Performance Evaluation for AVS2 Scene Video Coding Techniques

Siwei Dong, Yonghong Tian, Tiejun Huang
Email: {dosdong, yhtian, tjhuang}@pku.edu.cn
National Engineering Laboratory for Video Technology
School of Electronics Engineering and Computer Science, Peking University, Beijing, China

***Abstract*—Different from conventional video coding techniques, the emerging AVS2 standard (the second generation of Audio Video coding Standard, AVS2 for short) introduces some new coding techniques mainly focuses on video coding on scene videos. Scene videos are videos with limited scenes, such as surveillance videos, conference videos, etc. For scene videos, due to the limited scenes, great redundancy especially in background region is retained in the video pictures. AVS2 scene video coding techniques aim at significantly reducing the redundancy so as to achieve higher compression performance. This paper introduces the main coding techniques in AVS2 and demonstrates the coding performance. Experimental results show that the coding efficiency has almost doubled over the main group on scene videos which brings new opportunities to the related industries on specific applications.**

Keywords*—*Scene videos coding, video coding standard, background modelling, background prediction, AVS2

I. INTRODUCTION

AVS2 leads the video coding efficiency to a new benchmark, doubling that of the first generation of AVS. Similar to traditional coding standard, the AVS2 design follows the classic block-based hybrid video coding approach. However, in order to improve coding efficiency, in the AVS2 coding framework, coding units (CU) are no longer fixed sized which employ the quad-tree structure. At the same time, the prediction units (PU) are not limited to symmetric partition while asymmetric PUs are also available. The sizes of transform units (TU) are independent from PU sizes which can provide more flexible coding processing. Besides, several creative techniques are adopted in AVS2 modules of prediction, transform, entropy coding, etc.

According to the recent research report by IDC [1], half of the video data are surveillance videos in 2010, and the proportion is increasing to about 65% in 2015. More than 5,800 EB of surveillance videos will be produced by 2020. It is quite a big challenge for storage and transmission. Therefore, the huge demand for efficient video compression techniques is rising urgently. There is much room to reduce the redundancy in scene videos including surveillance videos, conference videos, etc. AVS2 with scene video coding techniques is such an emerging standard to fill the gap.

As we know, background regions appear frequently in scene videos which lead to great temporal and spatial redundancy. To compress the background data more efficiently, AVS2 adopts several new techniques including background picture (GB picture) and background-prediction picture (S picture), background modelling and background residual prediction [2]. Fig. 1 describes the video coding architecture.

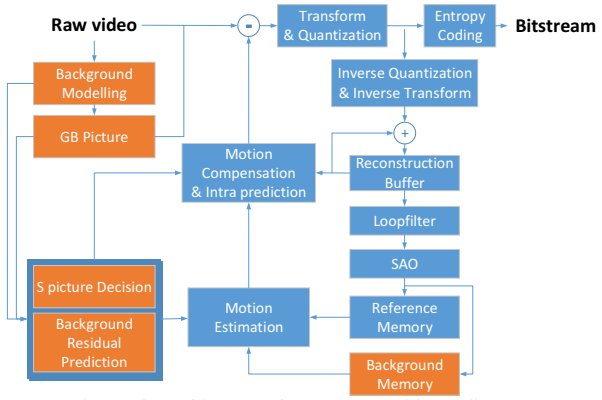

Fig. 1. The architecture of AVS2 scene video coding

This paper introduces and analyses the latest AVS2 scene video coding techniques mentioned above. The rest of the paper is organized as follows. Sec. II introduces the methods of reducing background redundancy. The performance of AVS2 scene video coding is shown with experimental results in Sec. III. Sec. IV concludes the paper.

II. METHODS OF REDUCING BACKGROUND REDUNDANCY

The background regions are dominant in the scenes of surveillance videos and reducing the background redundancy can improve the coding performance efficiently. The long-term reference technique, the background prediction based techniques (GB picture, S picture and background residual prediction) are adopted in AVS2 to reduce the background redundancy [3].

A. The long-term reference technique

The surveillance videos are captured with the fixed cameras and thus the scenes are static with moving foreground objects. There are nearly no changes in the background regions over

978-1-4799-8688-0/15 $31.00 © 2015 IEEE
DOI 10.1109/BigMM.2015.18

time and these regions can be well referenced by following frames.

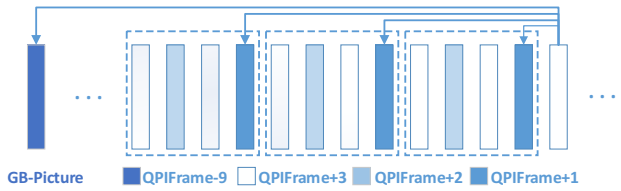

Fig. 2 The long-term reference technique

Traditionally, the current frame can only be inter-predicted by the previous frames in the group of pictures (GOP), which means that the distance between the current picture and the reference frame is restricted in a relative short distance. As a result, the background regions cannot be well utilized as reference frame. Shown in Fig. 2, the background picture is encoded with a relatively small quantization parameter (QP), which can provide better background prediction as a long-term reference. Each subsequent inter-predicted frame utilizes the background picture as one of its references to reduce the background region redundancy significantly. Thus, the long-term reference technique breaks the restriction of the reference distance. If the background regions are encoded with high quality, the corresponding regions can be referenced by the following frames with long distance when adopting the long-term reference technique. As the quality of the background regions is high, the quality of the regions in following frames predicted by them will increase, consequently, the bitrate will decrease. Thus the total performance will increase for the whole surveillance videos.

B. GB picture

To support the long-term reference technique, a new picture type is adopted in AVS2, which is called non-output intra decoding picture (GB picture). The GB picture is encoded by intra mode only but it does not output when displaying. The reason is that at the encoder, the GB picture is a background picture generated by the training frames and thus it is just for being referenced rather than for viewing.

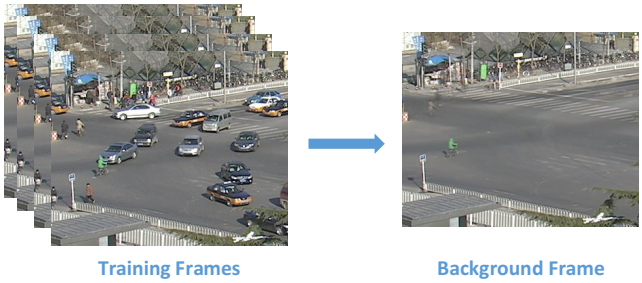

Fig. 3. The training frames and the background frame

The generation of GB picture, is out of the scope of AVS2 standard. Any background modelling method is compatible if the generated background picture is in consistent with the syntax of GB picture. To meet the requirement of real-time transmission and storage for scene videos, the background modelling method in AVS2 should have advantages of low complexity and high efficiency. As a default implementation, in AVS2 Reference Design (RD), a method of segment-and-weight based running average (SWRA) [4] is utilized to generate the GB picture. SWRA approximately generates the background by assigning larger weights on the frequent values in the averaging process. Technologically, SWRA divides the pixels at a position in the training pictures into temporal segments with their own mean values and weights, and then calculates the running and weighted average result on the mean values of the segments. In the process, pixels in the same segment have the same background/foreground property and the long segments have larger weights. Experimental results [4] show that SWRA can achieve good performance yet without suffering a large memory cost and high computational complexity. An example of the constructed background frame and the training frames are shown in Fig. 3.

When encoding the GB picture, a smaller QP is selected to get high quality for GB picture. And the GB picture can be well referenced by the following pictures. Firstly, the GB picture is a background picture where the whole picture is background regions. The background regions of the following pictures can always find the matching regions in the GB picture. Secondly, the GB picture is of high quality and it will provide prediction blocks of high quality and the residuals of the following background regions will be small. Thus the coding performance of the following picture will be better. Although the GB picture takes a lot of bits, the bitrate savings brought by the following frames are much larger than the cost and thus the total performance becomes better.

The GB picture is stored in an independent buffer. For each inter-picture (P picture), the reference picture set consists of past several pictures in neighbourhood and the long-term GB-picture if the P-picture is decided to adopt the long-term reference technique.

The GB picture can be adaptively updated to ensure that it can always provide the best reference. When the background update module is triggered and a new background picture is generated, the current GB picture is about to be replaced by the new one. As the GB picture memory is independent, it is easy to accomplish the replacement without any side effects to other neighbor reference pictures.

C. S picture

Provided with the GB picture and the long-term reference technique, another picture type called S picture is designed for balancing the coding performance and purpose of random access. The S picture only uses the reconstructed GB picture for reference [7] and only intra, SKIP and P2N×2N modes with zero motion vectors are available in S picture. The other inter prediction modes are forbidden since there is no motion in S picture and the other partition modes only cause an increase in the bit cost brought by the partition flag without a more precise prediction.

Traditionally, when achieving random access, the picture at the random access point should be decoded independently from the previous frames. In the AVS2, the S picture is set as the random access point in replace of intra-predicted picture (I

picture). The reason is illustrated as follows. The latest decoded GB picture is provided all the time and when decoding the S picture, the GB picture is firstly got and then the S picture is decoded jointly with the existing GB picture. Besides, the zero motion vectors in the S picture make sure that there is no need in consideration of the motion vector prediction (MVP) and thus even the temporal MVP is adopted, the S picture can still be correctly decoded. The S picture can be decoded independently from the previous frames in the GOP structure and thus it can be set as the random access point.

The S picture outperforms I picture when adopted as the random access point since the inter prediction is adopted in the S picture and the prediction performance is better. With the S picture, the performance of the random access can be improved on a large scale.

D. Background residual prediction

Based on the background picture, some coding units (CU) contains both foreground pixels and background pixels, which are called hybrid coding units (HCU). For HCU, the usual way of motion estimation is not that perfect. Due to the stationary background in scene videos, the background region in HCU is more likely to be matched in the similar position of GB picture in the reference picture set. Meanwhile, the foreground region matches better in the other reference pictures.

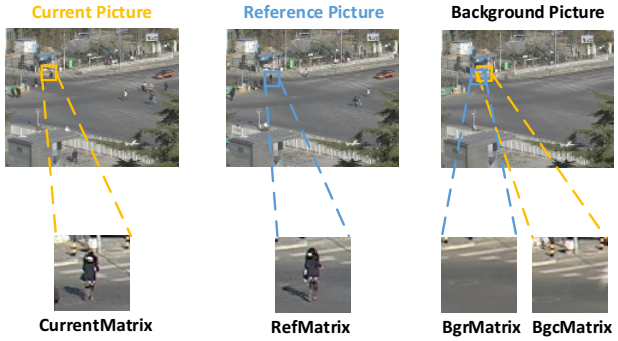

Fig. 4. Background residual prediction

In short, the traditional prediction is not efficient enough for HCU. In Fig. 4, the *CurrentMatrix* represents the pixel matrix of a HCU which contains a pedestrian. The region of the foreground object (we mean the pedestrian here) can be well matched in the *RefMatrix*. However, the background region (the zebra crossing) is not the perfect match if we simply utilize the selected reference picture for prediction. Considering the spatial variation, for the background region, the spatial corresponding matrix in background picture, i.e. the *BgcMatrix* will provide a better prediction rather than the *RefMatrix*. At the same time, the *BgrMatrix* in background picture is the spatial corresponding matrix of the *RefMatrix*. Therefore, we can use the different pixel matrices to acquire more accurate prediction, the *RefMatrix-BgrMatrix* for foreground prediction while the *BgcMatrix* for background prediction. The technique described above is called the background residual prediction.

```
for each pixel in the matrix
  if(abs(RefMatrix[x,y]-BgrMatrix[x,y]) <= BgDiffPredThreshold)
    PredMatrix[x,y] =
        RefMatrix[x,y]-BgrMatrix[x,y]+BgcMatrix[x,y]
  else
    predMatrix[x,y] = RefMatrix[x,y]
```

Fig. 5. The algorithm of background residual prediction

Fig. 5 describes the algorithm of background residual prediction. Given a CU, if the pixel differences between the *RefMatrix* and the *BgrMatrix* exceed the threshold, the background residual prediction technique is available and the predicted matrix of current CU is constructed via *RefMatrix-BgrMatrix+BgcMatrix*. In order to obtain better prediction, the technique of background residual prediction is adopted by AVS2. A flag named *cu_bdp_flag* indicates whether current CU utilizes background residual prediction.

III. PERFORMANCE OF AVS2 SCENE VIDEO CODING

A. Experiment Setup

To evaluate the scene video compression performance more extensively, five typical scene videos are selected as the common test sequences, including three SD ones with resolution of 720×576 and two HD ones with resolution of 1600×1200, as shown in Fig. 6 [5]. Crossroad is a three-way junction with lots of people crossing the road. In the middle part of the video, when the traffic light turns green, cars and buses start to move. Office describes an indoor scene, in which the staffs occupy most of the camera field of view. Overbridge is shot in a snowy morning in which pedestrians are walking on the bridge while buses are running under it. Intersection and Mainroad are captured by traffic cameras with resolution of 1600×1200 during rush hours.

TABLE I
THE COMMON TEST CONDITIONS OF AVS2 SCENE VIDEO CODING

Parameter	LDP	RAB	RAP
QPIFrame	27, 32, 38, 45		
QPPFrame	QPIFrame+1		
QPBFrame	-	QPIFrame+4	-
SeqHeaderPeriod	0	1	1
IntraPeriod	0	32	32
NumberBFrames	0	7	0
FrameSkip	0	7	0
BackgroundQP	QPIFrame-9		
BackgroundEnable	1		
ModelNumber	120		

The common test conditions [6] define a set of encoder configurations used in experiments. These configurations include the following:

1) Low delay with P slices (LDP)
2) Random Access with B slices (RAB)
3) Random Access with P slices (RAP)

Table I shows the common test conditions of AVS2 scene video coding. The *BackgroundQP* for GB picture is equal to

that of I picture minus 9. The training picture number of background modelling is 120.

We conduct the experiments on RD 7.0, which is the latest released reference software for AVS2. The coding performance is evaluated on two sets of experiments, one of which keeps the scene video coding techniques enabled (RD 7.0 Scene) and the other is kept disabled (RD 7.0 General).

Resolution	Sequence	Scene Shot
720×576	Crossroad	
	Office	
	Overbridge	
1600×1200	Intersection	
	Mainroad	

Fig. 6. The common test sequences for scene video coding in AVS2

B. *Experimental Results*

Table II shows the performance between RD 7.0 Scene and RD 7.0 General. According to the experimental result, RD 7.0 Scene reduces 27.0% (LDP), 52.3% (RAB) and 46.4% (RAP) bitrates in average against RD 7.0 General in 720×576 videos and 40.8% (LDP), 53.1% (RAB) and 49.0% (RAP) on 1600×1200 videos. Among the video sequences, Office has large foreground objects and it is hard to generate clear background picture, so the coding performance is relatively lower than others. In average, RD 7.0 Scene can obtain 32.5% (LDP), 52.6% (RAB) and 47.4% (RAP) bitrate savings on all common test sequences.

TABLE II
THE PERFORMANCE COMPARISON BETWEEN RD 7.0 SCENE AND RD 7.0 GENERAL

Resolution	Sequence	RD 7.0 Scene vs. RD 7.0 General (BD-Rate)		
		LDP	RAB	RAP
720×576	Crossroad	-27.6%	-50.5%	-42.8%
	Office	-14.9%	-34.8%	-28.9%
	Overbridge	-38.5%	-71.6%	-67.4%
	Average	**-27.0%**	**-52.3%**	**-46.4%**
1600×1200	Intersection	-20.5%	-33.2%	-31.1%
	Mainroad	-61.1%	-72.9%	-66.9%
	Average	**-40.8%**	**-53.1%**	**-49.0%**
All	**Average**	**-32.5%**	**-52.6%**	**-47.4%**

IV. CONCLUSIONS

Due to limited scenes, there is huge redundancy in scene videos. Most of traditional video coding methods are not able to remove this kind of redundancy well. AVS2 is the latest coding standard with efficient scene video coding techniques aiming at high efficiency video coding for scene videos. This paper introduces and analyses the major representative techniques, including the long-term reference technique and the background prediction based techniques (GB picture, S picture and background residual prediction).

With background prediction based coding methods to reduce the scene redundancy, AVS2 can achieve 32.5% (LDP), 52.6% (RAB) and 47.4% (RAP) coding efficiency in average on scene videos. Along with the development of the Internet of Things and the rapid construction of smart city, a growing number of scene videos will be captured. Undoubtedly, AVS2 is a perfect solution to the massive video compression, which as well brings in fresh perspectives to related research and applications.

ACKNOWLEDGMENT

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

REFERENCES

[1] IDC, “IDC Digital Universe study: Big Data, Bigger Digital Shadows & Biggest Growth in the Far East”, Dec. 2012.

[2] L. Zhao, S. Dong, P. Xing and X. Zhang, “AVS2 surveillance video coding platform”, in AVS M3221, Dec. 2013.

[3] F. Liang, “Information Technology – Advanced Media Coding Part2: Video (CD)”, in AVS N2046, May 2014.

[4] X. Zhang, Y. Tian, T. Huang, W. Gao, “Low-complexity and High-efficiency Background Modelling for Surveillance Video Coding,” Proc. IEEE Int’l Conf. Visual Communication and Image Processing, Nov 2012.

[5] S. Dong, L. Zhao, “AVS2 surveillance test sequences”, in AVS M3168, Sept. 2013.

[6] S. Dong, “Common test conditions of AVS2-P2 surveillance profile”, in AVS N2021, Jan. 2014.

[7] R. Wang, Z. Ren, H. Wang, “Background-predictive picture for video coding,” in AVS M2189, Dec. 2007