PATTERN BASED VIDEO CODING WITH UNCOVERED BACKGROUND

Manoranjan Paul, Weisi Lin, Chiew Tong Lau, and Bu-sung Lee

School of Computer Engineering, Nanyang Technological University, Singapore-639798, Singapore

E-mail: {m_paul, wslin, asctlau, ebslee}@ntu.edu.sg

ABSTRACT

The pattern-based video coding (PVC) outperforms the H.264 through better exploitation of block partitioning and partial block skipping. In the PVC scheme the best pattern is determined against the moving regions (MRs) in a macroblock (MB) of the current frame against the co-located MB in the reference frame; motion estimation (ME) and motion compensation (MC) are carried out using the pattern covered MRs, and the rest of the regions are treated as skipped areas. The MRs can be due to the object areas and the uncovered background (UCB) areas. Thus, the ME & MC by the pattern for the MRs of the UCB would not be accurate if there is no similar region in the reference frame. As a result no coding gain can be achieved for the UCB. Recently a dynamic background frame termed as the McFIS (the most common frame of a scene) has been generated using Gaussian mixture models for object detection. In this paper we propose a new PVC technique which will use the McFIS as a reference frame to determine the MRs where only object areas will be captured as the MRs. Thus, the proposed technique overcomes the mismatch problem of the UCB for ME&MC. The experimental results confirm the superiority of the proposed scheme in comparison with the existing PVC and McFIS-based methods by achieving significant image quality gain.

Index Terms—Video coding, uncovered background, light change, repetitive motion, H.264, motion estimation, pattern based video coding, and multiple reference frames.

1. INTRODUCTION

The latest video coding standard H.264 [1] outperforms its competitors such as the H.263, MPEG-2, MPEG-4, etc. due to a number of innovative features in the intra and inter-frame coding techniques. Variable block sizes (VBS) motion estimation (ME) and motion compensation (MC) is one of the most prolific features. In the VBS scheme, a 16×16-pixel macroblock (MB) is partitioned into small several rectangular or square shape blocks. ME & MC are carried out for all possible combinations, and the ultimate block size is selected based on the Lagrangian optimization (LO) using the bits and distortions of the corresponding blocks. The real world objects, by nature, may be in any arbitrary shapes, thus, the ME&MC using only rectangular/square shaped blocks may roughly approximate the real shape and thus the coding gain would not be satisfactory. A number of research works are conducted using non-rectangular block partitioning [2][3][4] using geometric shape partitioning, motion-based implicit block partitioning, and L-shaped partitioning. Requirement of excessively high computational complexity in the segmentation process and the marginal improvement over the H.264 make them less effective for real time applications. Moreover, the requirement of precious extra

bits for encoding the area covering almost static background makes the above mentioned algorithms inefficient in terms of rate-distortion performance.

To exploit the non-rectangular block partitioning and partial block skipping for static background area in an MB, the pattern-based video coding (PVC) [5][6][7] scheme partitions the MBs via a simplified segmentation process that again avoids handling the exact shape of the moving objects, so that the popular MB-based ME could be applied. The PVC algorithm focuses on the moving regions (MRs) of the MBs, through the use of a set of regular 64-pixel pattern templates (see Figure 1). The MR is defined as the difference between the current MB and the co-located MB of the reference frame. The pattern templates were designed using ‘1’s in 64 pixel positions and ‘0’s in the remaining 192 pixel positions in a 16×16-pixel MB. The MR of an MB is defined as a region comprising a collection of pixel positions where pixel intensity differs from its reference MB. Using some similarity measures, if the MR of a MB is found well covered by a particular pattern, then the MB can be classified as a region-active MB (RMB) and coded by considering only the 64 pixels of the pattern, with the remaining 192 pixels being skipped as static background. Embedding PVC in the H.264 standard as an extra mode provides higher compression for RMBs as larger segment with static background is coded with the partial skipped mode.

Figure 1: The pattern codebook of 32 regular shaped, 64-pixel patterns, defined in 16×16 blocks, where the white region represents ‘1’ (motion) and the black region represents ‘0’ (no motion) [6].

The MR generated from the difference of the current MB and the co-located MB from the *traditional* reference frames (i.e., the immediate previous frame or any frame which is previously encoded) may contain moving object and uncovered background (UCB) (detailed in Figure 2). The ME&MC using pattern covered MR for UCB would not be accurate if there is no similar region in the reference frames. As a result no coding gain can be achieved for the UCB using the PVC. Similar issues occur for any other H.264 VBS modes due to the lack of suitable matching region in the reference frames. Thus, we need a reference frame where we will find the UCB for the current MB if once that region was evidenced. Only a *true* background of a scene can be the best

choice to be the reference frame for UCB. Moreover, an MR generated from the true background against the current frame represents only moving object instead of both the moving object and the UCB. Thus the selection of the best matched pattern against the newly generated MR is the best approximation of the object/partial object in an MB. The ME&MC using the best matched pattern carried out on the immediate previous frame will provide more accurate motion vector and thus minimum residual errors for the object/partial object of the MB. The rest of the area (which is not covered by the pattern) is copied from the true background frame. The immediate previous frame is used for ME&MC assuming that the object is visible in the immediate previous frame. The other modes of the H.264 can also use true background as well as the immediate previous frame (in the multiple reference frames technique [1]) as two separate reference frames. The LO will pick the optimal reference frame.

A *true* background is not available in most cases. Thus we need to generate background from a video scene. Recently a dynamic background frame termed as the McFIS (the most common frame of a scene) [14] has been developed from the Gaussian mixture based *dynamic background modeling* (DBM) [11][12][13] for video coding. In this paper, the McFIS is to be used as a long term reference (LTR) frame [9][10] assuming that the background of the current frame will be referenced from the McFIS and the foreground will be referenced from the immediate previous frame. To be more specific, we will use the McFIS to generate new MR for the PVC and also use it as an LTR frame for both the PVC and the H.264 modes. The ultimate mode will be selected using the LO. The experimental results confirm that the proposed method outperforms two recent and relevant algorithms by improving significant image quality.

The rest of the paper is organized as follows: Section 2 describes the proposed coding scheme. Section 3 demonstrates the experimental set up and results. Section 4 concludes the paper.

2. PROPOSED SCHEME

The proposed scheme is based on the H.264 coding standard by embedding the pattern mode with two reference frames: one is the immediate previous frame, and another is the McFIS assuming that motion areas and normal/uncovered static areas will be referenced from the immediate previous frame and the McFIS respectively. The McFIS is generated by dynamically background modeling using Gaussian mixture models [13][14]. It is constructed from the decoded frames at the encoder and decoder using same technique so that it needs not to be transmitted from the encoder to the decoder. When a frame is decoded at the encoder/decoder, the McFIS is updated using the newly decoded frame. The detailed procedure will be described in Sub-section 2.2. To exploit the non-rectangular MB portioning and partial skipped mode, a pattern mode is incorporated as an extra mode into the conventional H.264 video coding standard, is defined as the PVC scheme [7]. Figure 1 has shown the pattern codebook (PC) [6] comprising 32 patterns which are used in the proposed scheme. Each pattern is a binary 16×16 matrix, where the white region indicates ‘1’ (i.e., foreground) and the black region indicates ‘0’ (i.e., background). Actually the pattern is used as a making to segment out the foreground from the background within a 16×16-pixel MB.

We need to determine MR for the current MB using the MBs from the current and reference frames. Then to find the best matched pattern from the pattern codebook through a *similarity metric* [7]. ME&MC are carried out using only pattern covered MR (i.e., the white region). In the proposed scheme we also introduce a

new pattern matching scheme for ME&MC so that we can overcome the occlusion problem in the existing schemes by exploiting uncovered background. The detailed procedures will be explained in the next sub-section.

2.1 New ME&MC for UCB areas

Let F_k^i and F_k^{i-1} be the k -th MBs from the i -th and $(i-1)$ -th frames respectively. According to the PVC scheme [7] the MR, M_k^i is defined as follows:

$$M_k^i(x, y) = |F_k^i(x, y) - F_k^{i-1}(x, y)|. \quad (1)$$

The similarity of a pattern $P_n \in PC$ (i.e., pattern codebook) with the MR in the k^{th} MB is defined as:

$$S_{k,n}^i(x, y) = \sum_{x=0}^{15} \sum_{y=0}^{15} |M_k^i(x, y) \times P_n(x, y)|. \quad (2)$$

The best matched pattern for a MR is then selected as:

$$P_j = \arg \max_{\forall P_n \in PC} (S_{k,n}^i). \quad (3)$$

Figure 2: Motion estimation and compensation problem using blocks or patterns when there is occlusion.

Figure 2 shows a current frame in (b), a reference frame in (a), the MR (marked as texture) according to Equation (1) in (c), and a true background without object (here a moving ball) in (e). In Figure 2 (b) C is the moving object, while C results from the MR (i.e., UCB) and corresponds to A in Figure 2 (c); there is no matched region for C in the reference frame (i.e., in Figure 2 (a)). Thus the PVC as well as the normal H.264 modes could not provide accurate ME&MC for C. This problem can be solved if we have generated true background (Figure 2 (e)) and ME&MC are carried out using the background as a reference frame using suitable H.264 mode or pattern mode (if the MR is best matched with any pattern). In this work, we use McFIS (actually a dynamic background frame) for referencing the UCB.

When a pattern is matching against the MR of B (in Figure 2 (c)) for the corresponding D in Figure 2 (b), ideally Pattern 11, 14, or 30 (see Figure 1) would be the best matched pattern but due to the MR for UCB Pattern 21 is the best matched pattern. However, ME&MC using Pattern 21 does not find proper reference region in any reference frames (i.e. Figure 2 (a) or (e)) and results in poor rate-distortion performance. To solve this problem we generate MR using the McFIS and the current frame (see Figure 2 (d)), then use the immediate previous frame for ME&MC using pattern covered region, and the rest of the region of the MB is copied from the co-located background (McFIS). In this process we need to replace F_k^{i-1} in (1) by k -th MB from McFIS (i.e., F_k^{McFIS}). We also use other two options using the traditional pattern matching

with Equation (1) and ME&MC, i.e., using the immediate previous frame (existing pattern matching) and using the McFIS (pattern matching using McFIS) to maximize the rate-distortion performance where a MR is not good matched by the best pattern.

Figure 3 (a) compares the average percentages of MBs selected by the LO as the reference MBs for three relevant techniques (i) the existing pattern matching [7] i.e., matching and ME&MC are carried out using immediate previous frame, (ii) pattern matching and ME&MC are carried out using the McFIS, and (iii) pattern matching using the McFIS but ME&MC is carried out using the immediate previous frame. The third technique is the newly introduced pattern matching and ME&MC approach while the second technique is for the multiple reference frame approach. The first 300 frames of six standard video sequences namely *Paris*, *Bridge Close*, *Silent*, *News*, *Salesman*, and *Hall Objects*, have used for the evaluation. The figure shows that the proposed technique (iii) selects the most number of MBs compared to the other two techniques. The higher percentage represents higher effectiveness for referencing. The results indicate that the proposed pattern matching and ME&MC technique is expected to perform better and this will be further evidenced by the rate-distortion performance in Section 3.

Figure 3: (a) Average percentages of MBs selected by the three techniques (i) the existing pattern matching, (ii) pattern matching using the McFIS, and (iii) the pattern matching using mixed (i.e., pattern matching using the McFIS but ME&MC is carried out using the immediate previous frame); (b) Percentages of areas where the McFIS and the LTR frames are referenced respectively.

2.2 McFIS generation

In a video scene, a pixel may be a part of different objects and backgrounds over the time. Each part can be represented by a Gaussian model expressed by pixel intensity variance, mean, and weight [11][12][13]. Thus, to model a pixel over the time Gaussian mixture models are used. Intuitively if a model has large weight and low variance, then most probably the model represents the most stable background. Based on this assumption a mean value of the best background model is taken as background pixel intensity for that pixel. In this way an entire background frame (i.e., McFIS) is constructed. Sometimes instead of the mean value, last satisfied pixel intensity (preserved when a pixel satisfies a model) is taken as the background pixel intensity to avoid artificial mean value [13]. As mentioned in [14], background generation using pixel mean (or pixel recent value) is not very effective in video coding applications as the McFIS is generated from the distorted image (i.e., decoded frame). To minimize this problem, neighboring pixel intensities within the McFIS (i.e., spatial correlation) are used to generate better McFIS [14]. In our experiment we have found that temporal correlation exploitation is even better to construct McFIS compared to the spatial correlation. Thus we modify the McFIS as follows assuming that D^i and D^{i-1} as the i th and $(i-1)$ th McFISes respectively

$$D^i(x, y) = \begin{cases} \tau D^i(x, y) + (1-\tau) D^{i-1}(x, y) & \text{if } |D^i(x, y) - D^{i-1}(x, y)| < T_p \\ D^i(x, y) & \end{cases} \quad (4)$$

where τ (<1) and T_p are the weighting factor and threshold respectively. It is obvious that there should be a strong correlation between consecutive McFISes especially in the stable region (i.e., background). If there is a small difference (i.e., T_p), we assume that it is due to the quantization error instead of different environments. Thus, to rectify this variance, a weighted average is formulated for the current McFIS from the previous McFIS. A large value of τ means we give emphasis for the current McFIS (i.e., recent changes compared to the learned model). In our implementation we use 0.5 and 10 for τ and T_p respectively.

2.3 Encoding and decoding

In the proposed scheme, the first frame of a video is encoded as an intra-frame, and the subsequent frames are encoded as inter-frames until scene change [8] occurs. When a frame is encoded and decoded at the encoder, the McFIS is updated using background modeling. When scene change occurs, the modeling parameters are reset and new McFIS will be generated. As the McFIS contains *stable* portion of a scene, sum of absolute difference (SAD) between the current frame and the McFIS is a good indicator for scene change. In this scheme a scene change is detected at i th frame if the ratio of SAD at i -th and $(i-1)$ -th frames is more than 1.7. For each MB, we have examined all modes including pattern mode using two reference frames, then the ultimate mode is selected based on the LO. Obviously in the proposed scheme we need more operations to generate the McFIS. But the experimental results confirm that it does not take more than 3% of the total encoding time.

3. OVERALL EXPERIMENTAL RESULTS

The ME&MC using multiple reference frames (MRFs) [1] is more effective in improving rate-distortion performance compared to the single reference frame for repetitive motion, uncovered background, non-integer pixel displacement, lighting change, etc. The number of reference frames in practical applications is limited due to the requirement of codes to identify the reference frames, computational time in ME which increases linearly with the number of reference frames, and memory buffer size also increases for storing decoded frames in both encoder and decoder. For trading-off, dual reference frames (long term reference frame and short term reference (STR) frame) are introduced [9][10] assuming that static regions and object regions (i.e., MRs) are referenced from the LTR and the STR frames respectively. When the i -th frame is being encoded, the $(i-1)$ -th frame is used as the STR frame and the $(i-N)$ -th frame (where $N>1$) is used as the LTR frame for N frames. The LTR frame then jumps forward by N frames and remains the same for encoding the next N frames. With the concept of dual reference frames, the LTR frame is the most relevant and therefore the competitor of the McFIS of the proposed scheme, and we have compared the proposed scheme with the LTR-based scheme using the pattern mode. We have also compared the proposed scheme with the algorithm in [14] where the McFIS is used as the second reference frame but no pattern mode is used. The algorithm in [14] also differs from the proposed scheme in the McFIS generation where spatial neighbouring pixels were used to modify the McFIS (i.e., unlike in Equation (4) where the previous McFIS is used).

Overall experiments are performed using a number of standard video sequences with QCIF and CIF resolutions. All sequences are encoded at 25 frames per second and 16 frames as the group of picture. Full-search quarter-pel ME with ± 15 as the search length is used. We have used IPPPP... format. In our implementation we use high quality LTR (HQLTR) [9][10] and high quality intra (I)-frame for better performance. To ensure this, we set the quantization parameters (QPs) for the HQLTR and the I-frame as $QP(I)=QP(HQLTR)=QP(P)-4$, where $QP(.)$ represents corresponding QP. Figure 3 (b) shows the average percentages (using the six standard videos listed in the Sub-section 2.1) of references using the McFIS and the LTR frames respectively (where the corresponding rest portions are referenced using the immediate previous frame). The results indicate that the McFIS captures more background areas compared to the conventional LTR frame. This translates into the improving rate-distortion performance by the proposed scheme, compared to that of the LTR frame. Fig 4 shows rate-distortion performance using the proposed (McFIS-PVC), the conventional HQLTR frame (LTR-PVC) [9], and only the McFIS [14] (McFIS) algorithms for the first 300 frames of the six aforementioned standard video sequences. The figure confirms that the proposed method consistently outperforms the relevant two algorithms by 0.20~1.25 dB.

4. CONCLUSIONS

In this paper, we proposed a new pattern-based video coding technique using dynamic background frame (i.e., the most common frame in a scene (McFIS)) as the long term reference frame to overcome the inaccurate motion estimation and compensation problem in the uncovered background areas by a new pattern matching and referencing technique. The experimental results showed that the proposed technique outperforms the two most relevant existing algorithms by improving 0.25~1.25 dB in coded image quality without noticeable computational increase.

5. REFERENCES

- [1] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, "Overview of the H.264/AVC Video Coding Standard," *IEEE TCSVT*, 13(7), 560-576, 2003.
- [2] O. Divorra Escoda, P. Yin, C. Dai, X. Li, "Geometry-adaptive block partitioning for video coding," *IEEE Int. Conference on Acoustics, Speech, and Signal Processing, (ICASSP-07)*, pp. I-657-660, 2007.
- [3] J. H. Kim, A. Ortega, P. Yin, P. Pandit, and C. Gomila, "Motion compensation based on implicit block segmentation," *IEEE Int. Conference on Image Processing*, pp. 2452-2455, (ICIP-08), 2008.
- [4] S. Chen, Q. Sun, X. Wu, and L. Yu, "L-shaped segmentations in motion-compensated prediction of H.264," *IEEE International Conference on Circuits and Systems, (ISCAS-08)*, 2008.
- [5] K. -W. Wong, K. -M. Lam, and W. -C. Siu, "An efficient low bit-rate video-coding algorithm focusing on moving regions," *IEEE TCSVT*, vol. 11(10), pp. 1128-1134, 2001.
- [6] M. Paul, M. Murshed, and L. Dooley, "A real-time pattern selection algorithm for very low bit-rate video coding using relevance and similarity metrics," *IEEE TCSVT*, vol. 15(6), pp. 753-761, 2005.
- [7] M. Paul and M. Murshed, "Threshold-free pattern-based low bit rate video coding," *IEEE ICIP-08*, pp. 1584-1587, 2008.
- [8] J. -R. Ding and J. -F. Yang, "Adaptive group-of-pictures and scene change detection methods based on existing H.264 advanced video coding information," *IET Image Processing*, 2(2), 85-94, 2008.
- [9] V. Chellappa, P. C. Cosman, and G. M. Voelker, "Dual Frame Motion Compensation with Uneven Quality Assignment," *IEEE TCSVT*, vol. 18, no. 2, pp. 249 - 256, 2008.
- [10] M. Tiwari and P. C. Cosman, "Selection of Long-Term Reference Frames in Dual-Frame Video Coding Using Simulated Annealing," *IEEE Signal Processing Letter*, vol. 15, pp. 249-252, 2008.
- [11] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," *IEEE Conference on CVPR*, vol. 2, pp. 246-252, 1999.
- [12] D.-S. Lee, "Effective Gaussian mixture learning for video background subtraction," *IEEE TPAMI*, 27(5), 827-832, May 2005.
- [13] M. Haque, M. Murshed, and M. Paul, "Improved Gaussian mixtures for robust object detection by adaptive multi-background generation," *IEEE Conference on PR*, 1-4, 2008.
- [14] M. Paul, W. Lin, C. T. Lau, and B. -S. Lee, "Video coding using the most common frame in scene," *IEEE ICASSP*, 2010.

Fig 4: Rate-distortion performance by the proposed (McFIS-PVC), the conventional high quality long term reference frame (LTR-PVC) [9] embedding the PVC, and dual frame reference frames using the most common frame in a scene (McFIS) [14] for six standard video sequences.