A Foreground-Background-Based CTU $\lambda$ Decision Algorithm for HEVC Rate Control of Surveillance Videos

Zhenglong YANG\(^a\), Member, Guozhong WANG\(^1\), and GuoWei TENG\(^{††}\), Nonmembers

SUMMARY Although HEVC rate control can achieve high coding efficiency, it still does not fully utilize the special characteristics of surveillance videos, which typically have a moving foreground and relatively static background. For surveillance videos, it is usually necessary to provide a better coding quality of the moving foreground. In this paper, a foreground-background CTU $\lambda$ separate decision scheme is proposed. First, low-complexity pixel-based segmentation is presented to obtain the foreground and the background. Second, the rate distortion (RD) characteristics of the foreground and the background are explored. With the rate distortion optimization (RDO) process, the average CTU $\lambda$ value of the foreground or the background should be equal to the frame $\lambda$. Then, a separate optimal CTU $\lambda$ decision is proposed with a separate $\lambda$ clipping method. Finally, a separate updating process is used to obtain reasonable parameters for the foreground and the background. The experimental results show that the quality of the foreground is improved by 0.30 dB in the random access configuration and 0.45 dB in the low delay configuration without degradation of either the rate control accuracy or whole frame quality.

key words: HEVC, segmentation, surveillance videos, $\lambda$ decision

1. Introduction

$\lambda$ domain rate control\(^1\) is an important part in high-efficiency video coding (HEVC). For the early $\lambda$ domain rate control, the bit rate is allocated according to the coding tree unit (CTU) prediction error. Then, the $\lambda$ domain rate control is improved by the adaptive bit allocation scheme\(^2\). As the surveillance videos are often captured by stationary cameras that always address the same scene\(^3\), the background is a representation of the scene with no moving objects, and the foreground typically arises from regional variation during the different adjacent frames.

The conventional rate control in HEVC clips involve every CTU $\lambda$ by the current frame $\lambda$. This approach neither improves the coding quality of the foreground nor reduces the bit rate from the background in a frame. Intuitively, it is necessary to exploit the special characteristics of the surveillance videos. A reasonable approach is to generate a high-quality background reference\(^4\). Chen et al.\(^4\) study the RDO optimization of the background reference CTUs without exploiting the RD characteristics of the foreground CTUs; in their work, the foreground CTU is extracted from the trained background frame, which inevitably brings coding complexity. Li et al.\(^5\) study the bit allocation scheme for the foreground and background separately, but the most important parameter ($\lambda$) is not considered carefully. The authors of\(^6\) calculate the $\lambda$ values for ROI and non-ROI with the independent bit rate allocation. However, there is no theoretical basis to verify the efficiency, and the calculated $\lambda$ may not be optimal without limitation.

In this paper, the foreground is extracted directly from the temporal adjacent frames to reduce complexity. Then, as the RD characteristics of the foreground and background differ but both obey the hyperbolic model, a separate optimal $\lambda$ decision for the foreground and background CTU with the RDO process of a frame is proposed. Finally, the parameters of the foreground and background are updated separately to satisfy regional variation.

2. Low-Complexity Pixel-Based Segmentation Algorithm

We use four continuous frames for the segmentation, which are the current coding frame and its three forward frames. The mean global forward difference (MGFD) between the current frame and the $i$-th forward frame is defined as

$$\text{MGFD}(i) = \frac{1}{W \cdot H} \cdot \sum_{x=1}^{W} \sum_{y=1}^{H} |p_{x,y} - \overline{p}_{x,y}|$$

(1)

where $W$ and $H$ are the width and height of the current frame, respectively. $(x,y)$ is the pixel location in the frame. $p_{x,y}$ and $\overline{p}_{x,y}$ are the pixel values of the frame and the collocated pixel values of the $i$-th $(i = 1, 2, 3)$ forward frame, respectively. Then, whether the pixel is associated with the foreground or the background is determined by

$$\text{FoB}_{x,y} = S(1)_{x,y} \& S(2)_{x,y} \& S(3)_{x,y}$$

(2)

where

$$S(i)_{x,y} = \begin{cases} 1, & |p_{x,y} - \overline{p}_{x,y}| \geq \text{MGFD}(i) \\ 0, & |p_{x,y} - \overline{p}_{x,y}| < \text{MGFD}(i) \end{cases}$$

(3)

If $\text{FoB}_{x,y} = 1$, the current pixel belongs to the foreground; otherwise, it is a background pixel. Finally, whether the current CTU belongs to the foreground or the background is determined by

$$\text{Flag}_{\text{CTU}} = \begin{cases} 1, & \sum_{i=1}^{n-1} \sum_{j=1}^{n} \text{FoB}_{x,y} \geq \omega \\ 0, & \sum_{i=1}^{n-1} \sum_{j=1}^{n} \text{FoB}_{x,y} < \omega \end{cases}$$

(4)
where \( n \) is the width of the CTU and \( \omega \) is set empirically as 1000. When \( \text{Flag}_{\text{CTU}} = 1 \), the current CTU belongs to the foreground; otherwise, it is a background CTU. The FoB maps and the \( \text{Flag}_{\text{CTU}} \) maps are shown in Fig. 1, in which the black regions are the foreground and the white regions are the background.

### 3. Separate CTU \( \lambda \) Decision Algorithm

We ran the HEVC encoder numerous times without rate control to observe the RD characteristics of the foreground and the background in random access (RA) and low delay (LD) configurations. The experimental results are shown in Fig. 2.

Regardless of the RA or LD configuration, the RD characteristics of the foreground and the background obey the hyperbolic model, and the characteristics of the foreground and background are different from each other. To minimize the distortion of a frame, the RDO problem can be modeled as

\[
\min_{(d_{\text{CTU}}(i))_{i=1}^{N_{\text{CTU}}}} D_F = \sum_{i=1}^{N_{\text{CTU}}} d_{\text{CTU}}(i) \\
\text{s.t.} \quad R_F = \sum_{i=1}^{N_{\text{CTU}}} r_{\text{CTU}}(i) \leq T_F \tag{5}
\]

where \( D_F \) and \( R_F \) are the distortion and the bit rate of the current frame, respectively. \( d_{\text{CTU}}(i) \) and \( r_{\text{CTU}}(i) \) are the distortion and the bit rate of the \( i \)-th CTU, respectively. \( N_{\text{CTU}} \) is the number of CTUs in the current frame. \( T_F \) is the target bit rate of the frame. Then, the Lagrange multiplier method can be converted to an equivalent unconstrained problem as

\[
\min_{(d_{\text{CTU}}(i))_{i=1}^{N_{\text{CTU}}}} \sum_{i=1}^{N_{\text{CTU}}} d_{\text{CTU}}(i) + \lambda_F \sum_{i=1}^{N_{\text{CTU}}} r_{\text{CTU}}(i) \tag{6}
\]

where \( \lambda_F \) is the Lagrange multiplier of the current frame.

Equation (6) can be solved by setting its derivative to zero:

\[
\frac{\partial}{\partial z} \sum_{i=1}^{N_{\text{CTU}}} d_{\text{CTU}}(i) + \lambda_F \frac{\partial}{\partial z} \sum_{i=1}^{N_{\text{CTU}}} r_{\text{CTU}}(i) = 0 \tag{7}
\]

where \( z \in [1, N_{\text{CTU}}] \). Whether the CTU belongs to the foreground or the background can be considered separately, so

\[
\sum_{i=1}^{N_{\text{CTU}}} \frac{\partial}{\partial z} d_{\text{CTU}}(i) + \lambda_F \frac{\partial}{\partial z} r_{\text{CTU}}(i) = 0 \tag{8}
\]

where \( \bar{N} + \tilde{N} = N_{\text{CTU}} \). \( \bar{N} \) and \( \tilde{N} \) are the number of the CTUs belonging to the foreground and the background, respectively. Then, every CTU in a frame is encoded separately. Equation (8) can be converted equally as

\[
\sum_{u=1}^{\bar{N}} \frac{\partial}{\partial z} d_{\text{CTU}}(u) + \lambda_F \frac{\partial}{\partial z} r_{\text{CTU}}(u) = 0 \tag{9}
\]

\[
\sum_{h=1}^{\tilde{N}} \frac{\partial}{\partial z} d_{\text{CTU}}(h) + \lambda_F \frac{\partial}{\partial z} r_{\text{CTU}}(h) = 0
\]

taking \( \lambda_{\text{CTU}}(i) = -\frac{\partial d_{\text{CTU}}(i)}{\partial r_{\text{CTU}}(i)} \) into Eq. (9). Finally, we can obtain

\[
\bar{N} \sum_{u=1}^{\bar{N}} \lambda_{\text{CTU}}(u) = \lambda_F \sum_{u=1}^{\bar{N}} \frac{\partial d_{\text{CTU}}(u)}{\partial r_{\text{CTU}}(u)} \tag{10}
\]

where Eq. (10) indicates that the mean \( \lambda \) of the CTUs that belong to the foreground or the background should be equal to \( \lambda \) of the current frame. As the RD characteristics of the foreground and the background are different from each other, the current CTU \( \lambda \) should be clipped separately. This term is calculated by
Fig. 3 Bit fluctuations of the foreground and the background with QP = 32 in the RA and LD configurations

Table 1 Values of $\delta_\alpha$ and $\delta_\beta$

<table>
<thead>
<tr>
<th>Target type</th>
<th>$\delta_\alpha$ RA</th>
<th>$\delta_\alpha$ LD</th>
<th>$\delta_\beta$ RA</th>
<th>$\delta_\beta$ LD</th>
</tr>
</thead>
<tbody>
<tr>
<td>App=0.03, g=0.08</td>
<td>0.05</td>
<td>0.02</td>
<td>0.01</td>
<td>0.005</td>
</tr>
<tr>
<td>0.015, g=0.2</td>
<td>0.1</td>
<td>0.1</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>0.2</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>0.5, g=0.2</td>
<td>0.4</td>
<td>0.4</td>
<td>0.4</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Clip(max($\sum_{i=1}^{u-1} \lambda_{CTU}(i)/(u - 1) \cdot 2^{\frac{1}{2}}$, $\lambda_{CTU}$), max($\sum_{i=1}^{h-1} \lambda_{CTU}(i)/(h - 1) \cdot 2^{\frac{1}{2}}$, $\lambda_{CTU}$))  

min($\sum_{i=1}^{u-1} \lambda_{CTU}(i)/(u - 1) \cdot 2^{\frac{1}{2}}$, $\lambda_{CTU}$), min($\sum_{i=1}^{h-1} \lambda_{CTU}(i)/(h - 1) \cdot 2^{\frac{1}{2}}$, $\lambda_{CTU}$))  

In Eq. (11), the current CTU $\lambda$ is affected not only by the frame $\lambda$ but also by the mean CTU $\lambda$, which belongs to the foreground or the background of the encoded CTUs. Thus, Eq. (10) can be approximately satisfied.

After encoding the current CTU, $\alpha_{CTU}$ and $\beta_{CTU}$ are updated. We also give the bit fluctuations of the foreground and the background without rate control. The results are shown in Fig. 3.

As shown in Fig. 3, the bits of the foreground fluctuate more greatly than those of the background. Thus, the updating parameters of the foreground should be larger than those of the background to follow the bit fluctuations. The updating parameters are calculated by

$$
\alpha_{CTU}^{new} = \alpha_{CTU}^{old} + \delta_\alpha \cdot (\ln \lambda_{CTU}^{real} - \ln \lambda_{CTU}^{cal}) \cdot \alpha_{CTU}^{old} \\
\beta_{CTU}^{new} = \beta_{CTU}^{old} + \delta_\beta \cdot (\ln \lambda_{CTU}^{real} - \ln \lambda_{CTU}^{cal}) \cdot \ln b_{pp\text{CTU}} 
$$

where the values of $\delta_\alpha$ and $\delta_\beta$ belonging to the foreground or the background are determined as addressed in Table 1.

4. Experimental Results

Because the rate control algorithm proposed in [1] allocates the bit rate based on the CTU prediction error, it assigns more bits to the foreground and less bits to the background. This approach is suitable for the characteristics of the surveillance videos. Thus, the default rate control of HM16.9 and the rate control of Li [1] are used as the comparison algorithms. The text sequences, which are basketballDrill_832 x 480, PKU_classsover_720 x 576, PKU_overbridge_720 x 576 and PKU_crossedroad_720 x 576, are used for the experiment. The bit rate accuracy is defined as

$$
M = \frac{|T_{\text{actual}} - T_{\text{target}}|}{T_{\text{target}}} 
$$

where $T_{\text{actual}}$ is the actual bit rate and $T_{\text{target}}$ is the target bit rate. The bit rate accurate results are shown in Table 3.

Table 3 shows the bit rate accurate values of the proposed algorithm, HM16.9 and Li [3], are 0.091%, 0.054% and 0.081% in the RA configuration and 0.002% and 0.002% and 0.118% in the LD configuration. Thus, the bit rate accuracies of all the rate control algorithms can be maintained at a high level.

The PSNRY indexes of the proposed rate control, HM16.9 and Li [1], are shown in Table 2, in which the PSNRY indexes of the foreground, the background, and the whole frame are expressed by F, B and W, respectively. The rate control of HM16.9 is used as the anchor to calculate the $\Delta$ values, which are shown in the last row of the table. Table 2 shows that the foreground qualities of the proposed algorithm and the Li [1] are improved in the RA and LD configurations. However, Li [1] decreases the coding qualities of the background and the whole frame, while the proposed algorithm maintains the coding qualities. As Li [1] allocates the bit rate by the CTU prediction error, more bit rates are assigned to the foreground CTUs, and the coding quality of the whole frame decrease. On the other, the proposed rate control always selects the optimal CTU $\lambda$ for every CTU with the separate CTU $\lambda$ decision and the separate parameter updating. As a result, the foreground and background quality are improved.

Figure 4 shows the bit fluctuations for every frame in the RA and LD configurations, in which the bit fluctuation of Li [1] varies more largely than the other cases. As Li [1] allocates the bit rate according to the CTU prediction error, although the proposed algorithm and HM16.9 allocate the bit rate adaptively, the bit fluctuations of the proposed algorithm and HM16.9 are similar to each other and steadier than that of Li [1].

Figure 5 shows the subjective comparisons of the proposed algorithm, HM16.9 and Li [1]. In Fig. 5 (c-1), there is a blocking artifact on the left foot. However, in Fig. 5 (b-1) and Fig. 5 (d-1), the same regions are clearer. The same situation appears in Fig. 5 (c-2), where the left elbow is missing some part, while they are relatively complete in Fig. 5 (b-2) and Fig. 5 (d-2). As mentioned above, the proposed algorithm and Li [1] both improve the qualities of the fore-
ground. However, Li [1] sacrifices the whole frame coding quality for improving the foreground, and the proposed algorithm maintains the whole frame coding quality. As a result, the proposed algorithm has the best rate control performance with the others on the surveillance videos.

5. Conclusion

In this paper, a separate CTU λ decision algorithm for surveillance videos is proposed. The proposed algorithm mainly improves the coding quality of the foreground and maintains the coding quality of the whole frame. First, we use low-complexity pixel-based segmentation to separate the foreground and the background in a frame. Then, with the RDO process according to the different characteristics of the foreground and the background, the separate CTU λ clipping method is used to obtain the optimal λ values of the foreground and background CTUs. Finally, a separate updating process is proposed to obtain reasonable rate control parameters for all the CTUs. The experimental results show that the proposed algorithm improves the coding quality of the foreground significantly without quality degradation of the whole frame.

References

