Video Analytical Coding: When Video Coding Meets Video Analysis

Yuyang Liu, Ce Zhu, Min Mao, Fangliang Song, Frédéric Dufaux, Xiang Zhang

Video Analytical Coding: When Video Coding Meets Video Analysis

Yuyang Liu[a], Ce Zhu[a], *Fellow, IEEE*, Min Mao[a], Fangliang Song[a],
Frederic Dufaux[b], *Fellow, IEEE*, Xiang Zhang[a]
[a]University of Electronic Science and Technology of China, Chengdu, China
[b]Laboratoire des Signaux et Systèmes, Paris, France

Abstract

Leveraging on the properties of human visual system, most of the well-designed video coding standards utilize rate-distortion optimization techniques by maximizing a fidelity cost function (e.g. peak signal noise ratio, PSNR) under an available bit rate budget constrain. However, a huge amount of video data is consumed by computers rather than by human beings in several application scenarios. In view of this, this paper proposes a new coding framework called video analytical coding (VAC) for video analysis. We use the term "analytical distortion" to denote the difference of video analysis performance when video quality degrades and analytical distortion is estimated by compression distortion. Meanwhile, we develop a new rate-analytical-distortion optimization (RADO) method to trade off the bit rate and the analytical distortion. Specifically, we consider moving object detection as the analysis task and develop a novel rate analytical distortion (RAD) model and a quantization parameter adaptation strategy for video coding, where the analytical distortion is related to the object detection performance represented as F1-measure. Experimental results show that the performance of the video analysis task can be significantly improved (up to 40% reduction of analytical distortion).

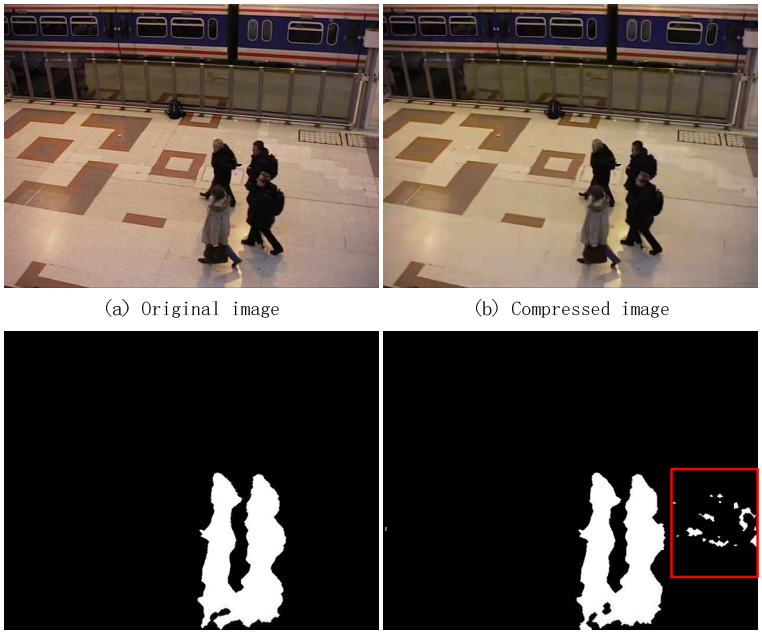
Index Terms: Video analysis; video analytical coding; analytical distortion; rate distortion optimization.

1 Introduction

The increasing availability of portable or installed cameras and the introduction of new multimedia applications to fulfill emerging needs, have given rise to new requirements on video compression and communication. As for many multimedia applications, e.g. surveillance, video content is not only presented to human beings but also analyzed by computers for variously applicable purposes, such as object detection, tracking, recognition, and so on. In other words, computers have become as viewers of the videos. Meanwhile, the considerable amount of generated videos need to be efficiently compressed due to the cost-effective storage and bandwidth limitation.

Currently, the main goal of most studies on video coding is to achieve high coding efficiency. Most of the widely deployed video coding standards, such as H.264/MREG-4 AVC (Advanced Video Coding) [1] and HEVC (High Efficiency Video Coding) [2], are designed under the assumption that human beings are the target viewers. Meanwhile, traditional rate distortion optimization (RDO) framework is applied into the video coding standards by optimizing the trade-off between the entropy of the discretized representation (rate) and the error arising from the quantization (distortion). However, applying the traditional RDO framework during video compression may be suboptimal when the video is intended for machine analysis. The critical issue is that resulting compression distortion may bring a negative impact on the video analysis performance. This point is shown in Fig. 1. More precisely, a foreground extraction

algorithm is run on the same frame without and with compression. Obviously, it can be observed that the extraction results are quite different, especially for the region in red box in Fig. 1(d). Generally, more compression distortion lead to more differences. Additionally, research on video analysis mainly aims to improve the video analysis performance and pays little attention to the negative impact introduced by compression distortion.



(a) Original image (b) Compressed image

(c) Foreground extraction result on original image (d) Foreground extraction result on compressed image

Fig. 1 Foreground extraction results on the same frame, without and with compression

In view of this, two problems should be addressed [3]: (i) How different the video analysis results will be depending on different video quality levels? (ii) How much bandwidth can be saved? Obviously, there is a trade-off between video quality and accuracy of the video analysis algorithms. To solve above problems, we propose a new coding framework, namely video analytical coding (VAC). We develop a new rate-analytical-distortion optimization (RADO) method, where the term "analytical distortion" represents the difference of the video analysis algorithms' performance when video quality degrades. The compression distortion is measure by Sum of Absolute Difference (SAD) in this paper and SAD of each non-I (P or B) frame is estimated by that of an I frame according to the temporal relationship. The analytical distortion is predicted by the compression distortion. Specifically, we choose to focus on one fundamental video analysis task, moving object detection. Accordingly, we develop a rate analytical distortion (RAD) model and an object based quantization parameter (QP) adaptation strategy for video coding. Our proposed method is fully standard compatible and the encoded bit-stream can be decoded by any HEVC decoder. This paper is an extension of our previous work in [4] and the following are the contributions of this paper.

- We propose a novel coding framework VAC for video analysis and accordingly develop a new RADO method to trade off video quality and accuracy of the video analysis algorithm.
- To avoid a two-pass coding procedure, we introduce a model to predict the compression distortion. Meanwhile, we use the predicted compression distortion to estimate analytical distortion.
- We propose an object based QP adaptation strategy where the object area is compressed using a relatively smaller QP compared with the background.

The reminder of the paper is organized as follows. We review related work in section 2. In section 3,

we first present the VAC framework and the compression distortion prediction model for non-I frame. Then, a brief review of our previous work in [4] is presented including the RADO method and the weighting parameter selection. Our proposed object based QP adaptation strategy is presented in section 3.5. Section 4 experimentally evaluates the proposed method. Section 5 concludes the paper.

2 Related Work

A considerable amount of work in video coding aims at improving the coding efficiency. The state-of-the-art video coding standard HEVC can achieve a near 50% bit rate reduction, while keeping comparable perceptual quality, compared with its predecessor H.264/MPEG-4 AVC. Besides, there are many developed RDO schemes for HEVC standard [5]-[7]. In [5], a multiple QP optimization scheme is introduced, where multiple QP candidates are checked to find the best one by minimizing the rate-distortion cost. Such multiple QP optimization scheme will definitely increase the coding complexity since the video codec needs extra time to find the best QP from the QP candidates. In our previous work [8-9], we investigated a RDO scheme taking the inter-frame dependency into account, where the impact of coding performance of the current coding unit on that of the following frames is considered.

However, the above RDO schemes do not consider the fact that human visual system pays more attention to region-of-interest (ROI) or foreground objects [10]. To address this problem, researchers investigate a few coding methods for some specific video applications, such as teleconference and surveillance.

- **Teleconference**: In conferencing applications, the face area in the video usually attracts the viewers and the background is stationary in most cases. Leveraging on this property, a few ROI based coding methods are developed to further save the transmission bandwidth. Liu *et al.* [11] introduced an efficient face ROI determination method using skin color combined with direct frame difference. A relatively larger portion of bits and more computational power were assigned to encode the detected ROI. Xiong *et al.* [12] used a motion based face detection method combined with an active contour model to find the well-located and compact face regions. Then, they proposed a facial feature priority based bit allocation method for ROI conversational video coding. Zhao *et al.* [13] proposed a ROI coding scheme for synthesized video aiming at achieving better and consistent quality given a target bit rate. However, there are two major problems in ROI coding methods: 1) the detection and segmentation of the ROI, and 2) more computing power at the encoder side.
- **Surveillance**: In video surveillance, there is typically little camera motion so that the background parts are mainly static. In order to take advantage of this property, some research has been done recently that selects or generates a picture as a special reference for coding the background regions. Tiwari *et al.* [14] proposed a long-term reference selection method using simulated annealing, where the selected reference is compressed with high quality. Pushkar *et al.* [15] selected a coded picture by taking the usage of skip mode into account. Paul *et al.* [16] modeled pixels of many pictures at the same position as a Gaussian mixture distribution and generated a background picture using the most probable pixel value. Zhang *et al.* [17] generated a background picture by simply averaging many pictures pixel by pixel. Chen *et al.* [18] proposed an approach that generates a background picture by updating pixels in some blocks instead of the whole picture. However, these methods have some drawbacks. First, it is often impossible to select a picture as the background reference, which contains the entire background content. Second, generating a picture needs extra processing time and delay to look ahead many future frames or wait for many decoded frames. Furthermore, the generated picture must be compressed with high quality and transmitted to the

decoder side. Finally, it is still a problem for these methods to deal with video sequence containing many moving foreground objects, as the foreground pixels are difficult to be filtered out from the background.

The above-mentioned schemes may be suboptimal when video content is consumed by computers rather than human beings. In this context, recent research work focuses on feature-preserving coding, where only feature descriptors are transmitted to the server side. It can be categorized into two paradigms: compress-then-analysis (CTA) and analysis-then-compress (ATC). Redondi et al. [19] compared the performance of CTA and ATC for image analysis in visual sensor networks and found out that the performance of ATC paradigm was better. Baroffio *et al.* [20] proposed a coding architecture for coding local features (e.g. SIFT, SURF) extracted from a video sequence, which can be adopted to implement the ATC paradigm. Both intra and inter-frame coding modes were applied in the proposed coding architecture and the final coding mode was determined by comparing the costs of two coding modes. In their later work [21], the coding architecture was applied to coding binary local features. As only features are transmitted, video content cannot be watched at the server side. It is inapplicable in some scenarios (e.g. video surveillance) where it is necessary to visualize the video content. By contrast, Chao and Steinbach [22] proposed a novel framework, in which keypoints extracted from a video were encoded and transmitted along with the compressed video. However, this framework needs more bits to transmit the feature descriptors.

Few studies have addressed the impact of video compression on video analysis. Korshunov and Ooi [3] proposed a formal rate-accuracy optimization framework, where the encoding parameters in distributed video surveillance systems could be determined given a target bit rate or accuracy. Furthermore, they denoted that there exists a sweet spot where reducing the bit rate would not significantly affect the accuracy of face recognition and tracking algorithms. Kokiopoulou and Frossard [23] proposed a supervised dimensionality reduction scheme which provides a tradeoff between compression and discriminant feature extraction. Liao *et al.* [24] proposed an analysis-oriented ROI based coding approach to reduce the impact of video compression on the performance of video analysis. However, the approach in [24] needs prior knowledge to detect ROI, which cannot always be obtained in practice.

3 Video Analytical Coding

In section 3.1, our proposed video analytical coding framework is presented. In the following, the compression distortion prediction model is introduced in section 3.2 and a brief review of our previous work is presented in section 3.3. In section 3.4, our proposed parameter adaptation procedure is described. Finally, section 3.5 illustrates the proposed object based QP adaptation strategy.

3.1 Video Analytical Coding Framework

The flow chart of our proposed video analytical coding framework is shown in Fig. 2. In our framework, the QP of an I frame will be refined according to our proposed QP adaptation strategy (see section 3.5) and traditional RDO technique is used to compress the I frame. Then, the compression distortion of an I frame denoted as SAD_I can be calculated after decoding. The compression distortion (SAD) of a non-I frame (P or B frame) is predicted by SAD_I according to the compression distortion prediction model (see section 3.2). The predicted compression distortion is then used to obtain the analytical distortion which denotes the difference of video analysis performance when video quality

degrades. Meanwhile, a simple frame subtraction algorithm is utilized to obtain the moving object area (foreground) of the current frame. The object area is used for updating the weighting parameter (see section 3.4) and the QP offset of the foreground coding blocks in P or B frame. Finally, the non-I frame is compressed by our proposed RADO method (see section 3.3).

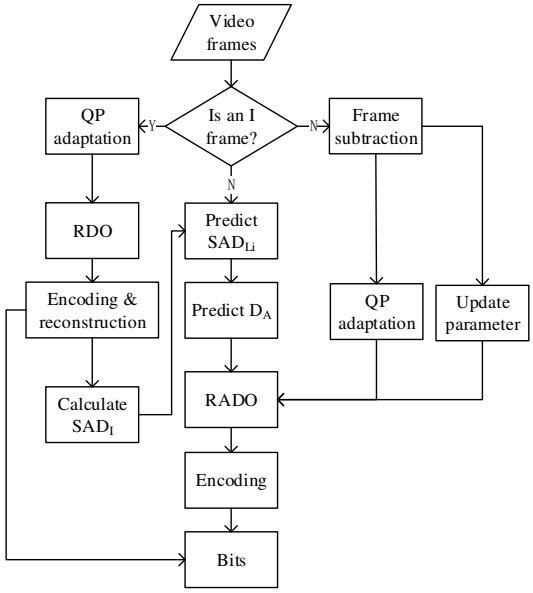


Fig.2 Flow chart of the proposed framework

In our previous work [4], a two-pass encoding procedure is applied to obtain the compression distortion. In order to reduce the encoding complexity, we introduce a prediction model to estimate the compression distortion of a non-I frame according to the temporal relationship. The following section will present the details of the prediction model.

3.2 Compression Distortion Prediction Model

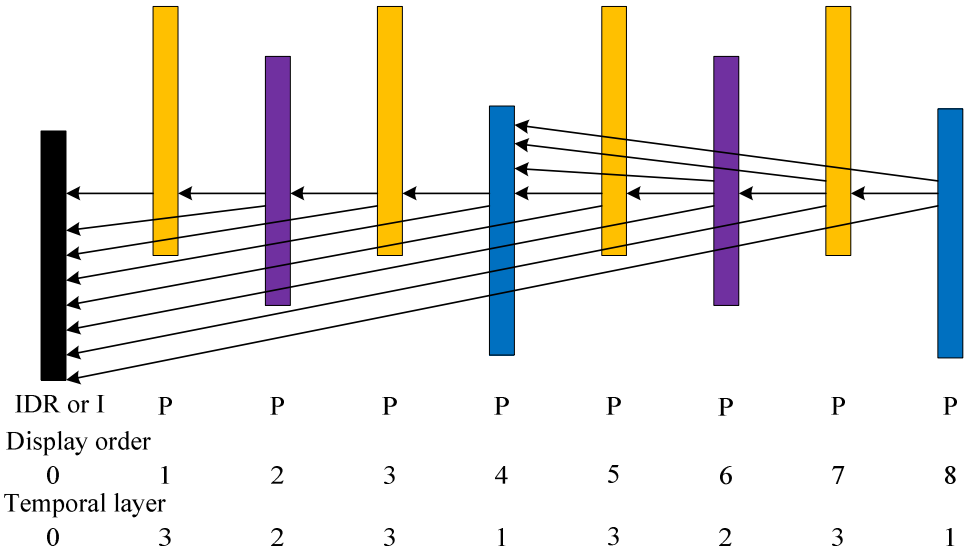


Fig. 3 Graphical presentation of hierarchical coding structure and the reference structure in HEVC low delay P configuration

In the state-of-the-art video coding standard HEVC, hierarchical coding structure is adopted to improve the coding efficiency. Fig. 3 shows the hierarchical coding structure under low delay P (LDP) configuration. Frames in different temporal layers are highlighted by different colors. In temporal layer

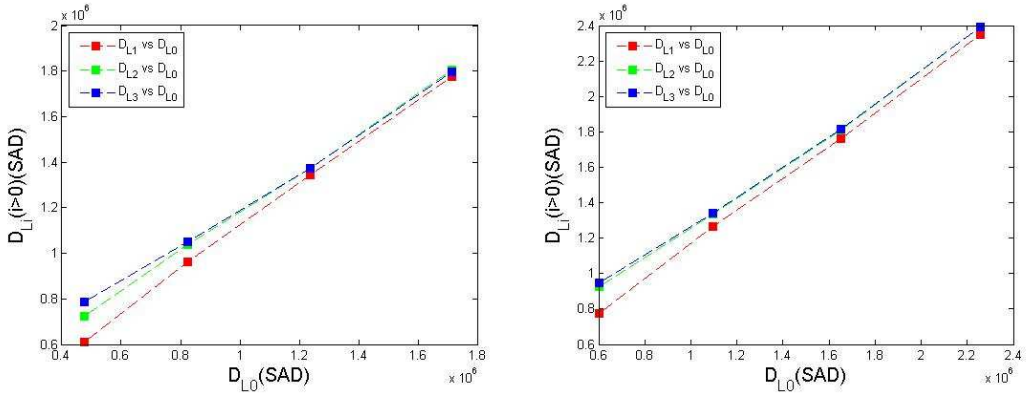
0, an I or IDR frame is denoted by a black bar. Frames in temporal layer 1, 2 and 3 are represented by blue, purple and orange bars respectively. In the following, an I frame and the following P frames are denoted by layers L_i, where i indexes the temporal layers from 0 to 3. Meanwhile, it can be easily seen that frames in a lower temporal layer (e.g. 0 and 1) are directly or indirectly referenced by frames in higher temporal layers.

In view of this, the temporal relationships are investigated in terms of distortion between I frame and other frames in different layers [25]. The experiments are conducted on the HEVC test model HM 16.7 [26]. Two video clips (Clip 1 and Clip 2) are encoded under LDP configuration. In the experiments, the QP is set as 22, 27, 32 and 37. The other coding parameters are set as the default use case. Meanwhile, the distortion is measured by SAD and the distortion of each layer ($i \geq 0$) at each QP point is averaged over the frames in the same layer.

The relationship in the average distortion between the layer L_0 and other layers ($i > 0$) is shown in Fig. 4. From Fig. 4, it can be seen that the average distortion of higher layer ($i > 0$) increases linearly with that of layer 0. The compression distortion prediction model is therefore formulated by a linear equation with $i \in \{1,2,3\}$

$$D_{L_i} = k_{L_i} \cdot D_{L_0} + b_{L_i}, \tag{1}$$

where D_{L_i} represents the average distortion of the i-th layers, k_{L_i} and b_{L_i} are the model parameters.



(a) Clip 1 (b) Clip 2

Fig. 4 Distortion dependency between temporal base layer and higher temporal layers

Using Eq. (1), the resulting predicted compression distortion is used to predict the analytical distortion [4]. Then, the predicted analytical distortion is used in our proposed RADO method which is presented in the following section.

3.3 Rate Analytical Distortion Optimization

Nowadays, a large amount of generated videos are consumed by computers running video analysis algorithms for some application-related purposes, such as face detection and recognition. However, most of the videos are compressed in a lossy way to further raise the compression ratio, which may decline the analysis algorithms' performance. In addition, the goal of the well-known RDO technology is to minimize the compression distortion under an available bit rate budget constrain [27, 28], without considering the above-mentioned scenarios. In this section, we address this issue and present our proposed RADO approach. Specifically, the difference of video analysis performance caused by video compression is denoted as analytical distortion and we use the term "D_A" to represent it.

In order to reduce the negative effect introduced by lossy compression while maintaining the video

coding efficiency, we formulate the RADO problem by jointly minimizing the compression distortion D_C in pixel domain and the analytical distortion D_A at a given bit rate R [1]. It can be written as

$$\min\ D_C + \tau D_A \quad s.t.\ R \le R_T \quad , \tag{2}$$

where τ is a weighting parameter and R_T is the available bit rate budget. Obviously, Eq. (2) will be reduced to the traditional RDO formulation when $\tau = 0$. Employing the Lagrangian multiplier method [29], the constrained RADO problem in Eq. (2) can be converted into an unconstrained form, which can be written as

$$\min\ \{J_{new}\}, \text{where } J_{new} = D_C + \tau D_A + \lambda_{new} R\,, \tag{3}$$

where J_{new} is the cost function and λ_{new} is the Lagrangian multiplier. As so far, it is difficult to get the optimal solution in Eq. (3) without modeling the interaction between D_A and R. In view of this, we empirically design a derivable RAD model, which can be expressed as

$$R = C_1\, \mathrm{e}^{C_2 D_A}\,, \tag{4}$$

where C_1 and C_2 are constant parameters. One more thing, we do not change the rate model in HEVC codec, so R is differentiable with respect to D_C. Consequently, the minimal cost J_{new} is obtained when

$$\frac{\partial J_{new}}{\partial R} = \frac{\partial D_C}{\partial R} + \tau \frac{\partial D_A}{\partial R} + \lambda_{new} = 0\,, \tag{5}$$

$$\lambda_{new} = -\frac{\partial D_C}{\partial R} - \tau \frac{\partial D_A}{\partial R}\,. \tag{6}$$

Actually, the derivation of D_C with respect to R can be directly obtained from the HEVC codec. When the weighting parameter $\tau = 0$, it represents the Lagrangian multiplier of traditional RDO problem which is represented by the term λ_{HM}. Then, Eq. (6) can be rewritten as

$$\lambda_{new} = \lambda_{HM} - \tau \frac{\partial D_A}{\partial R}\,. \tag{7}$$

Finally, we build the relationship between λ_{new} and λ_{HM}, which can obtain a fast solution of λ_{new}. Furthermore, once the weighting parameter τ is determined, the Lagrangian multiplier λ_{new} can be calculated from Eq. (7). In the subsequent section, the weighting parameter adaptation procedure is presented in details.

---

[1] In this paper, we do not aim at studying the absolute performance of a video analysis algorithm itself. Rather, we are more specifically concerned with how the analysis algorithm behaves when video quality degrades. For instance, if the video analysis algorithm can achieve the same performance when applied on the full quality video and the degraded video, the analytical distortion is considered to be zero. Besides, the analytical distortion is predicted by a linear model [4].

3.4 Weighting Parameter Adaptation

Obviously, the optimal Lagrangian multiplier λ_{new} in Eq. (10) cannot be obtained without setting the weighting parameter τ. Therefore, the weighting parameter τ plays an important role in our proposed RADO method. In order to explore the impact of τ on the proposed framework, we test a set of values

where τ ranges from 0.1 to 0.9. Four QP values (22, 27, 32, 37) are selected and all the experiments are conducted under LDP configuration. Besides, four reference video clips (Clip 1-4) are selected, including three indoor video clips (Clip 1-3) from PETS2006 [30] and one outdoor video clip (Clip 4) from PETS2009 [31]. Clip 1 to Clip 4 are captured by stationary cameras without any zooming. Meanwhile, the selected video clips all have 600 frames at a frame rate of 30fps, and have resolution of 720x576 except for Clip 4 which is of size 768x576. Fig. 5 shows the thumbnails of the four reference video clips.

(a) Clip 1 (b) Clip 2

(c) Clip 3 (d) Clip 4

Fig. 5 Thumbnail for each video clip

Specifically, we consider a task of moving object detection, which is a fundamental component in many application scenarios. The moving object detection algorithm in [32] is selected to study the trade-off between the accuracy of moving object detection and the coding rate. Meanwhile, we choose the F1-measure to evaluate the performance of moving object detection task [33]. Besides, we consider the moving object detection problem as a binary classification scheme [34]. Then, positives and negatives are counted at the pixel level. In the following, True and False Positives (denoted by *TP* and *FP* respectively) refer to the number of detected positives according to ground truth, and similar for True and False Negatives (denoted by *TN* and *FN*). In particular, we take the detection result of pristine video as the ground truth by considering two aspects. On the one hand, the difference of detection results (analytical distortion) between the pristine video and the compressed video can be directly indicated in this way, since we are more specifically concerned with analytical distortion as aforementioned. On the other hand, it is difficult to obtain the ground truth in practice.

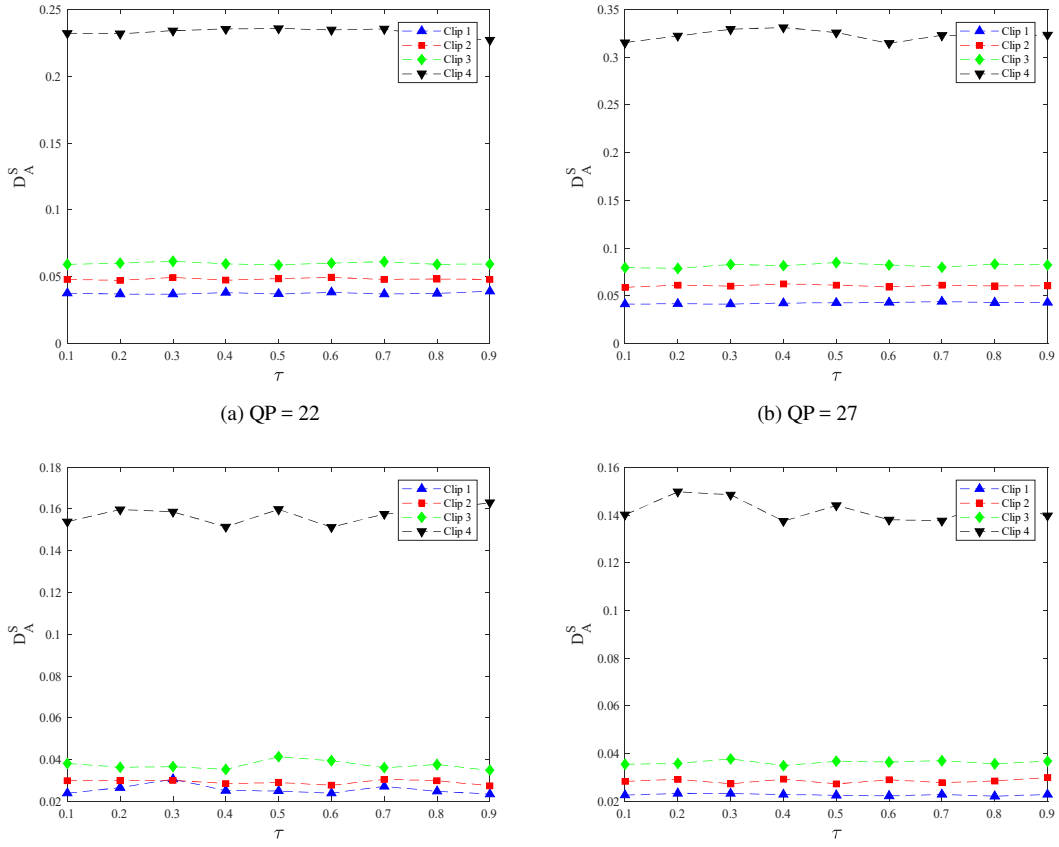
Basically, the F1-measure consists of two parts, namely recall and precision, which are calculated by the number of *TP*/*FP* and the number of *TN*/*FN*. The precision *pr* and recall *re* can be calculated by Eq. (8) and Eq. (9) respectively.

$$pr = \frac{TP}{TP + FP}, \tag{8}$$

$$re = \frac{TP}{TP + FN}. \tag{9}$$

We use F1-measure given by

$$F = 2 \times \frac{pr \times re}{pr + re}. \tag{10}$$



(a) QP = 22 (b) QP = 27

(c) QP = 32 (d) QP = 37

Fig. 6 Experimental results of D_A^S with four video clips under different values of τ. (a), (b), (c) and (d) show the results at four QPs respectively.

Using F1-measure, the analytical distortion of the *i*-th frame D_A^i can be expressed as

$$D_A^i = 1 - F_i, \tag{11}$$

where F_i the F1-measure value of the *i*-th frame. Then, the average analytical distortion D_A^S over the video sequence *S* can be written as

$$D_A^S = \frac{1}{N} \sum_{i=1}^{N} D_A^i, \tag{12}$$

where *N* is the total number of frames. In the following of this paper, the analytical distortion is represented as D_A^S.

Fig. 6 shows the experimental results of four video clips under different values of τ. As a first

observation, the average analytical distortion D_A^S of three indoor video clips (Clip 1-3) is more or less constant when τ increases. The second observation is that the average analytical distortion D_A^S of Clip 4 oscillates along with the increase of τ. There are two reasons. First, illumination changes will reduce the robustness of object detection algorithms. The analytical distortion will increase along with the increase of the differences between the detection results of the frame subtraction algorithm applied in our proposed method and that of the object detection algorithm in [32]. Otherwise, the analytical distortion will be reduced as shown in the experimental results. Second, the weighting parameter τ has an impact on the mode decision process, which will influence the detection results of the object detection algorithm in [32]. Meanwhile, we find that the object area also has a great impact on the analytical distortion and the weighting parameter τ is updated according to the detected object area of each frame [4]. The weighting parameter τ is updated by

$$\tau = \rho \times \exp(\frac{Area}{w \times h}), \tag{13}$$

where w and h denote the frame width and frame height, $Area$ represents the total detected object area of each frame and ρ is a constant parameter. In this paper, ρ is set to be 0.1 according to the experimental results shown in Fig. 6.

3.5 Object Based QP Adaptation Strategy

According to the work in [10], compression distortion has more influence on analytical distortion in the object area than that in background. In view of this, we develop an object based QP adaptation strategy. It can be expressed as

$$QP_i = \begin{cases} QP_f - \Delta QP, & if \ \ i \in object \\ QP_f + \Delta QP, & if \ \ i \notin object \end{cases}, \tag{14}$$

where QP_i and QP_f represent the quantization parameter of the i-th coding tree unit and the frame level QP, and ΔQP denotes the QP offset. Specifically, according our previous work in [8, 9], the coding performance of an I frame may have a strong impact on that of the following frames. Therefore, we also set a QP offset to the I frame in this paper, which can be expressed as

$$QP_I = QP_B - \Delta QP_I, \tag{15}$$

where QP_I and QP_B represent the quantization parameter of an I frame and the initial QP, and ΔQP_I denotes the QP offset.

4 Experimental Results and Discussion

In this section, we conduct comparative experiments on the HEVC test model (HM 16.7). Video sequences are encoded under LDP configuration. The other coding parameters are set as the default case. Four QP values (22, 27, 32, 37) are selected. Both QP offsets ΔQP and ΔQP_I (see Eq. (14) and Eq. (15)) are set to be 2. The rate-distortion performance of the proposed method is measured in terms of BD-rate saving over the HM 16.7.

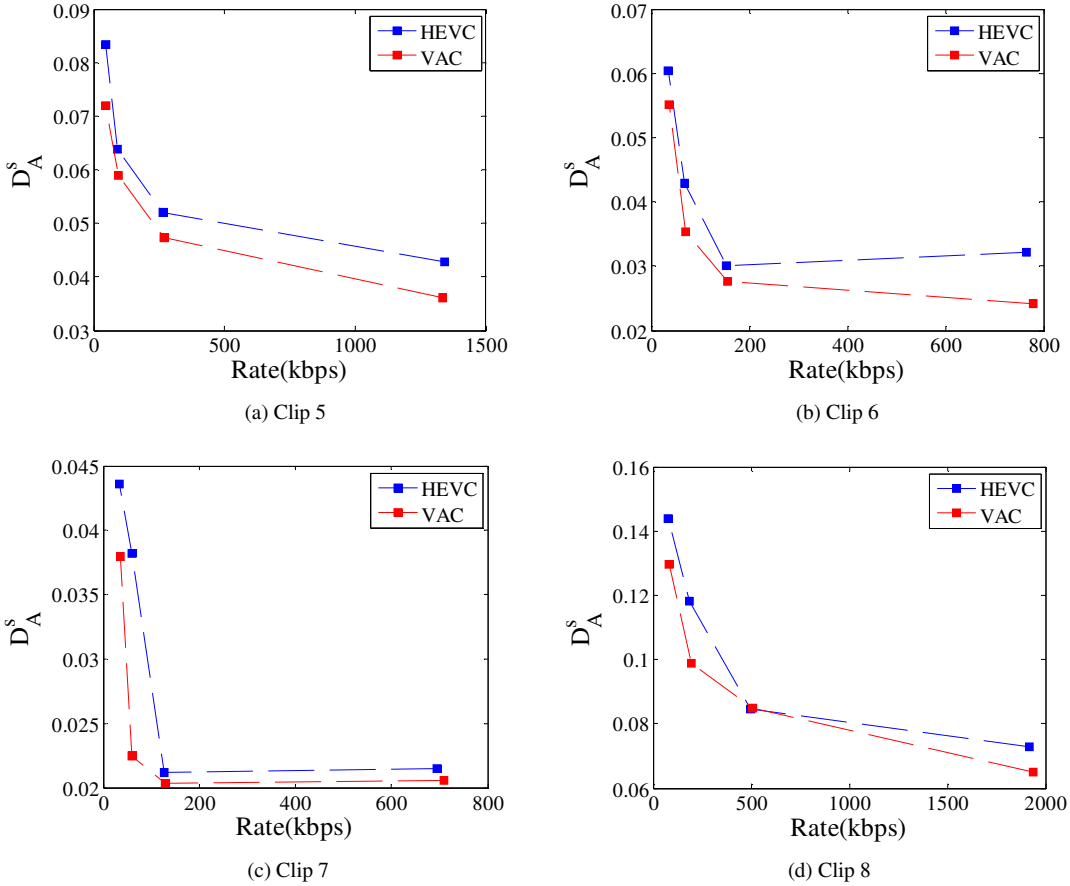
In order to show the rate analytical distortion performance of VAC, another 4 test video clips (Clip 5-8) are chosen, including 3 indoor video clips from PETS2006 and 1 outdoor video clip from CAVIAR [35]. Clip 5 to Clip 8 are captured by stationary cameras without any zooming. Besides, Clip 5 to Clip 7

have 600 frames at a frame rate of 30fps, and have resolution of 720x576. Clip 8 has 200 frames at the same frame rate and has resolution of 800x600.

4.1 Rate-Analytical-Distortion Performance

The RAD curve comparison of VAC against HEVC on each video clip is shown in Fig. 7. It is worth noting that the bit-streams generated by the proposed scheme is still HEVC compliant as none of the syntax structures is changed in our proposed scheme.

As a first observation in Fig. 7, it can be seen that VAC can reduce the analytical distortion effectively. In the best scenario (Fig. 7(c)), up to 40% reduction in terms of average analytical distortion can be achieved. Meanwhile, Fig. 8 illustrates the detection results of VAC and HEVC. It can be seen that the detection results of VAC are better than that of HEVC. Due to the proposed QP adaptation strategy, the object area is compressed by a relatively smaller QP, which makes the reconstruction quality higher compared with that in HEVC. This could help improve the performance of object detection algorithm.



(a) Clip 5 (b) Clip 6

(c) Clip 7 (d) Clip 8

Fig. 7 RAD curve comparison of VAC against HEVC in LDP configuration

As a second observation, it should be pointed out that the RAD performance of VAC is a little bit worse than that of HEVC under low bit rate in Fig. 7(d). From Fig. 8 (see the third row), it can be seen that the performance of VAC may not be satisfactory when the object is small. There are two main reasons. First, prediction errors exist in the prediction steps for the estimation of SAD and the analytical distortion, which will negatively influence the performance of VAC. Second, due to the frame subtraction algorithm used in this paper, small objects could not be detected in some cases. The resulting undetected object will be compressed by a relatively lager QP, which will decrease the reconstruction quality and further

influence the object detection algorithm.

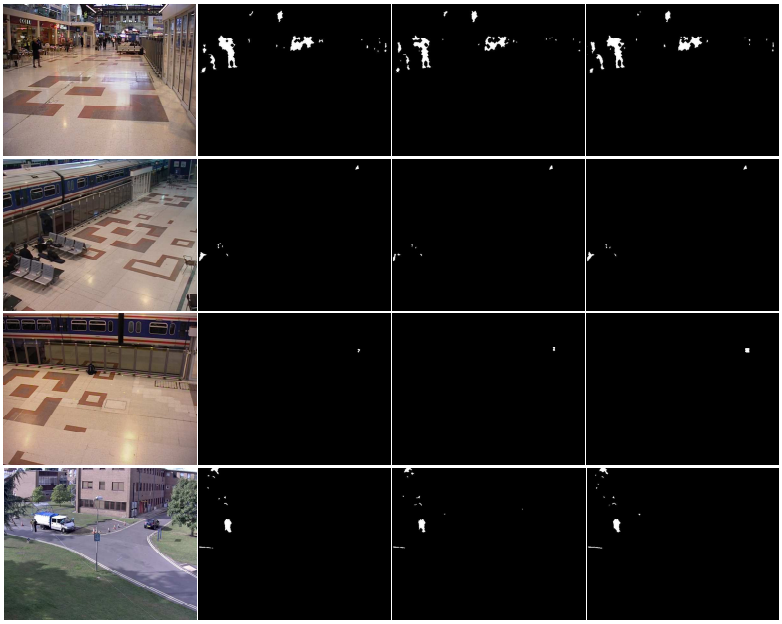


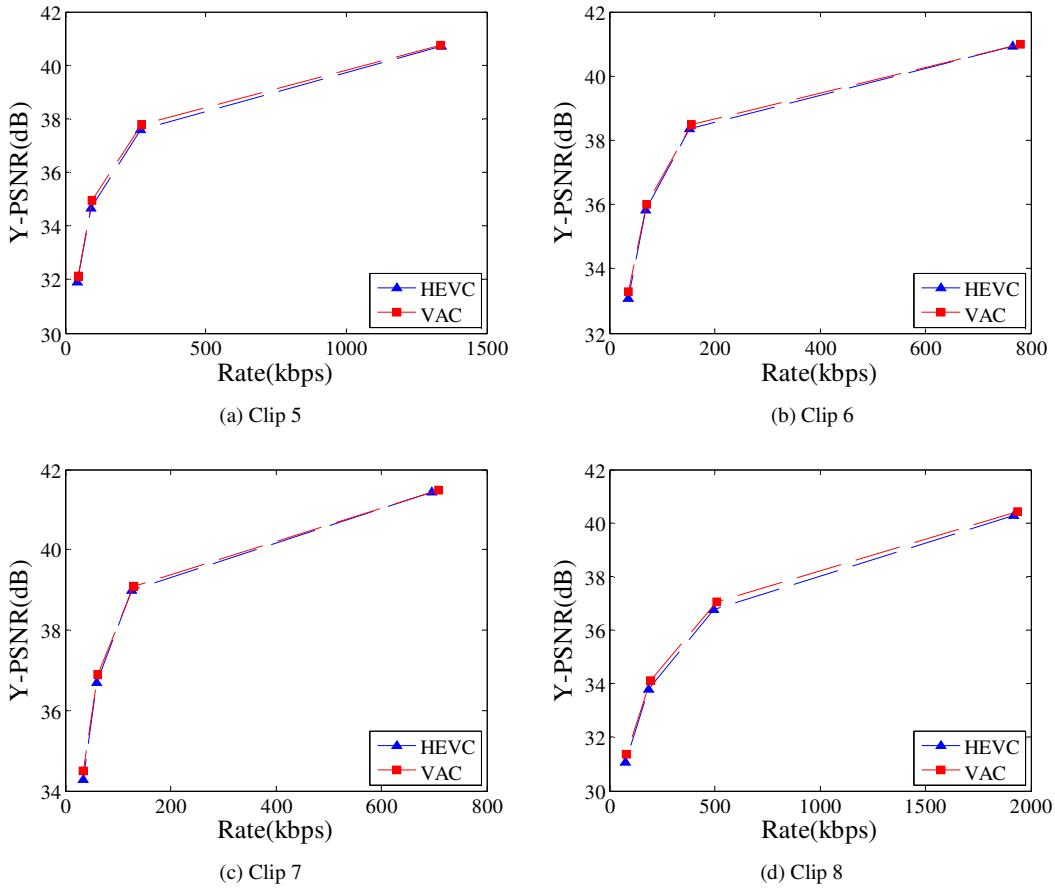
Fig. 8 Illustration for the detection results of four test video clips. The first column shows the thumbnails of Clip 5-8 respectively. The second column shows the detection results of the original frames. The third and the fourth column denote the detection results on their compressed versions when using HM 16.7 and VAC respectively.

4.2 Rate-Distortion Performance

Table 1 illustrates the RD performance gain of VAC over HEVC in terms of BD-rate saving. It can be observed that VAC can achieve about 4.2% BD-rate savings in average over the HEVC HM 16.7, with over 4.9% and 6.3% BD-rate savings for Clip 5 and Clip 8 respectively. Fig. 9 shows the RD curve comparisons on four test video clips. It can be observed that VAC can achieve better RD performance that HEVC. The main reason is that our proposed object based QP adaptation strategy refines the QP of an I frame as well as that of the background and the object area.

Table 1 RD Performance gain of VAC over HEVC in terms of BD-rate saving

Video Clip	**BD-Rate (%)**
Clip 5	-4.9
Clip 6	-2.5
Clip 7	-3.0
Clip 8	-6.3
Average	-4.2



(a) Clip 5 (b) Clip 6

(c) Clip 7 (d) Clip 8

Fig. 9 RD curve comparison of VAC against HEVC in LDP configuration

4.3 Complexity Evaluation

Table 2 shows the complexity comparison between the HEVC HM 16.7 and the proposed scheme in terms of encoding time under LDP coding structure. The test is done with Intel Core i5-4570 CPU and only one core is used to run the programs for making the running of all sequences under a similar workload. Encoding time for each clip is averaged over four test QPs. From the experimental results, it can be observed that the proposed scheme requires a slightly increased encoding time compared with HM 16.7, since some additional time is spent on the additional procedures such as frame subtraction and compression distortion prediction. However, this increase remains reasonable.

Table 2 Complexity comparison between VAC and HEVC in terms of encoding time (sec.)

Encoding Time (Sec.)	Clip 5	Clip 6	Clip 7	Clip 8
HEVC	4106.001	4384.378	3905.945	1486.276
VAC	4241.163	4472.652	3980.182	1537.996

5 Conclusion

In this paper, we propose a new coding framework called video analytical coding for video analysis. We use the term “analytical distortion” to denote the difference of video analysis performance when video quality degrades and develop a new rate-analytical-distortion optimization (RADO) method. Typically, analytical distortion is estimated by compression distortion. To show the effectiveness of our proposed method, we consider moving object detection as the analysis task and develop a novel rate

analytical distortion (RAD) model for video coding, where the analytical distortion is related to the object detection performance represented as F1-measure. Experimental results show that the performance of the video analysis task can be significantly improved. In our future work, we will extend our framework to other scenarios such as object detection in camera motion case, tracking and recognition. Furthermore, we will also focus on the parameter adaptation, including for the weighting parameter and QP offset.

6 Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 61571102 and in part by the Other Funds under Grant 2015AA015903 and Grant ZYGX2014Z003.

7 Reference

[1] 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., 13(7) (2003), pp. 560-576.

[2] G. J. Sullivan, J. Ohm, W. J. Han and T. Wiegand, “Overview of the high efficiency video coding (HEVC) standard,” IEEE Trans. Circuits Syst. Video Technol., 22(12) (2012), pp. 1649-1668.

[3] P. Korshunov and W. T. Ooi, “Critical video quality for distributed automated video surveillance,” ACM Multimedia, 6-11 Nov. 2005, Singapore, pp. 151-160.

[4] Y. Liu, C. Zhu, M. Mao, F. Song, F. Dufaux, X. Zhang, “Analytical distortion aware video coding for computer based video analysis,” IEEE Workshop on Multimedia Signal Processing (MMSP), 16-18 Oct. 2017, London-Luton, UK, in press.

[5] B. Li, D. Zhang, H. Li, J. Xu, “QP determination by lambda value,” JCTVC-I0426, Geneva, May 2012.

[6] B. Li, J. Xu, D. Zhang, and H. Li, “QP refinement according to Lagrange multiplier for high efficiency video coding,” IEEE International Symposium on Circuits and Systems (ISCAS), 19-23 May 2013, Beijing, China, pp.477-480.

[7] H. Zeng, K. N. Ngan, and M. Wang, “Perceptual adaptive Lagrangian multiplier for high efficiency video coding,” IEEE Picture Coding Symposium (PCS), 8-11 Dec. 2013, San Jose, California, USA, pp. 69-72.

[8] Y. Gao, C. Zhu, S. Li, T.W. Yang, “Temporally dependent rate-distortion optimization for low-delay hierarchical video coding,” IEEE Trans. Image Process., 26(9) (2017), pp. 4457-4470.

[9] S. Li, C. Zhu, Y.B. Gao, Y.M. Zhou, F. Dufaux, M.T. Sun, "Lagrangian multiplier adaptation for rate-distortion optimization with inter-frame dependency," IEEE Trans. Circuits Syst. Video Technol., 26(1) (2016), pp. 117-129.

[10] L. Itti, “Automatic foveation for video compression using a neurobiological model of visual attention,” IEEE Trans. Image Process., 13(10) (2004), pp. 1304–1318.

[11] Y. Liu, Z. G. Li, and Y. C. Soh, “Region-of-interest based resource allocation for conversational video communication of H.264/AVC,” IEEE Trans. Circuits Syst. Video Technol., 18(1) (2008), pp. 134–139.

[12] B. Xiong, X. Fan, C. Zhu, X. Jing, Q. Peng, "Face Region Based Conversational Video Coding," IEEE Trans. Circuits Syst. Video Technol., 21(7) (2011), pp. 917-931.

[13] W. B. Zhao, J. J. Fu, Y. Lu, S. P. Li, D. B. Zhao, “Region-of-interest based coding scheme for synthesized video,” Visual Communications and Image Processing (VCIP), 13-16 Dec. 2015, Singapore, pp. 1-4.

[14] M. Tiwari and P. C. Cosman, “Selection of long-term reference frames in dual-frame video coding using simulated annealing,” IEEE Signal Process. Lett., 15 (2008), pp. 249–252.

[15] G. Pushkar and B. Amrutur, “Skip decision and reference frame selection for low-complexity H. 264/AVC surveillance video coding,” IEEE Trans. Circuits Syst. Video Technol., 24(7) (2014), pp. 1156–1169.

[16] M. Paul, W. Lin, C. Lau and B. Lee, “A long-term reference frame for hierarchical B-picture-based video coding,” IEEE Trans. Circuits Syst. Video Technol., 24(10) (2014), pp. 1729–1742.

[17] X. Zhang, Y. Tian, T. Huang, S. Dong, and W. Gao, “Optimizing the hierarchical prediction and coding in HEVC for surveillance and conference videos with background modeling,” IEEE Trans. Image Process., 23(10) (2014), pp. 4511–4526.

[18] F. D. Chen, H. Q. Li, L. Li, D. Liu, F. Wu, “Block-composed background reference for high efficiency video coding,” IEEE Trans. Circuits Syst. Video Technol., 27(12) (2016), pp. 2639-2651.

[19] A. Redondi, L. Baroffio, M. Cesana, and M. Tagliasacchi, “Compress-then-analysis vs. analysis-then-compress: Two paradigms for image analysis in visual sensor networks,” in IEEE International Workshop on Multimedia Signal Processing (MMSP), 30 Sep. - 2 Oct. 2013, Santa Margherita di Pula, Sardinia, Italy, pp. 278-282.

[20] L. Baroffio, M. Cesana, A. Redondi, M. Tagliasacchi, and S. Tubaro, “Coding visual features extracted from video sequences,” IEEE Trans. Image Process., 23(5) (2014), pp. 2262-2276.

[21] L. Baroffio, J. Ascenso, M. Cesana, A. Redondi, and M. Tagliasacchi, “Coding binary local features extracted from video sequences,” in IEEE International Conference on Image Processing (ICIP), 27-30 Oct. 2014, Paris, France, pp. 2794-2798.

[22] J. Chao, E. Steinbach, “Keypoint encoding for improved feature extraction from compressed video at low bitrates,” IEEE Trans. Multimed., 18(1) (2016), pp. 25-39.

[23] E. Kokiopoulou and P. Frossard, “Semantic coding by supervised dimensionality reduction,” IEEE Trans. Multimed., 10(5) (2008), pp. 806-818.

[24] L. Liao, R. Hu, J. Xiao, G. Zhan, Y. Chen, J. Xiao “An analysis-oriented ROI based coding approach on surveillance video data,” the 17th Pacific-Rim Conference on Multimedia (PCM), 15-16 Sept. 2016, Xi’an, China, pp. 428-438.

[25] K. Yang, S. Wan, Y. Gong, H. Wu, Y. Feng, “An efficient Lagrangian multiplier selection method based on temporal dependency for rate-distortion optimization in H.265/HEVC,” Signal Process. Image Commun., doi: 10.1016.j.image.2017.05.006.

[26] Fraunhofer Heirich Hertz Institute, https://hevc.hhi.fraunhofer.de/ (accessed 26 Nev 2016).

[27] G. J. Sullivan and T. Wiegand, “Rate-distortion optimization for video compression,” IEEE Signal Process. Mag., 15(11) (1998), pp. 74-99.

[28] A. Ortega and K. Ramchandran, “Rate-distortion methods for image and video compression: An overview,” IEEE Signal Process. Mag., 15(11) (1998), pp. 23-50.

[29] H. Everett, “Generalized Lagrange multiplier method for solving problems of optimum allocation of resources,” Operation Research, 11(3) (1963), pp. 399-417.

[30] D. Thirde, L. Li, F. Ferryman, “Overview of the PETS2006 challenge,” the Proceeding of 9th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS), 18 Jun. 2006, New York, USA, pp. 47-50.

[31] J. Ferryman, A. Shahrokni, “Pets2009: dataset and chanllenge,” IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS), 5-7 Aug. 2009, Seattle, WA, USA, pp. 1-6.

[32] B. Lei, A. Leonardis, and B. Schiele, “Robust object detection with interleaved categorization and segmentation,” J. Comput. Vis., 77(1) (2008), pp. 259-289.

[33] D.M.W. Powers, “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness & correlation,” J. Mach. Learn. Technol., 2(1) (2011), pp. 37-63.

[34] Y. Dhome, N. Tronson, A. Vacavant, T. Chateau, C. Gabard, Y. Goyat, D. Gruyer, “A benchmark for background subtraction algorithms in monocular vision: a comparative study,” International Conference on Image Processing, Theory, Tools and Applications (IPTA), 7-10 Jul. 2010, Paris, France, pp. 66-71.

[35] http://homepages.inf.ed.ac.uk/rbf/CAVIAR/ (accessed 22 September 2016).