HEVC Compressed Domain Moving Object Detection and Classfication

Liang Zhao[1], Debin Zhao[1], Xiaopeng Fan[1], Zhihai He[2]
[1]Department of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
Email: {liang.zhao, dbzhao, fxp}@hit.edu.cn
[2] Department of Electrical and Computer Engineering, University of Missouri, Columbia, MO, USA
Email: hezhi@missouri.edu

***Abstract*—Compressed domain moving object segmentation and classification plays an important role in many real-time applications, such as video indexing and intelligent video surveillance. Compared with the previous international video coding standards, such as H.264/AVC, HEVC introduces a host of new coding features. Therefore, moving object segmentation and classification directly from HEVC compressed videos represents a new challenge. In this paper, we develop a method for segmenting and classifying moving objects, specifically, persons and vehicles, in the HEVC compression domain. We first train a classifier to determine if an image patch belongs to the foreground objects or background using HEVC syntax features. This will generate a bounding box which locates the object in the video frame. We then train a second classification model to classify the moving objects, either persons or vehicles, using bags of spatial-temporal HEVC syntax words. Our extensive experimental results demonstrate that the approach provides the remarkable performance and can classify moving person and vehicles accurately and robustly.**

Keywords—compressed domain, object segmentation, object classification, HEVC

I. INTRODUCTION

Object segmentation and classification from videos is a challenging problem with a variety of important applications, such as intelligent surveillance, video indexing and retrieval, etc. With access to original pixels, a wide variety of methods and tools have been developed in the literature to extract features and descriptors to characterize the video content for highly efficient object segmentation and classification [1-5]. Note that most video content are received or stored in compressed formats encoded with international video coding standards, such as HEVC [11]. To obtain the original video frame, we have to perform video decoding, which is a computation-intensive task, especially for high-resolution videos. To address this issue, compression-domain approaches have been explored for direct video content analysis which extracts features directly from the bit stream syntax, such as motion vectors and block coding modes, using a minimum amount of decoding efforts [6, 7]. The major advantage of compression-domain approaches is their low computational complexity since the full-scale decoding and reconstruction of pixels are avoided. In this paper, we focus on moving object detection and classification from compressed surveillance videos.

Recently, a number of moving object segmentation and classification algorithms in the H.264/AVC [14] compression domain have been reported [6-10]. Mezaris et al. use the motion information to locate spatiotemporal objects [6]. Babu et al. [7] introduce a method to accumulate motion vector (MV) information over time for moving object segmentation. Temporally accumulated MVs are further interpolated spatially to obtain a dense field. It should be noted that motion vectors extracted from the compressed bit stream may not represent the true object motion, since they are determined to optimize the coding efficiency. To address this issue, other information, such as DCT coefficients and macroblock (MB) partition, are used to detect and track moving objects [8-10]. For example, the number of bits used by each 4×4 block is used to detect moving object in [8] from H264 videos. This method is able to segment the object with a relatively accurate shape since it operates on 4×4 blocks. Porikli et al. [9] present a segmentation method that takes advantage of the inter-frame motion and intra-frame spatial frequency information embedded in MPEG videos. Pei et al. have developed an efficient moving object segmentation and tracking method by adaptively using the information from motion vectors, DCT coefficients and prediction modes [10].

HEVC is the newest international standard for video coding [11-13]. Very little work has been done on video object analysis directly from HEVC compressed videos. In addition, new features and syntax elements in HEVC, such as coding, prediction, and transform units, can be exploited to further improve the overall performance. Compared to existing methods in the literature, the major contributions of this work lie in the following aspects. First, we have developed and evaluated new HEVC syntax features in the spatiotemporal domain to achieve efficient object classification and maintain motion coherence and spatial compactness. Second, we have successfully extended the conventional "bag of words" model in the pixel domain to the compression domain.

The rest of paper is organized as follows. Section 2 presents the compression-domain moving object segmentation and classification using HEVC syntax features. Section 3 presents the experimental results of the proposed algorithm. Section 4 concludes the paper.

II. PROPOSED MOVING OBJECT SEGMENTATION AND CLASSIFICATION ALGORITHM

The overall objective of this paper is to develop a framework for compression-domain moving object detection and classification directly from HEVC videos. The overall algorithm framework of our system is illustrated in Fig. 1. It consists of two stages: moving object segmentation and person-vehicle

978-1-4799-5341-7/16/$31.00 ©2016 IEEE

classification. Each stage involves a training phase to learn the model from the training data and a testing phase to apply the learned model to test videos.

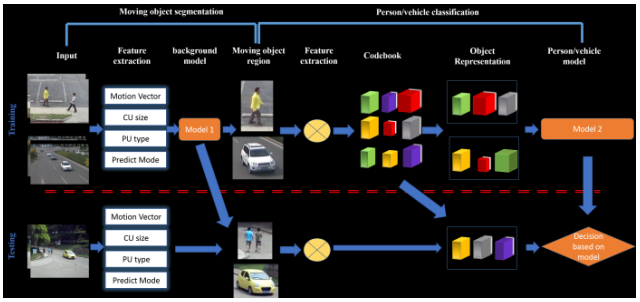

Fig. 1 the proposed framework for HEVC compression-domain object detection and classification

A. HEVC Compression-Domain Moving Object Segmentation

HEVC has introduced a quad-tree based coding approach where each picture is divided into square coding tree units (CTUs) [12]. Each CTU is the root of a coding tree, which is further divided into coding units (CUs). Their sizes can be adaptively chosen by using a quad-tree based partitioning with the tree leaves representing the CUs. Each CU is a root for a prediction tree. The prediction tree has only one level and describes how a CU can be further split into so-called prediction blocks (PUs), where 8 different partition modes are used for inter-coded CUs and only two modes are used for intra-coded CUs. These are important features in HEVC which significantly improves the overall coding efficiency.

Figs. 2(a) and (d) show an example of block partitions for a surveillance video frame with moving persons and vehicles. Here, the largest square blocks, smaller square blocks and rectangular blocks represent CTU, CU and PU, respectively. We observe that moving objects often have smaller CUs and PUs when compared to background image regions. In addition, object parts with non-rigid motions often have much smaller CUs and PUs. This is because they have difficulty in finding good matches from previous frames. This provides an important cue for the rigidness of object motion. The squares in Figs. 2(b) and (e) represent blocks with non-zero motion vectors whereas the square blocks in Figs. 2(c) and (f) represent intra coding modes. We can see that some parts of persons and vehicles are coded with non-zero motion vectors. However, some other parts of persons and vehicles are coded with intra modes or zero motion vectors. In addition, some background blocks, which are supposed to be static, have non-zero motion vectors due to noise or background changes.

Therefore, instead of solely relying on motion vector information for compression-domain moving object detection and segmentation, our proposed method combines various HEVC syntax elements, including motion vector information, CU and PU sizes, and intra/inter/skip coding modes to form a comprehensive feature vector to characterize the moving object in the compression domain. Our foreground-background classification operates on 4×4 blocks to achieve fine-grain segmentation of the moving objects. If the CU and PU have block partitions larger than 4×4, all 4×4 blocks inside are using the same motion vectors and coding modes as their containing block. Since the motion field is generally smooth, we also include the motion information of blocks in the spatiotemporal neighborhood to characterize the current block.

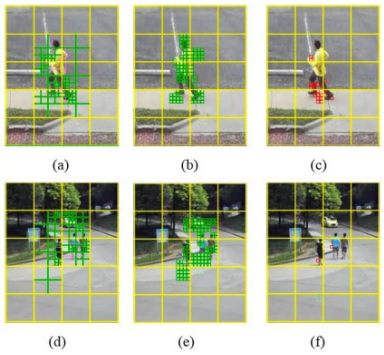

Fig. 2: (a) block partitions of a running man; (b) non-zero motion vectors of the moving object in (a); (c) intra coding modes of the moving object in (a); (d) block partitions of the moving person and vehicles; (e) non-zero motion vectors the moving objects in (d); (f) intra coding modes of the moving objects in (d).

A linear SVM classifier is trained to classify all 4×4 blocks into foreground and background blocks. During moving object detection and segmentation, the SVM classifier is applied to each 4×4 block in the current frame, and then connected component analysis is used to group all foreground blocks into foreground objects.

B. HEVC Compression-Domain Moving Object Classification

For object classification in surveillance videos, we aim to classify the detected moving objects into persons and vehicles using HEVC syntax features in the compression domain. In this work, we propose to explore an approach called bag of HEVC syntax words. The "bag of words" representation has been successfully used for object classification in the pixel domain [15]. Due to its histogram-type description, it is able to handle large variations in object locations, sizes, and viewing angles, as well as occlusions. This is particularly important for object detection-classification since detected object patches may not aligned well with the object boundary. The major contribution of this work is to establish a bag of words model in the HEVC domain for object classification.

This method has the following major steps: (1) describing each coding block within the moving object region using HEVC syntax features; (2) constructing a codebook using a clustering method; (3) representing each moving object using a normalized histogram of codewords from this codebook; and (4) train a two-class classifier to classify the moving objects into persons and vehicles. The main challenge is to determine effective features in the compression domain which have sufficient discrimination power between persons and vehicles. In this work, through extensive experiments and performance evaluations, we have identified four types of features, namely, the absolute value of motion vectors, CU sizes, prediction modes, and motion vector difference. The absolute value of the motion vector relates to the velocity of the object, which is a simple yet important for discrimination feature for persons and vehicles. Since the motion field within the rigid objects, such as vehicles, is often smooth, motion vector of spatial and temporal neighborhood blocks are also used as an important feature.

Although both persons and vehicles are outlined with smaller CU sizes, we have found that the distribution of different CU sizes within the object region is one of the most distinctive features for persons and vehicles classification, since vehicles often exhibit consistent motion within the object regions. Specifically, large CUs often appear on the boundary of the moving persons. However, for vehicles, they appear both on the boundary and at the center.

For effective person-vehicle classification in the HEVC compression domain, prediction modes and motion vector differences are also used. We observe that persons often undergo non-rigid deformations. In this case, it is harder to find a good match for the CU and PU. More blocks within the region of person are coded with intra prediction modes when compared to those of slow-moving and fast-moving cars. Because motion vectors within the moving vehicle are more consistent than those in persons, motion vector differences between neighborhood blocks of vehicles are smaller than those of persons. The motion vector differences of the x component and y component are computed separately. Specifically, the motion vector difference of current block is computed as follows:

$$MVD_{N_{i,j}} = \begin{cases} abs\left(MV_{N_{i,j}}^{x} - MV_C^{x}\right) + abs\left(MV_{N_{i,j}}^{y} - MV_C^{y}\right), & if\ MV_{N_{i,j}} \neq 0 \\ 0, & else \end{cases} \quad (1)$$

$$MVD\_C_0^t = \frac{\sum_{i=0,j=0}^{i=2,j=2} MVD\_N_{i,j} + count/2}{count} \quad (2)$$

Where $MV\_N_{i,j}$ denotes the motion vector of the current block, $MV\_N_{i,j}^{x}$ denotes the x component of $MV\_N_{i,j}$, $MV\_N_{i,j}^{y}$ denotes the y component of $MV\_N_{i,j}$, MV_C denotes the motion vector of the current block, MV_C^{x} is the x component of MV_C, MV_C^{y} denotes the y component of MV_C, $MVD\_N_{i,j}$ denotes the motion vector difference between current block and its neighborhood block $N_{i,j}$, $count$ is the number of neighborhood blocks with non-zero motion, and $MVD\_C_0^t$ is the total motion vector difference of current block and all of its neighborhood blocks. Meanwhile, $MVD\_C_0^{r_1}$ and $MVD\_C_0^{r_2}$ denote the motion vector difference of its temporal neighborhood blocks in reference frames r_1 and r_2, respectively.

Once the features of all blocks in the training datasets have been extracted, we apply k-mean clustering to these feature vectors into M clusters. The center of each cluster becomes a codeword. In total, the codebook will have M codewords. For example, in our experiments, we set M=600. Each moving object, either a person or a vehicle, will contain a large number of blocks. We compute its feature distance between each block Bn and each codeword Cm. We find the codeword which has the minimum distance to Bn and cast Bn to the bin of this codeword. In this way, we can generate a codeword histogram for all blocks in the object. After normalized by its size, the histogram is used as the feature to describe the moving object. With this feature description scheme and the training data, we train a two-class linear SVM classifier for person-vehicle classification.

III. Experimental Results

To evaluate the performance of our proposed HEVC compression-domain moving object detection and classification scheme, we have collected 5 surveillance videos which represent typical surveillance scenarios. Fig. 3 shows example frames of these test videos. We use separate surveillance videos for training. Videos are in the YUV 4:2:0 format at 25 frames per second. All sequences were encoded using the HEVC HM v10.0 encoder. HEVC syntax features, such as motion vectors, prediction modes, and partition information, CU sizes, and PU types, are extracted from the HEVC compressed bit stream.

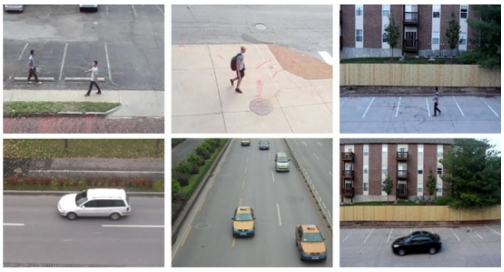

Fig. 3: Example frames of test videos

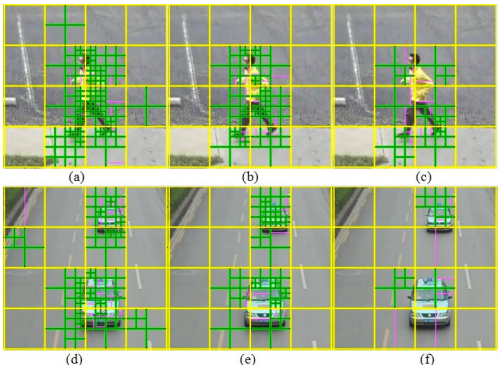

Fig. 4: (a) CU sizes of moving person with QP = 22; (b) CU sizes of moving person with QP = 27; (c) CU sizes of moving person with QP = 32; (d) CU sizes of moving vehicles with QP = 22; (e) CU sizes of moving vehicles with QP = 27; (f) CU sizes of moving vehicles with QP = 32.

In order to train the models for moving object detection and classification, we use HEVC surveillance videos of persons and vehicles of a wide variety of moving speeds, patterns, and poses. The codebook is generated from this training data and fixed during testing. We observe that, during HEVC coding, when the quantization parameter (QP) increases, more prediction residue coefficients are quantized to zeroes and the encoder prefers to select larger CU sizes to reduce the overhead. Figs. 4 (a)-(f) shows how the CU sizes of moving persons and vehicles change with the QP where 3 different QPs (22, 27 and 32) are used to encode the video sequence. We can see that moving objects are outlined with much smaller CU sizes when coded with smaller QPs. The CU size within the moving object increases with the QP. Therefore, we need to train different set of models for moving object detection and classification for different QPs. In this work, we have found that two models for QP = 22 and 32 are sufficient to handle this variation due to QP changes. To demonstrate the robustness of our proposed algorithm, we also test our model on QP = 27 and 30 respectively.

Fig. 5 shows several examples of moving object detection and segmentation results using our learned SVM model with HEVC syntax features. Here, (a) and (d) are the original video

frames of a walking person and a moving vehicle. (b) and (e) are the segmentation results generated by our algorithm. (c) and (f) shows the detected bounding boxes (image patches) of the moving object. We can see that the bounding boxes are well aligned with the moving object.

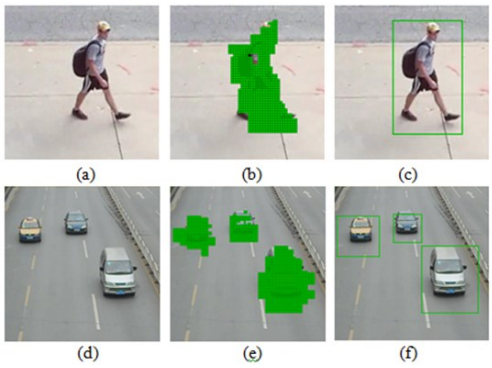

Fig. 8: (a) video data for a walking man with slight shadow; (b) foreground segmentation of (a) (c) bounding box of the walking man in (a); (d) video data in a speedway; (e) foreground segmentation of (a); (f) bounding box of the moving vehicles in (c).

Tables I shows the performance of person and vehicle classification results on different test videos and quantization settings. For performance evaluations, we manually label each moving object, either a person or a vehicle, as the ground truth. In total, there are 1290 persons and 1342 vehicles. We use the accuracy as the performance metric, which is defined as

$$Accuracy = \frac{TP}{TP+FP} \quad (3)$$

Table I. System performance summary

Sequence	QP = 22	QP = 27	QP = 30	QP = 32
Seq_1	95.6	90.8	95.6	92.2
Seq_2	96.4	89.1	96.4	95.5
Seq_3	98.0	99.5	98.0	99.0
Seq_4	97.5	97.5	97.5	97.8
Seq_5	95.0	95.7	95.0	98.1
Average	**96.0**	**95.8**	**96.0**	**97.5**

Here, TP and FP are true positive and false positive rates, respectively. We can see that the overall classification accuracy of the proposed method is over 95% for more than 2500 moving objects. Our system achieves consistent performance across different QPs. Note that we have only used two classification models trained for QP=22 and 32.

IV. Conclusion

In this paper, we have presented a novel approach to segment and classify the moving objects in a HEVC-compressed video. The only data from the compressed stream used in the proposed method are the motion vectors and the associated coding modes. Firstly, moving object region is automatically segmented by using the feature vectors extracted from the HEVC compression domain. Then, we explored the possibility of applying the representation of "bag of temporal-spatial words" to classify the moving objects in HEVC compression domain. The proposed method has a fairly low processing time, yet still provides high accuracy.

Acknowledgment

This work was supported in part by the Major State Basic Research Development Program (973 Program) of China under Grant 2015CB351804; in part by the National Science Foundation of China under Grants 61472101 and 61390513.

References

[1] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," IEEE Trans. Pattern Anal. Mach. Intell., vol. 25, no. 5, pp. 564–575, May 2003.

[2] H.F. Xu, A.A. Younis, M.R. Kabuka, Automatic moving object extraction for content-based applications, IEEE Trans. Circuits Syst. Video Technol., vol. 14, no. 6, pp. 796–812, Jun. 2004.

[3] Y. Wang, K.-F. Loe, T. Tan, and J.-K. Wu, "Spatiotemporal video segmentation based on graphical models," IEEE Trans. Image Process., vol. 14, no. 7, pp. 937–947, Jul. 2005.

[4] M. Grundmann, V. Kwatra,M. Han, and I. Essa, "Efficient hierarchical graph-based video segmentation," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 2141–2148, Jun. 2010.

[5] Hidetomo Sakaino, "Video-Based Tracking, Learning, and Recognition Method for Multiple Moving Objects", IEEE Trans. Circuits Syst. Video Technol., vol. 14, no. 5, pp. 1661–1674, Oct. 2013

[6] V. Mezaris, I. Kompatsiaris, N. V. Boulgouris, and M. G. Strintzis, "Real-time compressed-domain spatiotemporal segmentation and ontologies for video indexing and retrieval," IEEE Trans. Circuits Syst. Video Technol., vol. 14, no. 5, pp. 606–621, May 2004.

[7] R. V. Babu, K. R. Ramakrishnan, H. S. Srinivasan, "Video object segmentation: A compression domain approach," IEEE Trans. Circuits Syst. Video Technol., vol. 14, no. 4, pp. 462–474, Apr. 2004.

[8] C. Poppe, S. De Bruyne, T. Paridaens, P. Lambert, and R. Van de Walle, "Moving object detection in the H.264/AVC compression domain for video surveillance applications," J. Visual Commun. Image Represent., vol. 20, no. 6, pp. 428–437, Aug. 2009.

[9] Fatih Porikli, Faisal Bashir, and Huifang Sun, "Compression domain Video Object Segmentation," IEEE Trans. Circuits Syst. Video Technol., vol. 20, no. 1, pp. 2–14, Jan. 2010.

[10] Pei Dong, Yong Xia, Li Zhuo and Dagan Feng, "Real-time moving object segmentation and tracking for H.264/AVC surveillance videos," in Proc. IEEE Int. Conf. Image Processing, Brussels, pp. 11–14, Sep. 2011.

[11] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard", IEEE Trans. Circuits Syst. Video Technol., vol. 22, no. 12, pp. 1649-1668, Dec. 2012.

[12] I.-K. Kim, J. Min, T. Lee, W.-J. Han, J. Park, "Block partitioning structure in the HEVC standard", IEEE Trans. Circuits Syst. Video Technol., vol. 22, no. 12, pp. 1697–1706, Dec. 2012.

[13] J. Ohm, G. J. Sullivan, H. Schwarz, T. K. Tan, T. Wiegand, "Comparison of the coding efficiency of video coding standards -- including high efficiency video coding (HEVC)", IEEE Trans. Circuits Syst. Video Technol., vol. 22, no. 12, pp. 1669–1684, Dec. 2012.

[14] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," IEEE Trans. Circuits Syst. Video Technol., vol. 13, no. 7, pp. 560–576, Jul. 2003.

[15] Csurka, Gabriella, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. "Visual categorization with bags of keypoints", In Workshop on statistical learning in computer vision, ECCV, pp. 1-22, May 2004.