Content-adaptive Traffic Surveillance Video Coding with Extended Spatial Scalability

Yunpeng Liu, Renfang Wang, Dechao Sun, Shijie Yao, Nayi Hong, Peng Jin
Zhejiang Wanli University/ College of Computer Science and Information Technology, Ningbo, China
Email: L35633@163.com

***Abstract*—Regions of interest (ROI) or visually salient regions are rarely considered in spatial scalable video coding, thus visually important content can not be better adapted to lower display resolutions. In this paper, we propose a content-adaptive spatial scalable coding for traffic surveillance video. First, the background image is extracted by an improved single Gaussian method based on the spatio-temporal model and updated from the latest static image. Then a background subtraction algorithm is present for detecting and tracking vehicles, the motion window of the leading vehicle is commonly referred to as ROI in traffic surveillance, and ROI is as a cropping window in extended spatial scalability (ESS) of the scalable video coding (SVC). Moreover, we employ a tracking-aware compression algorithm to remove more low tracking interest bit rate by ROI-based quantization strategy and frequency coefficient suppression technique, so tracking accuracy is used instead of PSNR as the compression criterion. The experimental results show that compared with conventional scaling coding the proposed algorithm can greatly improve the visual perception of the decoded base layer video with limited loss in the rate-distortion performance, and allows for about 60% bit rate savings while maintaining comparable tracking accuracy.**

***Index Terms*—scalable video coding, extended spatial scalability, traffic surveillance, content-adaptation**

I. Introduction

Scalable video coding (SVC) is highly desirable for surveillance applications, in which video sources not only need to be viewed on multiple devices ranging from high-definition monitors to videophones or PDAs, but also need to be stored and archived [1]. Spatial scalability (SS) is defined via linear scaling operations between different resolutions. Changes of aspect ratios are possible via different scaling in both dimensions. Multiple-resolution videos can be compressed into one bit stream [2-4]. And extended spatial scalability (ESS) of SVC can support cropping and non-dyadic scaling between different aspect ratios [5].

Video retargeting is another technology that is receiving a lot of attention. A given video is adapted to a different target resolution and an aspect ratio, by involving non-linear scaling operations. This is done in a way that visually important content is preserved while distortions are hidden in visually less important areas, in which warping [6] and seam carving [7] techniques get more study and attention for better performance and practicality.

Reference [8-10] integrate the non-linear warping [6] operations into spatial scalability framework, which includes two new building blocks, i.e. non-linear warping prediction and warp coding. For generic audiovisual services and broadcast television applications, the method can address the content adaptation of low-resolution terminals with a limited increase in the bit rate and complexity. But different from the generic audiovisual services, the surveillance video commonly has a long-term static background image, and a different type of surveillance has different monitoring objects with the specific video features. Hence, content-adaptive spatial scalability of a surveillance video has a certain degree of particularity.

In this paper, our study refers to traffic surveillance videos. In general, people are only interested in the leading vehicles (i.e. visually salient vehicles) in the valid monitor area. The motion window (i.e. the minimum window containing vehicles) of the leading vehicles is considered as the visually important content in the monitor screen and as the cropping window. Hence, content-adaptive retargeting of the existing high resolution traffic video is achieved only by cropping and scaling operation without the complex non-linear warping or seam carving. The downsampling can be operated in real time because of the low computational complexity, and spatial scalable coding fully complies with ESS profile of SVC standard.

The rest of this paper is organized as follows: Section 2 introduces the whole coding system structure. In Section 3, locating the leading vehicles algorithm is presented. Section 4 describes regions of interest (ROI) - based coding in enhancement spatial layer to improve the rate-distortion performance. The experimental results are provided in Section 5. Finally Section 6 concludes the paper and gives suggestions for future work.

II. System Coding Structure

Fig. 1 illustrates the system coding structure. Compared with the conventional spatial scalable encoder,

Manuscript received December 1, 2013; revised January 1, 2014; accepted March 1, 2014.
Corresponding author, Yunpeng Liu.

doi:10.4304/jcp.9.11.2595-2602

traffic videos analysis module is added, which is to extract a visually important area containing the leading vehicles, and to output the position and size of cropping window (see details in Section 3). Moreover, ROI-based quantization and frequency coefficient suppression modules are added to improve the rate-distortion performance of enhancement spatial layer (see details in Section 4).

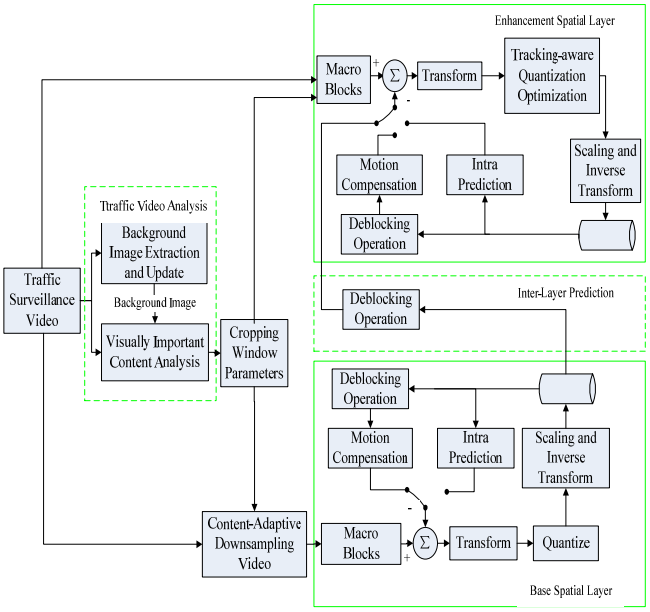


Figure 1. Content-adaptive spatial scalable coding structure.

III. CONTENT-ADAPTIVE SPATIAL SCALABILITY

The high resolution source video downsamples the lower resolution videos by two methods as illustrated in Fig. 2. One is the conventional scaling; the other is the content-adaptive cropping and non-dyadic scaling by ESS. The relations between enhancement and base layers in ESS are illustrated in Fig. 3 [5], where the cropping window is not necessarily aligned with the pictures macroblock structure (i.e. cropping origin (x_0,y_0) is not positioned on the top left pixel of a macroblock). Moreover, cropping and scaling parameters may vary for each picture (picture-level ESS), which is desirable for the base layer pictures focusing on the main ROI of enhancement pictures. The main ROI of traffic videos (i.e. the leading vehicles) are located by the following steps: the background image extraction and update, detecting and tracking vehicles, locating the leading vehicles.

We use the improved single Gaussian model [11] to model the background for surveillance videos, which can attain the ideal background image even when more cars are parked or driving slowly. A region-based vehicles tracking algorithm has a better real-time and anti-noise performance. Reference [12] gives an effective method to segment more vehicles with occlusion, which is used in our paper for tracking vehicles.

Vehicles go into the monitor screen area until they exit illustrated in Fig. 4. It can be seen that there is only one vehicle from frame i to i+18, hence the motion window of the vehicle is ROI as a cropping window. From frame i+19 to i+39, other vehicles appear, but the black one has a much bigger motion window than the others, it is still the leading vehicle. The downsampling results of ROI from frame i to i+39 are shown in Fig.5.

Figure 2. Downsampling comparison of two methods.

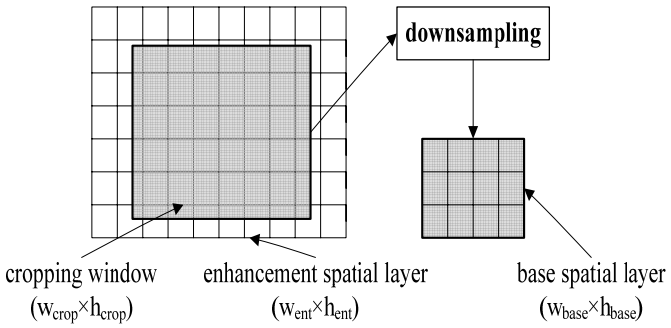


Figure 3. Relations between enhancement and base layers.

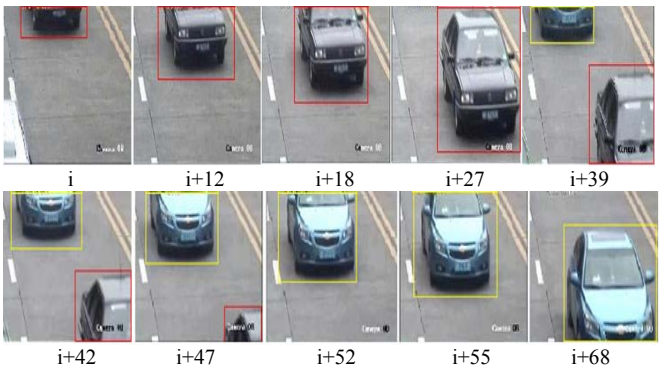


Figure. 4 Tracking vehicles.

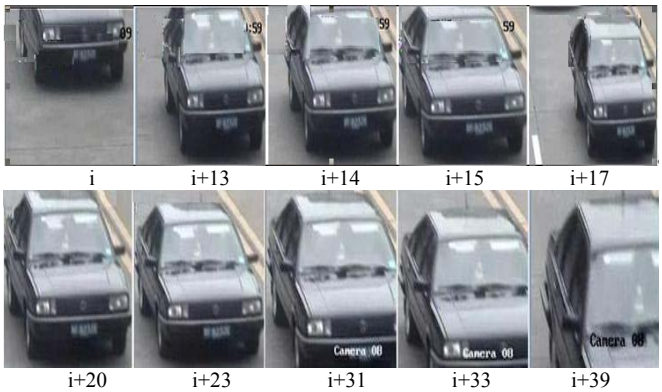


Figure. 5 Downsampling results of ROI.

The different downsampling deformation of the same vehicle results from the different size of the motion window and inaccurate vehicle detection. Hence the

motion window of the same vehicle shall remain the same size to reduce the differences of deformation. Our method is to choose the maximal motion window of the same leading vehicle as standard; others extend to the standard size. And the detailed algorithm is shown below.

Step1. Two important data structures *VehicleList* and *TraceList* are defined to effectively implement this algorithm. *VehicleList* denotes the list of the all traced vehicles in each frame, which is implemented and updated by the tracking algorithm. *TraceList* is used to save *VehicleList* in each read-in frame. When one vehicle disappears from the monitor screen, the detection whether this vehicle is the leading one during the monitoring interval begins. After the leading vehicle is detected in one frame, the corresponding *VehicleList* of the frame is deleted from *TraceList*.

Step2. Judge when the vehicle is not in the monitoring scope.

$$Vehicle\begin{cases}\text{Out MonitorScope,}\\ \quad \text{if } Vehicle \notin TraceList[current+Length]\\ \text{In MonitorScope,else.}\end{cases} \quad (1)$$

where *Vehicle* denotes the tracking one, *current* denotes the current frame, *Length* denotes the continuous detected frames behind *i* frame. For example, when *Length* is equal to 3, *VehicleList* in the continuous 3 frames behind position of *current*=i+48 in Fig. 4 are detected, because the dark vehicle cannot be detected, it is regarded as not in the monitoring scope. After one moving vehicle disappears in the monitor, it will be judged whether the vehicle is the leading one by the below steps.

Step3. Judge the leading vehicle.

$$Vehicle \in \begin{cases}\text{Leading Vehicle,}\\ \quad \text{if } Count(i,current,Vehicle,VehicleList)=1,\\ \quad or\ Area(Vehicle) >> Area(other)\\ \text{None Leading Vehicle,else.}\end{cases} \quad (2)$$

where *i* denotes the position when the vehicle appears, *Count* is the function which detects the number of tracking vehicles. Still taking Fig. 4 for instance, check *VehicleList* in each frame from i to i+47 frame, the dark vehicle which drives off the monitor discussed in Step 2 only exists from i to i+18. So the dark one is regarded as the leading vehicle. *Area* is the function which calculates the area of moving window. >> denotes much greater than. For example, the other vehicle appears from i+19 to i+39 frames in the monitor, but the dark vehicle is still the leading one because of the much bigger moving window area.

Step 4. Find the biggest moving window to which other windows extend, and attain the frame number, the moving window position and size etc., which are used to downsample.

Step5. If the tracking vehicle is almost the same size as others, i.e. no leading vehicle exists. Then the traditional downsampling method is used, the whole monitor screen is ROI.

Step6. The process is ignored if the tracking vehicle is not the leading one.

Step7. The *VehicleList* of the tracking vehicle is deleted from *TraceList*.

Step8. Loop detecting next leading vehicle by the same steps.

IV. Tracking-optimal Compression

Base spatial layer video coming from cropping and scaling will cause the inter-frame correlation to become weaker, hence the rate-distortion performance of base layer turns worse. For enhancement spatial layer, the prediction mode commonly is the inter-frame and inter-layer adaptive, i.e. choosing the optimal coding mode between the inter-frame and inter-layer predictions. There is no change in the inter-frame prediction of enhancement layer, but for the inter-layer prediction, the prediction performance is worse due to the weaker inter-layer correlation and the worse rate-distortion performance in base layer. Therefore, the rate-distortion performance of the whole coding system deteriorates.

Visual perception based ROI coding can improve the rate-distortion performance of ROI and retain the subjective quality of the decoded video. The perception of Human Visual System (HVS) for the video scene is selective, and different regions or objects in the video scene have diverse levels of visual importance. According to the features of traffic surveillance videos, the visual perception analysis diagram is illustrated in Fig. 6.

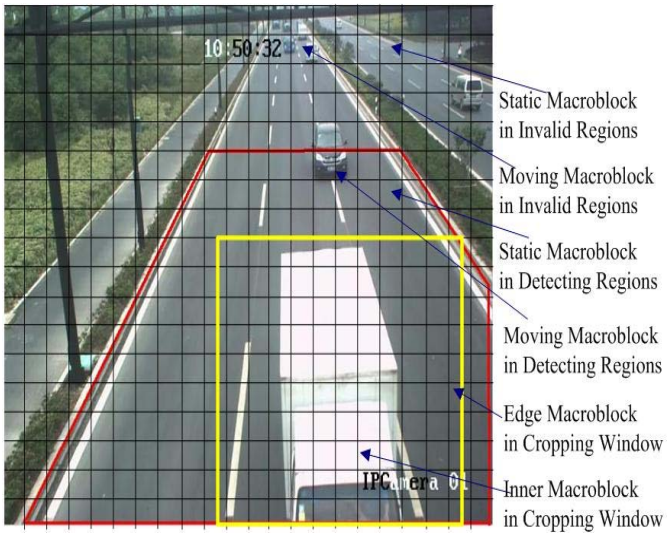


Figure. 6 Visual perception analysis of traffic image

The degree of importance in turn increases from "Static Macroblock in Invalid Region" to "Inner Macroblock in Cropping Window" in Fig.6. The coding quality is controlled by allocating more bit resources to the visually important macroblocks and less to visually less important macroblocks. Two methods are often used to control the bit resource allocation. One is quantization parameter (QP) control, i.e. the more important the macroblocks are, the less QP is; the other is frequency coefficient suppression control, Fig.7 illustrates 18 different frequency coefficient suppression matrixes

(abbreviation SM). It is easy to know that the more important the macroblocks are, the more the number of 1 in SM is.

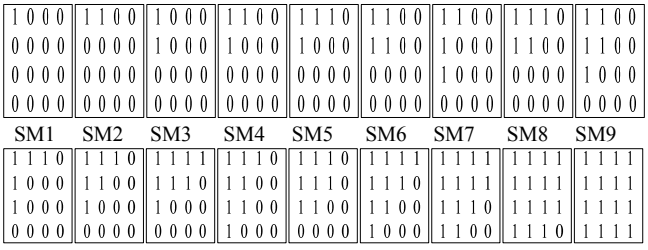


Figure. 7 Frequency coefficient suppression matrix

Whether QP or SM control is used, only very limited rate-distortion performance is improved, because the bit rate reduction will inevitably bring a certain loss in quality. Given that the main application of traffic surveillance, is the high-level semantic analysis based on the low-level object detection and tracking, we use tracking accuracy instead of PSNR as the compression criterion, the compression distortion of tracking and quantization optimization are as follows.

A. Compression Distortion of Tracing

The surveillance video performance metrics are reviewed in [13], of which the overlap (OLAP), precision (PREC) and sensitivity (SENS) metrics are chosen, with the ground truth defined as the tracking results generated using the uncompressed video.

$$OLAP = 1/N \sum_{i=1}^{N} (GT_i \cap AR_i)/(GT_i \cup AR_i) \quad (3)$$

$$PREC = TP/(TP + FP) \quad (4)$$

$$SENS = TP/(TP + FN) \quad (5)$$

where GTi denotes the i-th object tracked in the uncompressed video, AR_i the i-th object tracked in the compressed video. TP, FP, FN respectively denotes the number of true positives, false positives, and false negatives. A true positive results from an object being present in both the GT and AR. A false positive results from an object being present in the AR but not in the GT, or if an object detected in the AR does not overlap an equivalent object in the GT. A false negative results from an object being present in the GT but not in the AR, or if an object detected in the GT does not overlap an equivalent object in the AR. The final accuracy A is defined as

$$A = (\alpha * OLAP) + (\beta * PREC) + (\gamma * SENS) \quad (6)$$

where α, β and γ are weighting coefficients, and $\alpha + \beta + \gamma = 1$, here ($\alpha$, β, γ) is (1/3, 1/3, 1/3).

B. Quantization Optimization

This process is to remove more low tracking interest bit rate by the ROI-based quantization strategy and frequency coefficient suppression technique. The detailed algorithm is as follows by modifying [14]:

Step1. QP initialization for different type of macroblock by

$$MB\_QP = \begin{cases} q + x, \ if \ \ current\ image\ is\ background\ or \\ MB \in invalid\ \ surveillance\ regions \\ q, \quad otherwise \end{cases} \quad (7)$$

where q is the input QP value. If x is too big, the inter prediction deteriorates more, if too small, no more useless bits are removed, here the experience value of x is 6. Valid surveillance regions are those of traffic surveillance interest such as roads or sidewalks, which would concentrate on tracking vehicles and pedestrians as opposed to other objects such as trees or clouds which are named invalid surveillance regions. Then the further quantization optimization is processed for macroblocks in valid surveillance regions of non-background.

Step2. Frequency coefficient suppression initialization by

$$SMA = [SM_1, SM_2, ..., SM_{18}], \ SMC = SM_1 = SMA[i], \quad i = 0 \quad (8)$$

where SMA is the array of SMs as shown in Fig.7, SMC represents the current chosen SM, the array index i is initiated as 0, i.e. the tracking measure begins from the least bit rate and the SMC is used for all macroblocks in valid regions.

Step3. The tracked sample video clips set initialization by

$$S = [S_1, S_2, ..., S_n], \quad Length(S_k) = len_k, \quad 1 \le k \le n \quad (9)$$

where S_k is the tracked sample video clip whose buffer size is decided by different delay requirements and encoding applications, commonly the buffer size cannot be too big. Using the set S instead of the single S_k is to enhance the accuracy by using the mean value of tracking metrics, but more clips will add more computational complexity too. Therefore, we choose two clips with a length respectively 5 and 10 based on experience, and then the resulting compressed bit rate and tracking accuracy are computed by

$$S = [S_1, S_2], \quad Length(S_1) = 5 \ and \ Length(S_2) = 10$$
$$(R_k, A_k) = f_{encode}(q, SMC, S_k), \quad k = 1, 2$$
$$(R, A) = ((R_1 + R_2)/2, (A_1 + A_2)/2) \quad (10)$$

Step4. After SMA index adding one i.e. i = i + 1 and getting next SM i.e. SMC = SMA[i], the (R_i, A_i) is computed by formula (10) again. In general, R_i will be greater than R_{i-1}, if $A_i <= A_{i-1} * p$ and $A_{i+1} <= A_i * p$, the loop is over, otherwise get the next index i and compute (R_i, A_i) until the end of SMA, here $0.95 <= p <= 1$. Because the noise is also more filtered when more frequency coefficients are removed, the tracking accuracy maybe is better than before. Hence, continuous A is compared twice to ensure the accuracy of the results. From these results, only those frequencies which provide the highest tradeoff of bits for tracking accuracy are kept.

V. EXPERIMENTAL RESULTS

A. Coding Efficiency

The tested traffic videos come from real traffic surveillance applications at home and abroad, and have

different characteristics such as the viewing angle, video quality and type of vehicle traffic observed. The details are as follows.

1) The *Close* sequence with resolution 720*480, shot in an American traffic application, showing a close main road with a normal traffic flow. The camera jitter is a problem when vehicles are passing. The video has little illumination change and little acquisition noise.

2) The *Intersection* sequence with resolution 352*288, shot in a Chinese traffic application, showing an urban busy intersection with steady traffic interrupted by a traffic signal. The video has some illumination changes and some acquisition noise.

3) The *Far* sequence with resolution 640*480, shot in a Chinese traffic application, showing a far highway with light traffic. The video has significant global illumination changes, acquisition noise and a slight camera jitter.

The experiments use the desktop computer with Pentium (R) Dual-Core 2.5GHz CPU, 2G memory and WindowXP operation system. The experimental parameters are configured as shown below. The encoder and decoder are JSVM9.19, frames are 1500, GOP Size is 8, intra period is -1, inter layer prediction mode is 2. For *Close*, *Intersection* and *Far* sequences, the enhancement and base layer format respectively is (720*480, 352*288), (352*288, 176*144) and (640*480, 320*240), of which the spatial scalability of *close* sequence is non-dyadic, *Intersection* and *Far* sequences are dyadic. The algorithm in [12] is used as the object tracker for the spatial enhancement layer. Input QPs of base layer and enhancement layer are (22, 24), (26, 28), (30, 32) and (34, 36).

First use the traditional method of rate-distortion metrics. Different encoding strategies are selected as shown in Table I. Standard two layer downsampling scalable coding scheme is referred to SD; content adaptive downsampling is referred to CE; CE_Q denotes that ROI coding is controlled by QP; CE_C denotes that ROI coding is controlled by transform coefficients; macroblock in the clipping window is CI, in the edge is ME; moving macroblock in the detection areas is DM, static macroblock in the detection areas is DS; moving macroblock in the invalid areas is IM, static macroblock in the invalid areas is IS. QPE is enhancement layer quantization parameter, SMn is referred by Fig. 7.

The comparisons of PSNR and bitrate in different coding strategies for QP (22, 24) is shown in Table II-V. The rate-distortion curves of ROI are shown in Fig. 8. The rate and tracking accuracy comparison curves of different sequences are illustrated as Fig.9.

TABLE I.
CODING STRATEGY CONFIGURATION

Coding Strategy	Different Macroblock Type					
	CI	ME	DM	DS	IM	IS
CE_Q	QPE-4	QPE	QPE+4	QPE+8	QPE+12	QPE +16
CE_C	SM18	SM15	SM 12	SM9	SM6	SM3

TABLE II
COMPARISON OF PSNR AND BITRATE (CLOSE SEQUENCE)

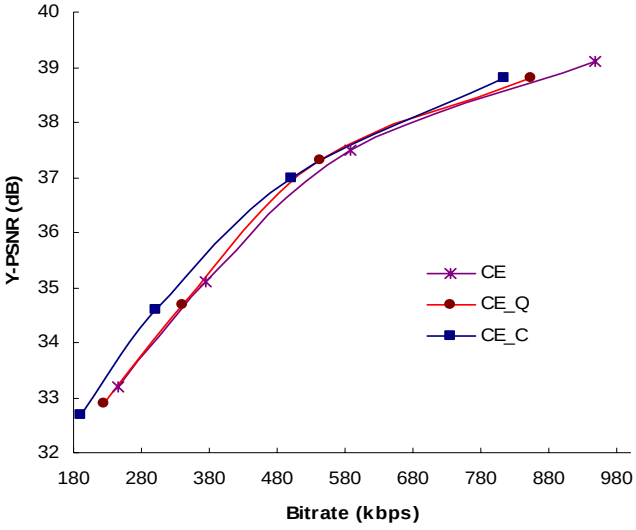
Coding Strategy	Close Sequence		
	Frame Average PSNR(dB)	ROI Average PSNR(dB)	Bitrate(kbps)
SD_1	42.8	39.1	794.4
SD_2	42.8	39.1	795.5
CE_1	42.9	39.0	946.8
CE_2	42.9	39.1	940.6
CE_Q_1	40.8	38.5	869.8
CE_Q_2	41.4	38.6	832.1
CE_C_1	41.7	38.6	870.1
CE_C_2	42.0	38.8	812.4

TABLE III
COMPARISON OF PSNR AND BITRATE (INTERSECTION SEQUENCE)

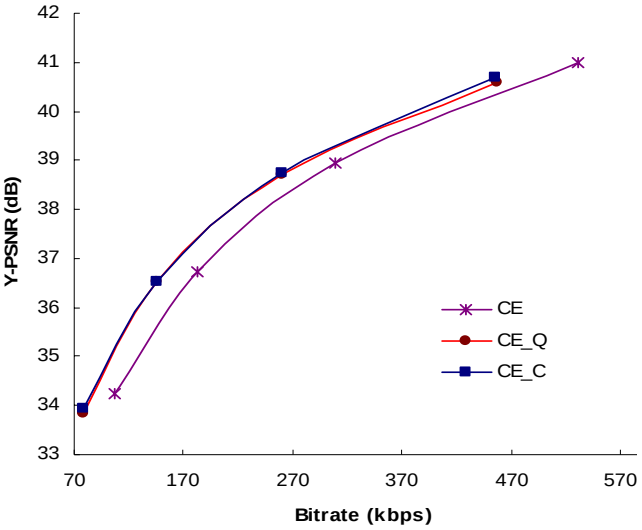
Coding Strategy	Intersection Sequence		
	Frame Average PSNR(dB)	ROI Average PSNR(dB)	Bitrate(kbps)
SD_1	44.1	41.0	455.7
SD_2	44.1	41.0	454.1
CE_1	44.1	41.0	530.6
CE_2	44.2	41.0	520.9
CE_Q_1	41.6	40.4	474.0
CE_Q_2	42.0	40.6	464.3
CE_C_1	42.0	40.6	468.3
CE_C_2	42.2	40.7	460.1

TABLE IV
COMPARISON OF PSNR AND BITRATE (FAR SEQUENCE)

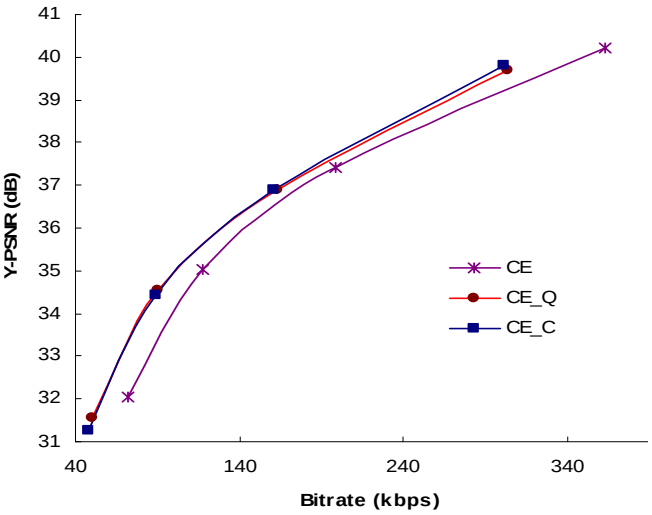
Coding Strategy	Far Sequence		
	Frame Average PSNR(dB)	ROI Average PSNR(dB)	Bitrate(kbps)
SD_1	43.8	40.3	297.2
SD_2	43.8	40.2	296.4
CE_1	43.9	40.2	362.9
CE_2	43.9	40.2	355.7
CE_Q_1	42.4	39.6	314.7
CE_Q_2	42.5	39.7	315.6
CE_C_1	42.6	39.9	305.6
CE_C_2	42.6	39.9	300.1



(a) Close sequence

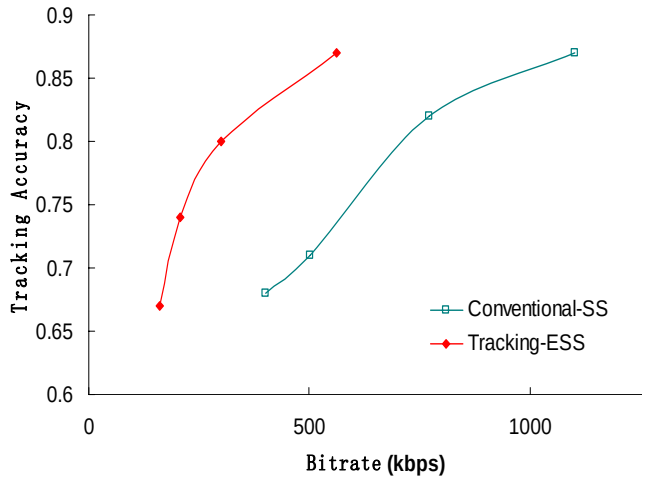


(b) Intersection sequence

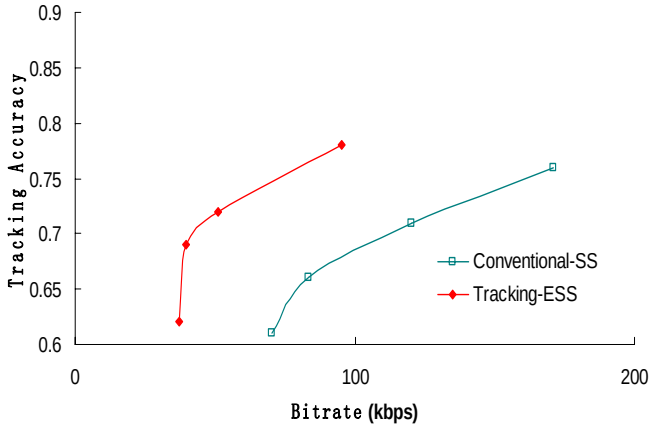


(c) Far sequence

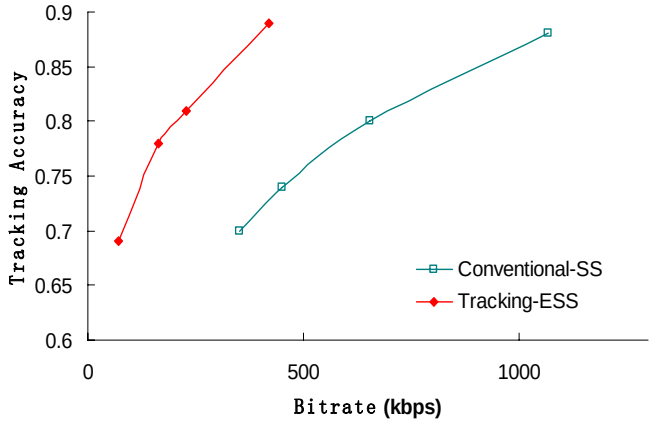
Figure. 8 Rate-distortion curves



(a) Close sequence



(b) Intersection sequence



(c) Far sequence

Figure. 9 Rate-tracking accuracy comparison curves

From Table II - IV and Fig. 8, there is a certain coding efficiency improvement, but not very significant. For rate tracking, as seen in Fig.9, the coding efficiency of *Far* sequence is best, the bit rate reduction is up to about 80% at the same tracking accuracy, and there are at least about 60% bit rate reduction. For *Close* and *Intersection* sequence, the average bit rate reduction respectively is about 60% and 50%. For all sequences, the average reduction is about 60%. The *Far* sequence also has the highest tracking accuracy at the same bit rate, this shows that as global illumination changes, acquisition noise and camera jitter will not affect the coding efficiency, light traffic brings better tracking accuracy and the compression ratio. This is because there are more background images in a light traffic scene, and invalid regions in far- distance surveillance are relatively bigger than close surveillance, we give bigger QP for background images and MBs in invalid regions in our paper, which improves the compression ratio. Moreover, there are more low tracking interest bits and more single vehicles in light traffic, which can improve the compression ratio and tracking accuracy.

B. Computation Complexity

The downsampling video comes from the original high resolution sequence. For content-adaptive ESS, not only the cropping and non-dyadic scaling are included, but

also background modeling and vehicles tracking of the spatial enhancement layer are included. However, this essentially has no effect on the overall computation complexity during downsampling as shown in Table V. The tested frames are 1500, the computer and system configurations are the same as those in Section 5.1, the downsampling and cropping oprerations use "Resampler" function in JSVM9.19. We can see that extra computation complexity introduced by the background modeling and vehicles tracking is very limited for different traffic video sequences, the real-time is maintained. On the one hand, the algorithms used in our paper are low-complexity; on the other hand, traffic videos always have a certain proportion of background images which reduce the computation of tracking vehicles.

TABLE V
COMPARISON OF AVERAGE DOWNSAMPLING TIME FOR EACH FRAME

Sequence Type	Conventional Downsample	Content-adaptive Downsample
Close (720*480, 352*288)	92 ms/frame	110ms/frame
Intersection (352*288, 176*144)	22ms/frame	35ms/frame
Far (640*480, 320*240)	64 ms/frame	80 ms/frame

In the encoding process of content-adaptive ESS, the iterative quantization optimization (IQO) is added which in fact consumes little time as shown in Table VI. Input QP is (26, 28), the tested frames are 100, and the codec and system configuration don't change.

TABLE VI
COMPARISON OF ENCODING TIME

Sequence Type	Conventional	Content-adaptive ESS		
		Total	IQO	IQO/Total
Close sequence (720*480, 352*288),	5075.4s	5113.5s	110.0s	2.15%
Intersection sequence (352*288, 176*144)	2080.7s	2250.1s	69.7s	3.11%
Far sequence (640*480, 320*240)	3560.7s	3588.1s	65.0s	1.81%

VI. CONCLUSION

This paper presents a content-adaptive spatial scalable coding for traffic surveillance videos, which improves the visual perception of decoded base layer videos. Moreover, we employ a tracking-aware compression algorithm to remove more low tracking interest bit rate by ROI-based quantization strategy and frequency coefficient suppression technique which allows for about 60% bit rate savings while maintaining comparable tracking accuracy. But the position information of the main object in ROI and the part background scene will be lost in base layer. This problem can be solved to some extent if combined with warping and seam carving. Under the proposed algorithm, users can real-time adjust the encoding parameters to decide conventional spatial scalability or content-adaptive, which greatly enhances the flexibility of spatial scalable coding, and better addresses the diverse application requirements of low-resolution terminals. The future research will focus on the allocation of computation resources and improvement in real-time encoding, as well as combing with video retargeting of warping and seam carving

ACKNOWLEDGMENT

The work is supported in part by the NSFC (Grant No. 61073074, 61303144), Ningbo Natural Science Foundation (Grant No. 2013A610069), Zhejiang Science and Technology Program (Grant No. 2012C21004), Projects in Science and Technique of Ningbo Municipal (Grant No. 2012B82003), Zhejiang Higher Education Reform Project (Grant No. jg2013135).

REFERENCES.

[1] H Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Trans. Circuits Syst. Video Technol*., vol. 5, no. 9, pp. 1103–1120, 2007.

[2] S C. Andrew and G. J. Sullivan, "Spatial scalability within the H.264/AVC scalable," *IEEE Trans. Circuits Syst. Video Technol*., vol. 7, no. 9, pp. 1121–1135, 2007.

[3] L Xin and G. R. Martin, "A hierarchical mode decision scheme for fast implementation of spatially scalable video coding," *in: Proc. of IEEE Visual Communications and Image Processing*, San Diego, CA, United states, November 2012.

[4] H Tobias, H. Philipp, and L. Haricharan, "An HEVC extension for spatial and quality scalable video coding," *in: Proc. of SPIE - The International Society for Optical Engineering*, Burlingame, CA, United states, February 2013.

[5] E Franois and J. Vieron, "Extended spatial scalability: a generalization of spatial scalability for non-dyadic configurations," *in: Proc. of International Conference on Image Processing*, Atlanta, GA, USA, October 2006.

[6] S S. Lin, I. C. Yeh, C. H. Lin, and T. Y. Lee, "Patch-based image warping for content-aware retargeting," *IEEE Trans. on Multimedia*, vol. 15, no. 2, pp. 359–368, 2013.

[7] B Yan, K. Sun, and L. Liu, "Matching-area-based seam carving for video retargeting," *IEEE Trans. Circuits Syst. Video Technol*., vol. 23, no. 2, pp. 302–310, 2013.

[8] A Smolic, Y. Wang, N. Stefanoski, M. Lang, A. Hornung, and M. Gross, "Non-linear warping and warp coding for content-adaptive prediction in advanced video coding applications," *in: Proc.of International Conference on Image Processing*, Hong Kong, China, September 2010.

[9] Y Wang, N. Stefanoski, X. Fang, and A. Smolic, "Content-adaptive spatial scalability for scalable video coding," *in: Proc. of Picture Coding Symposium*, Nagoya, Japan, December 2010.

[10] Y Wang, N. Stefanoski, M. Lang, A. Hornung, and A. Smolic, "Extending SVC by content-adaptive spatial scalability," *in: Proc. of International Conference on Image Processing*, Brussels, Belgium, September 2011.

[11] Y P. Liu, S. Y. Zhang, R. F. Wang, and Y. Zhang, "Inter-frame fast coding algorithm in temporal scalability for traffic video," J*ournal of Zhejiang University (Engineering Science)*, vol. 47, no. 3, pp. 400- 408, 2013.

[12] X P. Ji and Z. Q. Wei, "Tracking method based on contour feature of vehicles and extended kalman filter," *Journal of Image and Graphics*, vol. 11, no. 2, pp. 267-272, 2011.

[13] M B. A. Baumann, J. Ebling, M. Koenig, H. S. Loos, W. N. M. Merkel, J. K. Warzelhan, and J. Yu, "A review and comparison of measures for automatic video surveillance systems," *EURASIP J. Image Video Process*, vol. 824726, pp. 30, 2008.
[14] E Soyak, A. Sotirios, and A. K. Tsaftaris, "Low-complexity tracking-aware H.264 video compression for transportation surveillance," *IEEE Trans. Circuits Syst. Video Technol*., vol. 21, no. 10, pp. 1378-1389, 2011.

Yunpeng. Liu received the Ph.D. from the School of Computer Science and Technology, Zhejiang University, in 2013, and is currently an associate professor at the School of Computer Science and Information, Zhejiang Wanli University, China. His research interests include video analysis, video coding , pattern recognition, and machine learning.

Renfang. Wang received the Ph.D. from the School of Computer Science and Technology, Zhejiang University, in 2008, and is currently a professor at the School of Computer Science and Information, Zhejiang Wanli University, China. His research interests include virtual reality, computer graphics, image engineering.

Dechao. Sun is a Ph.D student of Signal and Information Processing, Ningbo University, and is also an senior engineer at the School of Computer Science and Information, Zhejiang Wanli University, China. His research interests include digital geometry processing and embedded system.

Shijie. Yao is an undergraduate student of College of Computer Science and Information Technology, Zhejiang Wanli University. His research interests include image engineering and video coding.

Nayi. Hong is an undergraduate student of College of Computer Science and Information Technology, Zhejiang Wanli University. His research interests include image engineering and video coding.

Peng. Jin is an undergraduate student of College of Computer Science and Information Technology, Zhejiang Wanli University. His research interests include image engineering and video coding.