MCFIS: BETTER I-FRAME FOR VIDEO CODING

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 conventional Intra (I-) frame is used for error propagation prevention, backward/forward play, random access, indexing, etc. This frame is also used as an anchor frame for referencing the subsequence frames. To get better rate-distortion performance a frame should have the following quality to be an ideal I-frame: the best similarity with the frames in a GOP, so that (i) when it is used as a reference frame for a frame in the GOP we need less bits to achieve the desired image quality; (ii) if any frame is missing at the decoding end we can retrieve the missing frame from it. In this paper we will generate a most common frame of a scene (McFIS) in a video sequence using dynamic background modelling and then encode it to replace the conventional I-frame. By using McFIS as an I-frame, we not only gain the above mentioned two benefits but also ensure adaptive GOP for better rate-distortion performance compared to the existing coding schemes. The experimental results confirm the superiority of our proposed scheme in comparison with the existing state-of-art methods by significant image quality and computation time.

Index Terms—Video coding, uncovered background, light change, repetitive motion, H.264, motion estimation, multiple sprites, sprite, MRF, and multiple reference frames.

1. INTRODUCTION

The latest video coding standard H.264 as well as other modern standards uses Intra (I-) and Inter frames for improved video coding [1]. I-frame is encoded using only its own information and thus can be used for error propagation prevention, backward/forward play, random access, indexing, etc. On the other hand, inter frame is encoded with the help of previously encoded I- or inter-frame(s) for efficient coding. In the H.264 standard, frames are encoded as a group of picture (GOP) comprises one I-frame with subsequent inter frames. The number of I-frame is fewer compared to the inter-frames because an I-frame requires two to three times more bits compared to the inter frames for the same image quality. I-frame is also used as an anchor frame for referencing the subsequence inter-frames of a GOP directly or indirectly. Thus, the encoding errors (due to the quantization) of I-frame are propagated and exaggerated towards the end of the frames of a GOP. By selecting first frame as an I-frame without verifying its quality to be an I-frame, we may avoid decoding delay but sacrifice overall rate-distortion performance.

A frame being the first frame of a GOP is not automatically qualified as a good I-frame. A frame will be an ideal I-frame if it has following quality: the best similarity with the frames in a GOP, so that (i) when it is used as a reference frame for inter-frames in the GOP we need fewer bits to achieve the desired image quality; (ii) if any frame is missing in the decoder we can retrieve the

missing frame from it. Moreover, if a video sequence does not contain any scene change or extremely high motion activity compared to the previous frames, insertion of I-frames reduces the coding performance. Therefore, we need to insert an optimal number of I-frames based on the *adaptive GOP* (AGOP) determination and scene change detection (SCD) algorithms.

The H.264 also introduces motion estimation (ME) and motion compensation (MC) using multiple reference frames (MRFs) [1][2]. It has been demonstrated that MRFs facilitate better predictions than using just one reference frame, for video with repetitive motion, uncovered background, non-integer pixel displacement, lighting change, etc. The requirement of index codes (to identify the particular reference frame used), computational time in ME & MC (which increases linearly with the number of reference frames), and memory buffer size (to store decoded frames in both encoder and decoder) limits the number of reference frames used in practical applications. The optimal number of MRFs depends on the content of the video sequences. Typically the number of reference frames varies from one to five (16 is the maximum recommended reference frames in the H.264). If the cycle of repetitive motion, exposing uncovered background, non-integer pixel displacement, or lighting change exceeds the number of reference frames used in MRFs coding system, there will be not any improvement and therefore, the related computation (mainly that of ME) and bits for index codes are wasted. Moreover, the existing MRFs based system experiences disaster in decoded picture quality if any frame is lost during transmission.

To deal with the major problem of MRFs, a number of techniques [3]-[6] have been developed for reducing the computation associated with. Most of the fast MRF selection algorithms including the above mentioned techniques used one reference frame (in the best case) when their assumptions on the correlation of the MRF selection procedure are satisfied or five reference frames (in the worse case) when their assumptions are completely failed. But it is obvious that in terms of rate-distortion performance, their techniques could not outperform the H.264 with five reference frames which is considered as optimal [1]. Moreover, their techniques also suffer disaster in image quality if any frame is missing during transmission. Uncovered background can also be efficiently encoded using sprite/multiple-sprite coding through computationally expensive object segmentation. Most of the video coding applications could not tolerate inaccurate video/object segmentations and expensive computational complexity incurred by segmentation algorithms.

Recently *dynamic background modeling* (DBM) [7]-[9] using *Gaussian mixture model* (GMM) has been introduced for robust and real time object detection from so called *dynamic environment* where *ground-truth* background is unavailable. Moreover, a static background model does not remain valid due to illumination variation over time, intentional or unintentional camera

displacement, shadow/reflection of foreground objects, and intrinsic background motions (e.g. waving tree leaves etc.) [9]. Object can be detected more accurately by subtracting the background frame (generated from the background model) from the current frame. We believe that a joint AGOP and SCD technique can be developed using only one *appropriate* I-frame containing most *common* and *enough* scene information and for computationally efficient and better rate-distortion performance.

In this paper, we have incorporated DBM into the video coding to generate *appropriate* I-frames to improve the SCD for AGOP, coding performance, and error concealment. First we generate a most common frame of a scene (McFIS) from the first few original frames of a video sequence using DBM. The McFIS is encoded as an I-frame and then used it as a second reference frame for all frames in a scene (the first reference frame is the already decoded immediate previous frame) unless SCD occurs. If SCD occurs, again we generate McFIS using the first few frames from the new scene and encode it as an I-frame. All of the frames of the scene are encoded as inter frames unless SCD occurs again. A simple SCD algorithm is also derived based on the McFIS as it contains *common* and *stable* information of a scene.

The rest of the paper is organized as follows: Section 2 describes the proposed coding scheme. Section 3 analyses the computational requirement. Section 4 demonstrates the experimental set up and results. Section 5 concludes the paper.

2. PROPOSED CODING SCHEME

In the proposed coding scheme a McFIS is generated using several original frames of a scene in a video sequence with the DBM [9]. Obviously our proposed DBM (pDBM) will differ from the traditional DBM (tDBM) as the former focuses on the efficient video coding and the later focuses on the background modeling for object detection. We encode the McFIS as an I-frame and all the frames (including first frame) of a scene are encoded as inter frames using the immediate previous frame (except the first frame) and the McFIS as two reference frames until a SCD occurs. When SCD occurs a new McFIS will be generated using a few frames of the new scene and encode it as an I-frame. All the frames of the scene are encoded as inter frames. Note that the McFIS is only used for the referencing but not for displaying at the decoding end. A computationally efficient SCD algorithm is also derived based on the McFIS (described in Section 2.2). The subsequent subsections will describe McFIS generation, SCD, and the architecture of the proposed scheme.

2.1 Generation of most common frame in scene (McFIS)

To get the best performance in object detection, the tDBM is performed at pixel level, i.e., each pixel of a scene is modeled independently by a mixture of K (normally three models are used) Gaussian distributions [7]-[9]. Each Gaussian model represents the intensity distribution of one of the different environment components e.g., moving objects, static background, shadow, illumination/cloud changes, etc. observed by the pixel in frames. If we assume that k -th Gaussian representing a pixel intensity is η_k with mean μ_k , variance σ_k^2 , and weight w_k such that $\sum w_k = 1$ for all k . The Gaussians are always ordered based on the w/σ in descending order assuming that the top Gaussian will provide most stable background [9]. The system starts with an empty set of models and then for every new observation X_t at the current time t , it is first matched against the existing models in order to find one (say the k th) such that $|X_t - \mu_k| \leq 2.5\sigma_k$. If such a model exists, its

associated parameters are updated. Otherwise, a new Gaussian is introduced with $\mu = X_t$, arbitrarily high σ , and arbitrarily low w by evicting η_k if it exists.

From the above mentioned models, background and foreground are determined using different techniques. Among them, Haque *et al.* [9] proposed the most effective approach. They used a parameter called *recentVal* to store recent pixel intensity value when a pixel satisfies a model in the Gaussian mixture. Then *recentVal* is used as background intensity if the corresponding model is selected as background model.

We have observed that *mean* and *recentVal* intensities are two extreme cases to generate true background intensity for better video coding. The *mean* is too generalized for pixel intensities over the time and the *recentVal* is too biased to the recent pixel intensity. Thus, in the pDBM method we used a weighting factor between the *mean* and *recentVal* to reduce the delay response (due to *mean*) and to speed up the learning rates (due to *recentVal*), which are sometimes desirable criteria for real-time object detection using dynamic background.

2.2 Scene change detection (SCD) and adaptive group of picture (AGOP)

Recently Ding *et al.* [2] combined AGOP and SCD for better coding efficiency based on the motion vectors and the sum of absolute transformed differences (SATD) using 4×4 pixels block. This method ensured 98% accuracy of SCD with 0.63dB image quality improvement. Most of existing methods used some metrics computed using already processed frames and the current frames. The McFIS is the most similar frame comprising stable portion of the scene (mainly background) compared to the individual frame in a scene. Thus the SCD is determined by a *simple* metric computed using the McFIS and the current frame. For SCD using McFIS, we randomly select 50% of the pixel position of a frame and find the sum of absolute difference (SAD) between McFIS and the current frame. If the SAD for the current frame is greater than that of the previous frame by 70%, we consider SCD occurs and regenerate McFIS using few original frames of the next scene and encode it as an I-frame; otherwise we continue inter-coding. To be more specific, our GOP size is the size of a scene. Unlike the other existing algorithm, we don't need any explicit algorithm for AGOP, and it can be achieved as an integrated part of the McFIS. Moreover, a scene change means the change of background of a video sequence. As the McFIS has the *history* of the scene we don't need a *rigorous* process (like Ding's algorithm) for SCD.

2.3 The proposed coding system

Fig 1 shows a comparison between the conventional and the proposed frame types and referencing techniques using two reference frames for the first four frames of a scene in a video sequence. The H.264 encoder and decoder are employed in the proposed system with exception that McFIS is encoded as an I-frame and used as the second reference frame. Thus the proposed system has two reference frames i.e., immediate previous frame and McFIS. Based on the rate-distortion Lagrangian optimization, final reference frame is selected for each block.

As the proposed McFIS would be a good choice of a reference frame especially for smooth areas, true background, and uncovered background areas compared to the other 4 reference frames, we have introduced a new skip macroblock (MB) definition as follows, with the standard definition.

Fig 1: Comparison between the (a) conventional and (b) proposed frame types and referencing techniques using two reference frames for the first four frames.

Let $C_k(x, y)$ and $R_k(x, y)$ denote the k^{th} MB of the current frame and corresponding McFIS of a video sequence, respectively with frame size W pixels $\times H$ lines where $0 \leq x, y \leq 15$ and $0 \leq k < W/16 \times H/16$. The moving region $M_k(x, y)$ of the k^{th} MB in the current frame is obtained as:

$$M_k(x, y) = \begin{cases} 1, & \text{if } |(C_k(x, y) \bullet B) - (R_k(x, y) \bullet B)| > 2; \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where B is a 3×3 unit matrix for the morphological closing operation $\bullet$, which is applied to reduce noise. If $\sum M_k(x, y) < QP/2$ where QP is the quantization parameter, a MB is skipped. By this new definition of skip MBs, the proposed coding technique classifies more MBs as skip MBs, and this is the one of the reasons for the lower bit requirement compared to the other existing methods. Moreover, this does not compromise image quality as the McFIS is a good reference frame. Note that under this definition if any MB is classified as a skip MB, we don't process any other modes so the encoding is speeded up.

As the McFIS plays an important role in the proposed scheme, we encode it with relative finer quantization compared to the inter frames. We have derived QP for I-frame (QP_{Intra}) from as follows

$$QP_{\text{Intra}} = \begin{cases} 40 & \text{if } QP_{\text{Inter}} > 40 \\ \lfloor e^{0.09 \times QP_{\text{Inter}}} \rfloor & \text{if } 20 \leq QP_{\text{Inter}} \leq 40 \\ 4 & \text{if } QP_{\text{Inter}} < 20 \end{cases} \quad (2)$$

from the QP of inter-frame (QP_{Inter}). The rationality of this formulation is that we try to get better quality of the McFIS (i.e., I-frame) compared to the inter-frame by finer quantization.

3. COMPUTATIONAL COMPLEXITY

Let ζ , d , δ , τ , and λ , be the total number of MBs in a frame, the total number of motion search points, the number of operations in each point, fraction of frames of a scene used for McFIS generation, and the average number of modes per MB per reference frame for ME respectively. For ME the H.264 requires $\zeta d \delta \lambda N^2$ operations while an MB comprises $N \times N$ pixels (for simplification we don't distinguish different operations). After ME we need operations for bits stream generation and so on. But these depend on the combination of DCT coefficients and variable length code tables. The ME, irrespective of a scene's complexity, typically comprises more than 60% of the processing overhead required to inter-encode a frame with a software codec using the DCT, when full search is used. From this, the H.264 requires $8.35 \zeta d \delta \lambda N^2$ operations for encoding of a frame using five reference frames.

Obviously the proposed technique needs some additional operations to generate McFIS. The whole process (consisting of five sub-processes: model upgrading/creation, model deleting,

normalizing, filtering, and background generating) takes 66, 6, 6, 8, and 4 operations per pixel respectively. We also need 6 operations for SCD in SAD calculations (with randomly selected 50% of pixels). Thus, in total we need $3.34 \zeta d \delta \lambda N^2 + 96 \zeta \tau N^2$ operations using the proposed approach with immediate previous frame and McFIS as two reference frames. Ding's algorithm needs extra ME and SATD calculations. Thus Ding's algorithm takes $3.34 \zeta d \delta \lambda N^2 + 0.8 \zeta d \delta N^2 + 10 \zeta N^2$ operations using two reference frames. We have compared computation time of the proposed algorithm and Ding's algorithm against the H.264 with fixed GOP size and five reference frames. Fig 2(a) confirms that the proposed algorithm and Ding's algorithm reduce the computational time by 61% and 58% respectively comparing with the H.264 when we have used first 32 frames of a scene for McFIS generation.

Fig 2: Computational time comparison of the proposed and Ding's algorithms against the H.264 with five reference frame (a); and percentages of references comparison between Ding's and the proposed algorithms (b).

4. EXPERIMENTAL RESULTS

Overall experiments are performed using a number of standard video sequences with QCIF, CIF, and 4CIF resolutions. All sequences are encoded at 25 frames per second. Full-search fractional ME with ± 15 as the search length is used. For comparison, we have selected Ding's algorithms and the H.264 with fixed 32 GOP size using five reference frames. Ding's algorithm is the best existing method in rate-distortion, SCD, and AGOP performance. For the completeness of comparison, we have also selected the H.264 using fixed GOP and five reference frames (termed as H.264). For the Ding's algorithms we have used two references (i.e., the immediate and 2nd immediate previous frames as the references).

Fig 2(b) shows the average percentages of reference using McFIS and 2nd frame for the proposed algorithm and Ding's algorithm respectively. The figure demonstrates that 21% and 11% references are selected by the proposed algorithm and Ding's algorithm respectively. This large number of referencing indicates rate-distortion improvement using McFIS as a reference frame over Ding's method. Fig 3 shows referencing map by the proposed scheme. A large number of areas (normal regions in Fig 3(b)) are referenced by the McFIS, and this indicates the effectiveness of the McFIS for improving coding performance.

Fig 4 shows rate-distortion performance using the proposed, Ding's, and the H.264 MRFs algorithms for two mixed video sequences (they are for SCD & AGOP as well) and other individual four sequences. Each mixed video sequence has 11 different videos with at least 50 frames. *Akiyo*, *Miss America*, *Claire*, *car phone*, *Hall Monitor*, *News*, *Salesman*, *Grand ma*, *Mother*, *Suzie*, and *Forman* are used in the mixed QCIF. *Silent*,

Fig 3: Referenced regions by the proposed method; (a) decoded 95th frame of *Paris* sequence, (b) black and other regions are referenced from the immediate previous frame and the McFIS respectively.

0.5~2.0 dB. It is mainly due to the large number of cases in which the McFIS is used as a reference frame for the background areas (see Fig 2(b) and Fig 3(b)) and bit saving for inter frame (instead of I-frame). The proposed scheme also successfully detects scene change to insert I-frame for improving coding efficiency.

5. CONCLUSIONS

In this paper, we proposed a new video coding technique using dynamic background frame (termed as a McFIS) as the I-frame and then used as the second reference frame to improve coding efficiency. The proposed McFIS is generated using real-time Gaussian mixture model. The proposed method used the McFIS's inherent capability of scene change detection for adaptive GOP. Using McFIS as the I-frame and the second reference frame we have the following advantages compared to the existing methods based on MRFs, SCD, and AGOP: (1) Only one McFIS is used instead of a number of reference frames so the overheads of index codes are reduced; (2) A McFIS enables the possibility of capturing a whole cycle of repetitive motion, exposing uncovered background, non-integer pixel displacement, or lighting change; (3) A simple mechanism for AGOP and SCD determination is possible using McFIS as it is the most common frame in that scene. SCD and AGOP are effectively determined by comparing difference between McFIS and the current frame, and thus the SCD (therefore AGOP) is integrated with reference frame generation;

Waterfall, Coastguard, Paris, Hall Monitor, Bridge far, Highway, Football, Bridge close, and Tennis are used in the mixed CIF video. The figure confirms that the proposed method outperforms the state of the art method (Ding's [2]) and the H.264 with five reference frames by

(4) Less computation in ME&MC is required using McFIS compared to multiple reference frames.

The proposed video coding technique outperforms the relevant state of the art algorithms, inclusive of the H.264 standard using fixed GOP and five reference frames, in terms of rate-distortion and computational requirement. The experimental results show that the proposed technique outperforms the existing algorithm with SCD and AGOP by at least by 0.5 dB with comparable computational time. The proposed algorithm also outperforms the H.264 by 0.5~2.0 dB, and saves 60% of computational time.

6. 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] 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.
- [3] Y. -W. Huang, B. -Y. Hsieh, S. -Y. Chien, S. -Y. Ma, and L. -G. Chen, "Analysis and complexity reduction of multiple reference frames motion estimation in H.264/AVC," *IEEE TCSVT*, 16(4), 507-522, 2006.
- [4] L. Shen, Z. Liu, Z. Zhang, and G. Wang, "An Adaptive and Fast Multi frame Selection Algorithm for H.264 Video Coding," *IEEE Signal Processing Letters*, vol. 14, No. 11, pp. 836-839, 2007.
- [5] T. -Y. Kuo, H. -J. Lu, "Efficient Reference Frame Selector for H.264," *IEEE Transaction on Circuits and Systems for Video Technology*, vol. 18, no. 3, pp. 400-405, 2008.
- [6] K. Hachicha, D. Faura, O. Romain, and P. Garda, "Accelerating the multiple reference frames compensation in the H.264 video coder," *Journal of Real-Time Image Processing*, Springer Berlin / Heidelberg, Vol. 4, No. 1, pp. 55-65, 2009.
- [7] 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.
- [8] D.-S. Lee, "Effective Gaussian mixture learning for video background subtraction," *IEEE TPAMI*, 27(5), 827-832, May 2005.
- [9] 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.

Fig 4: Rate-distortion performance by the proposed, Ding's and the H.264 with 5 MRFs algorithms for mixed and other QCIF/CIF/4CIF video sequences.