boston, MA.

[9] Gorur, P., & Amrutur, B. (2014). Skip decision and reference frame selection for low-complexity H. 264/AVC surveillance video coding. *IEEE Transactions on Circuits and Systems for Video Technology*, *24*(7), 1156-1169.

[10] Vetro, A., Wiegand, T., & Sullivan, G. J. (2011). Overview of the stereo and multiview video coding extensions of the H. 264/MPEG-4 AVC standard. *Proceedings of the IEEE*, *99*(4), 626-642.

[11] Pan, Z., Zhang, Y., & Kwong, S. (2015). Efficient motion and disparity estimation optimization for low complexity multiview video coding. *IEEE Transactions on Broadcasting*, *61*(2), 166-176.

[12] Tech, G., Chen, Y., Müller, K., Ohm, J. R., Vetro, A., & Wang, Y. K. (2016). Overview of the multiview and 3D extensions of high efficiency video coding. *IEEE Transactions on Circuits and Systems for Video Technology*, *26*(1), 35-49.

[13] Schwarz, H., & Wiegand, T. (2012, May). Inter-view prediction of motion data in multiview video coding. In *Picture Coding Symposium (PCS), 2012* (pp. 101-104). IEEE.

[14] Purica, A. I., Mora, E. G., Pesquet-Popescu, B., Cagnazzo, M., & Ionescu, B. (2016). Multiview plus depth video coding with temporal prediction view synthesis. *IEEE Transactions on Circuits and Systems for Video Technology*, *26*(2), 360-374.

[15] Xiao, J., Hu, R., Liao, L., Chen, Y., Wang, Z., & Xiong, Z. (2016). Knowledge-based coding of objects for multisource surveillance video data. *IEEE Transactions on Multimedia*, *18*(9), 1691-1706.

[16] Bossen, F. (2013). Common test conditions and software reference configurations. *JCTVC-L1100*, *12*.

[17] Bjontegaard, G. (2001). Calculation of average PSNR differences between RD-curves. *VCEG-M33*.

[18] Pavlakos, G., Zhou, X., Chan, A., Derpanis, K. G., & Daniilidis, K. (2017, May). 6-dof object pose from semantic keypoints. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on* (pp. 2011-2018). IEEE.