A NOVEL PATTERN IDENTIFICATION SCHEME USING DISTRIBUTED VIDEO CODING CONCEPTS

Manoranjan Paul and Manzur Murshed
Gippsland School of Information Technology, Monash University, Churchill, Vic-3842, Australia
E-mail: \{Manoranjan.paul, Manzur.Murshed\}@infotech.monash.edu.au

ABSTRACT

Pattern-based video coding focusing on moving region in a macroblock has already established its superiority over recent H.264 video coding standard at very low bit rate. Obviously, a large number of pattern templates approximate the moving regions better however, after a certain limit no coding gain is observed due to the increase number of pattern identification bits. Recently, distributed video coding schemes used syndrome coding to predict the original information in decoder using side information. In this paper a novel pattern identification scheme is proposed which predicts the pattern from the syndrome codes and side information in decoder so that actual pattern identification number is not needed in the bitstream. The experimental results confirm that this new scheme successfully improves the rate-distortion performance compared to the existing pattern-based video coding as well as H.264 standard. This new scheme will also open another window of syndrome coding application.

Index Terms—Video coding, pattern recognition, side information, distributed video coding, H.264, low bit rate.

1. INTRODUCTION

Applications such as video conference, video call, and Internet video streaming through limited bandwidth channel require very low bit rate (VLBR) video coding technique. The recent block-based video coding standard H.264/AVC [1] has introduced variable block sizes (\(16 \times 16, 16 \times 8, 8 \times 16, 8 \times 8, 4 \times 4, 4 \times 8, 8 \times 4\) pixels) motion estimation (ME) and motion compensation (MC) strategy to capture various motions within a \(16 \times 16\) pixels macroblock (MB). However, this strategy failed to exploit intra-macroblock temporal redundancy (ITR) (see Fig 1) which is static in successive frames. Indeed, few bits are used to signal zero residual error and zero motion for ITR to decoder, which is obviously significant when a sequence is encoded at VLBR. Pattern-based video coding (PVC) [2]-[5] exploited the ITR using binary patterns (see Fig 1) in partitioning the suitable MB via a simplified segmentation process that avoided handling the exact shape of the moving objects, so that popular MB-based ME&MC techniques could be applied. ME and MC are carried out using only pattern covered region. Thus, a significant compression is achieved compared to the H.264 as the pattern size is one-fourth of the original MB. A pattern codebook (PC) of 32 pattern templates is given in Fig 2.

The moving region (MR) of an MB has different shape as it depends on the part of object contained in the MB. Obviously, a large number of pattern templates approximate the MR more accurately but, after a certain limit no coding gain is observed due to the increase number of pattern identification codes (PIC). Variable length PIC using pattern frequency information and pattern co-occurrence matrix [4] used 4.62 and 4.15 bits per pattern represented MB (which is popularly know as region-active MB (RMB)) instead of 5 bits for 32 patterns respectively. Hence, any strategy that reduces the PIC concomitantly improves the overall encoding performance by classifying more RMBs.

Fig 1: An example on how pattern-based coding can exploit the intra-macroblock temporal redundancy (ITR) to improve coding efficiency.

Very recently, computational complexity distribution of the existing video coding standards modeled with a heavy encoder and light decoder is challenged by a stark contrast distributed video coding (DVC) architecture [6][7][8][9] which is modeled with a light encoder and a heavy decoder. The DVC exploits the joint source-channel coding theorems developed by Slepian and Wolf [10] and Wyner and Ziv [11]. The main idea is to divide the source data into a finite number of sets, popularly known as cosets, and the original data can be retrieved in the decoder from the coset index and predicted data by applying nearest neighborhood technique. The predicted data is generated from the side information, for example, previously decoded data. If the maximum distance between actual and predicted data is \(E_{\text{max}}\) and the minimum distance of each coset at level \(i\) is \(d_{\text{min}}=2^{i-1}\), then co-set theorem states that the actual data will be retrieved without error if \(E_{\text{max}} < d_{\text{min}}/2\) is satisfied. In this case we need to divide data set into \(i+1\) cosets. Compression of the scheme stems from transmitting only the coset index (syndrome) instead of the actual value. Although the compression ratio of the DVC can theoretically be as efficient as that of the conventional heavy encoders, this upper bound can only be achieved with the accurate estimation of inter-frame correlation structures. In real videos however the correlation structures are highly spatially-varying, potentially unbounded in magnitude, and imprecisely known leading to significant imperfection. Moreover, being a block-based codec, this new technique is obviously unable to exploit the intra-block temporal redundancy as alluded for the existing standards. So far no distributed video coding schemes are comparable to the conventional coding in terms of quality.

In this paper, we use distributed coding concept to encode PIC of 32 patterns. If we applied this theorem straightforward to encode 32 patterns, we will not get any compression as the maximum prediction error is 16 for 32 patterns which require 32 cosets. Obviously smaller number of cosets will provide compression for those RMBs where accurate pattern retrieval is possible. Fig 2 shows all 32 patterns. It is interesting to observe...
that the gravitational centres of moving region of all patterns are almost uniformly distributed in 360° with respect to centre. We will exploit this angular distance for coset formation as the dissimilarity between two patterns is directly proportional to the angular difference between them. We will use side information from the previously decoded frames as we experimentally observed that there is a strong correlation between two successive co-located MBs in terms of pattern similarity. As the same predicted system can be designed in the both encoder and decoder, we will not lose any information if we discard those RMBs where pattern could not be retrieved accurately.

Fig 2: The pattern codebook of 32, 64-pixel patterns (numbering from left to right and top to bottom), defined in 16×16 blocks, where the white (moving) region represents 1 and black (static) region represents 0 [4].

In this paper a novel PIC scheme is proposed which successfully exploited distributed video coding concepts for retrieving the original pattern in decoder without knowledge of its index. Using this approach a significant reduction of the PIC intractable problem as we need to consider all 64

\( \times \) combinations. Thus, coset formation technique needs to be converted into a simple problem where the gravitational center of a pattern will represent corresponding whole pattern. Although, by doing this we could not utilize all available information in coset formation, it simplifies the formation step. It is easily observed that the gravitational centers (GCs) of patterns (considering only white moving region) are almost symmetrical around 360° (see Fig. 3).

Fig 3: Positions of gravitational centers of patterns in 16×16 block and the numbers indicate the patterns.

Table 1: Various size of cosets using angular distance and corresponding minimum distance among the cosets.

<table>
<thead>
<tr>
<th>Coset</th>
<th>Patterns</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>[19,14,5,32,3,16,24,26,12,1,7,27,18,10,22,29]</td>
<td>13°</td>
</tr>
<tr>
<td>4</td>
<td>[19,5,3,24,12,7,18,22] [30,14,3,12,5,29,28,13]</td>
<td>37°</td>
</tr>
<tr>
<td>8</td>
<td>[19,3,12,18] [30,31,25,28] [14,16,1,10] [21,8,23,6]</td>
<td>85°</td>
</tr>
<tr>
<td>16</td>
<td>[20,4] [3,18] [31,28] [16,10] [8,6] [24,22,2,13]</td>
<td>175°</td>
</tr>
</tbody>
</table>

The Table 1 shows the minimum angular distances 13°, 37°, 85°, and 175° using 2, 4, 8, and 16 cosets respectively. Theoretically the distances would be 22.5°, 45°, 90°, and 180°, but already designed patterns are not fully symmetrical around the circle.

3. SIDE INFORMATION GENERATION

If \( C(x,y) \) and \( R_y(x,y) \) denote the \( k^\text{th} \) MB of the current and reference frames respectively of a video sequence, where \( 0 \leq x, y \leq 15 \), the moving region \( M_y(x,y) \) of the \( k^\text{th} \) MB in the current frame is obtained by:

\[
M_y(x,y) = T(C_x(y) \bullet B - R_y(x,y) \bullet B) \quad (1)
\]

where \( B \) is a 3×3 unit matrix for the morphological closing operation \( \bullet \), which is applied to reduce noise, and the thresholding function \( T(y) = 1 \) if \( y > 2 \), and is 0 otherwise. If we generate a binary matrix for whole current frame with respect to reference frame, we will get moving region for entire current frame. Fig 4 shows two moving region between frame 1 & 3 in (a) and frame 3 & 5 in (b). If we carefully observed these two
moving regions we will see the similarity between two in terms of shape and size in MB level. From this observation we are motivated that side information could be generated from previously decoder successive two frames. Note that we consider alternative frames as we use 15 frames per second at rate-distortion calculation. Obviously more similarity will be observed between adjacent two frames.

![Fig 4: The moving region between frame 1 and 3 (a) and frame 3 and 5 (b) for Miss America video sequence using (1).](image)

This co-relation can be quantified using pattern similarity. Let $|Q|$ be the total number of ‘$s in the matrix Q. The dissimilarity of a pattern $P_s \in PC$ with the MR in the $k$th MB can be defined as:

$$D_{k,n} = |M_k| - |P_s \wedge M_k|.$$  

(2)

Clearly, higher the similarity lower will be the value of $D_{k,n}$. The best similar pattern $P_i$ will be selected for the MR $M_k$ as follows:

$$P_i = \arg \min_{s \in PC} D_{k,n}.$$  

(3)

4. PATTERN RETRIEVE

Sometimes after motion estimation the residual error is almost zero or less significant due to the smooth translational motion and coarse quantization. Moreover, zero motion vector is observed for a significant number of cases in the typical video sequences. These two features we can exploit in our pattern prediction scheme to increase the pattern retrieval accuracy. In the existing PVC scheme, only the pattern covered MR is considered for the ME&MC. The rest of the region of an MB is copied from the co-located MB as a static region from the decoded frame. When a pattern could not accurately approximate the MR of an MB, the pattern mode does not provide the best rate-distortion performance due to the pattern mismatch error. Thus, our proposal is to consider the whole MB instead of pattern covered MR for ME&MC. By doing this, we not only eliminate pattern miss-approximation error but also reduce the extra ME computational cost as we already have motion vector for the entire MB from the 16x16 mode. Thus, when residual error is zero after quantization, retrieving a wrong pattern through our prediction scheme could not affect in rate-distortion performance.

In our proposed scheme, we determine the best pattern for the current MB by (2) and (3). Then we find the coset index of this pattern. We also find the best pattern (i.e., side information) from the motion-translated co-located MB in the decoded frame. From the coset index and side information, we apply nearest neighborhood technique to find the predicted pattern. If the predicted pattern and the actual pattern (i.e., the best pattern for the current MB) are same pattern, we will encode this MB by the best pattern. Otherwise we will try with second/third/fourth best pattern at the same way. If no one is successful, we will encode this MB by other H.264 modes instead of pattern mode.

5. SIMULATION RESULTS

The existing PVC, proposed scheme (termed as PVCcosetX {2, 4, 8, 16}), and the H.264 algorithms are implemented based on the Baseline profile of H.264/AVC [12] with full search fractional pel motion estimation for the number of standard QCIF [12] video sequences at 15 frames per second (fps). The experimental results reveal that 85–100% correct decision can be achieved up to 20 quantization parameter (QP) [12] (see Fig 6) using all four best patterns. For obvious reason, the prediction is improving with the number of cosets (see Fig 6 (left)). Whereas only the best pattern can predict more than 60% patterns accurately when $QP = 20$ (see Fig 6 (right)). It is clear that, using four patterns instead of one will increase the computation, however, it can be reduced using already available motion estimation information of 16x16. Thus, proposed scheme is comparable with the existing PVC in terms of computational complexity. The number of RMBs is decreasing with bit rates (see Fig 7). Thus, rate-distortion gain using the proposed and the
PVC schemes have diminishing return with the bit rates compared to the H.264.

The new PVCosetX scheme reduces the bitstream by 29% (by 16 cosets) to 33% (by 2 cosets) and 8% (by 16 cosets) to 13% (by 2 cosets) on average (see Table 2) compared to the H.264 and the existing PVC which is significant especially at very low bit rate. Note that the PVC using more cosets performs better with bit rates as it accommodates more RMBs but the PVC with 4 cosets will provide reasonable performance for all bit rates. The PVC with 2 cosets will outperform in all cases if we can use better side information. A rate-distortion performance is also shown for Miss America video sequence using different cosets in Fig 8. The figure confirms that more than 0.25 dB PSNR is improved at 64 kbps using 2 cosets.

Table 2: Coding performance using PVCoset, PVC, and H.264

<table>
<thead>
<tr>
<th>Video</th>
<th>PSNR (dB)</th>
<th>Bit Rate (kbps)</th>
<th>PVCoset2</th>
<th>PVCoset4</th>
<th>PVCoset8</th>
<th>PVCoset16</th>
<th>H.264</th>
</tr>
</thead>
<tbody>
<tr>
<td>Miss America</td>
<td>36</td>
<td>14.5</td>
<td>15.5</td>
<td>16.5</td>
<td>17.0</td>
<td>19</td>
<td>33</td>
</tr>
<tr>
<td>Foreman</td>
<td>40</td>
<td>26.5</td>
<td>27.5</td>
<td>27.5</td>
<td>28.5</td>
<td>31</td>
<td>43</td>
</tr>
<tr>
<td>Claire</td>
<td>43</td>
<td>55.5</td>
<td>56.2</td>
<td>56.5</td>
<td>57.2</td>
<td>59</td>
<td>67</td>
</tr>
<tr>
<td>Salesman</td>
<td>44</td>
<td>69.8</td>
<td>70.2</td>
<td>70.6</td>
<td>71.2</td>
<td>73</td>
<td>79</td>
</tr>
<tr>
<td>Foreman</td>
<td>27</td>
<td>30.5</td>
<td>32.0</td>
<td>33.0</td>
<td>34.0</td>
<td>39</td>
<td>58</td>
</tr>
<tr>
<td>Claire</td>
<td>30</td>
<td>44.0</td>
<td>45.0</td>
<td>46.0</td>
<td>47.0</td>
<td>51</td>
<td>67</td>
</tr>
<tr>
<td>Salesman</td>
<td>32</td>
<td>64.0</td>
<td>65.0</td>
<td>66.0</td>
<td>67.0</td>
<td>71</td>
<td>84</td>
</tr>
<tr>
<td>Claire</td>
<td>34</td>
<td>80.5</td>
<td>81.0</td>
<td>81.2</td>
<td>82.0</td>
<td>86</td>
<td>98</td>
</tr>
</tbody>
</table>

Fig 7: Average number of RMBs with different QPs using eight standard video sequences.

7. REFERENCES