VIDEO CODING USING THE MOST COMMON FRAME IN SCENE

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

Motion estimation (ME) and motion compensation (MC) using variable block size, fractional search, and multiple reference frames (MRFs) help the recent video coding standard H.264 to improve the coding performance significantly over the other contemporary coding standards. The concept of MRF achieves better coding performance in the cases of repetitive motion, uncovered background, non-integer pixel displacement, lighting change, etc. The requirement of index codes of the reference frames, computational time in ME&MC, and memory buffer for pre-coded frames limits the number of reference frames used in practical applications. In typical video sequence, the previous frame is used as a reference frame with 68~92% of cases. In this paper, we propose a new video coding method using a reference frame (i.e., the most common frame in scene (McFIS)) generated by the Gaussian mixture based dynamic background modelling. The McFIS is not only more effective in terms of rate-distortion and computational time performance compared to the MRFs but also error resilient transmission channel. The experimental results show that the proposed coding scheme outperforms the H.264 standard video coding with five reference frames by at least 0.5 dB and reduced 60% of 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 H.264 advanced video coding standard improves rate-distortion performance significantly compared to its predecessors and competitors by introducing a number of innovative ideas in Intra- and Inter-frame coding [1]. Major performance improvement is taken place by means of motion estimation (ME) and motion compensation (MC) using variable block sizes, sub-pixel search, and multiple reference frames (MRFs) [2]. Many conditions demonstrate that MRFs facilitate better predictions than a system 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 reference frame), 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. If the cycle of repetitive motion, exposing uncovered background, non-integer pixel displacement, or lighting change exceeds the number of reference frames used in the MRF coding system, we will not get any coding improvement and a lot of computation time is wasted. Moreover, the existing MRF system also experiences disaster in image quality if any frame is missing during transmission.

A number of techniques including [3]-[6] are found in literature to reduce computational time associated with MRFs without jeopardizing image quality. Huang *et al*. [3] searched either the previous or every reference frame based upon the result of the intra prediction and ME from the previous frame. This approach can reduces 76-96% of computational complexity by avoiding searching for unnecessary reference frames. Moreover, this approach is orthogonal to conventional fast block matching algorithms, and they can be easily combined to achieve further efficient implementations. Shen *et al*. [4] proposed an adaptive and fast MRF selection algorithm based on the hypothesis that homogeneous areas of video sequences probably belong to the same video object, move together as well, and thus have the same optimal reference frame. Simulation results show that this algorithm can deduct 56~74% of computation time in ME. Kuo *et al*. [5] proposed a fast MRF selection algorithm based on the initial search results using 8×8-pixel block. Hachicha *et al*. [6] used *Markov Random Fields* algorithm relying on robust moving pixel segmentation. This approach saved 35% of coding time by reducing the number of reference frames at three instead of five without jeopardizing image quality.

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 real time video coding applications could not tolerate inaccurate video/object segmentations and expensive computational complexity incurred by the above mentioned algorithms.

Recently dynamic background modelling (DBM) [7]-[9] using *Gaussian Mixture Model* (GMM) is introduced for robust object detection from so called dynamic environment where ground-truth background is impossible due to the practical reasons such as a busy train station, airport, etc. Moreover, static background model does not remain valid due to illumination variations 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 background frame (generated from the background model) from the current

frame. In this paper, we have incorporated DBM into the video coding architecture to improve the coding performance through ME&MC using McFIS as a second reference frame. First we generate McFIS from the pre-decoded frames using DBM, and then use it as a second reference frame (first reference frame is the immediate previous frame). The same McFIS generation technique is used at the encoder and decoder so that we don't need to send background model to the decoder. When scene change is detected for a video sequence we only reset the DBM model (hence McFIS), otherwise we continue updating the McFIS. Using McFIS as a reference frame we have the following advantages compared to the existing methods based on MRFs, adaptive GOP (AGOP), scene change detection (SCD), and background coding:

- Only one McFIS is used instead of a number of reference frames as it can capture a whole cycle of repetitive motion, exposing uncovered background, light changes, etc.
- Since a McFIS is generated from the history of already decoded frames, intrinsically it has better error recovery capacity as the McFIS has already contained pixel intensity history of the frames.
- Any simple mechanism for SCD&AGOP determination by comparing difference between McFIS and the current frame will be more effective.
- Less computation time in ME&MC is required using McFIS compared to the three or more reference frames.

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

2. PROPOSED CODING SYSTEM

The McFIS is generated using DBM based on the GMM [7]-[9]. Obviously traditional DBM (tDBM) would be different from our customized DBM (cDBM) as the tDBM primarily focuses on object detection, whereas the cDBM focuses on rate-distortion optimization when McFIS is used as an extra reference frame for encoding uncovered background, repetitive motion, etc. Moreover, the tDBM has used original video frames to construct background frame (BF), whereas the proposed cDBM will use decoded frames, which are quantized at different levels based on the available bit-rate, to generate McFIS. It is also used for SCD. If the SCD occurs then we have to reset McFIS, otherwise we will update current McFIS using the recent decoded frame. The subsequence subsections will describe McFIS generation, SCD, and the architecture of the proposed scheme.

2.1 McFIS generation

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 kth) such that $|X_t - \mu_k| \le 2.5\sigma_k$. 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 ω by evicting η_K if it exists. From the above mentioned models, background and foreground are determined using different techniques. Stauffer *et al.* [7] and Lee *et al.* [8] used two different kinds of user-defined thresholds based on the background and foreground ratio, whereas, Haque *et al.* [9] used *classical* background subtraction method which identifies an object if the value of the current intensity differs from the recent value of the *best* background model by a well-studied threshold to avoid delay response of the above two models.

As we have mentioned earlier, due to the quantization the original pixel intensity and decoded pixel intensity may vary. This variation may effect in McFIS generation, and ultimately in coding performance. To minimize this effect we have modified current pixel intensity in the decoded frame using the average value of the neighboring pixel intensities. Let D_t be the t-th decoded frame to generate (t+1)-th McFIS. For a given pixel position (x, y) in D_t, we modified $D_t(x, y)$ as $D'_t(x, y)$:

$$D'_t(x,y) = \begin{cases} \tau D_t(x,y) + (1-\tau)\overline{D}_t(x,y) & if \left| D_t(x,y) - \overline{D}_t(x,y) \right| < T_p \\ D_t(x,y) & otherwise \end{cases} \quad (1)$$

where τ and T_p are the weighting factor and threshold respectively and $\overline{D}_t(x,y) = 1/4 \sum_{i=0}^{1}\sum_{j=0}^{1} D_t(x+i, y+j)$. (2)

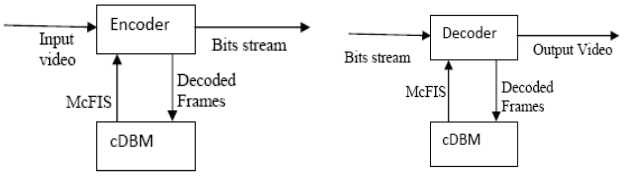
Unlike the tDBM, we have used D'_t instead of D_t to update the model. In our experiment we have used $\tau = 0.5$ and $T_p = 3$. We have observed that generation of background image i.e., McFIS using recent value sometimes does not work properly. It is due to the pixel intensity fluctuation causes by the course quantization. To minimize this variation we have used same (i.e., τ) weighting factor between mean and recent value.

2.2 Scene change detection (SCD) and 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 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 1.7, then we consider SCD occurs and insert an I-frame, otherwise we continue inter-coding (no other AGOP). This would be effective compared to the existing algorithms as the McFIS is *equivalent* to a group of already processed frames. 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 Architecture of the proposed coding system

Fig 1 shows the architecture of the proposed coding system. The H.264 encoder and decoder are employed in the proposed system



(a) (b)

Fig 1: Block diagram of the proposed encoder (a) and decoder (b) where both use McFIS, which is generated from the decoded frames.

with only exception that McFIS is used as the second reference frame instead of using immediately previous five reference frames. 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, and 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.

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 $0 \le x,y \le 15$ and $0 \le k < W/16 \times H/16$. The moving region $M_k(x,y)$ of the k^{th} MB in the current frame is obtained as:

$$M_k(x,y) = \begin{cases} 1, & \text{if } |(C_k(x,y) \bullet B) - (R_k(x,y) \bullet B)| > 2; \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

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

3. COMPUTATIONAL COMPLEXITY

Let ζ, d, δ, and λ be the MBs in a frame, total motion search points, total number of operations in each point, 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 for N×N-pixel MB as each search points requires δ operations (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 researchers already claimed that ME, irrespective of a scene's complexity, typically comprises more than 60% of the processing overhead required to encode an inter picture with a software codec using the DCT, when full search is used. From this fact it can be well argued that the H.264 requires $8.35\zeta d \delta \lambda N^2$ operations for encoding of a frame using five reference frames.

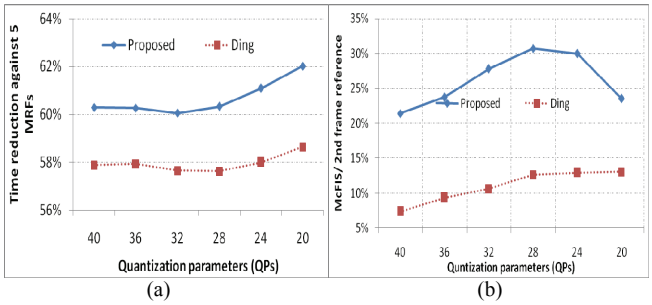
Obviously the proposed technique will take some extra operations to generate McFIS and interpolate McFIS for encoding each frame using fractional ME. To analyze this, we divided the whole process into five sub-processes such as model upgrading/creation, model deleting, normalizing, filtering, and background generating. Those take 66, 6, 6, 8, and 4 operations per pixel respectively. We also need 15 operations per pixel for fractional ME and 6 operations for SCD in SAD calculations and randomly selected 50% of pixels. Thus, in total we need $3.34\zeta d \delta \lambda N^2 + 108\zeta N^2$ operations using the proposed approach with immediate previous frame and McFIS as two reference frames Ding's algorithm needs extra ME and SATD calculations. Although this extra ME sometimes (on average 20% of cases) can be used if inter frame coding is decided through SCD and AGOP, it is considered an over head at the 80% of cases. 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

By ignoring smaller term of operations, we can theoretically predict that around 60% of computations saving can be possible using any of the proposed and Ding's algorithms against five MRFs. For more specific, we can also predict using the smaller parts of the operations that the proposed algorithm is slightly better than the Ding's algorithm. Fig 2(a) demonstrates that the proposed and Ding's algorithms reduce 61% and 58% on average respectively. This result confirms our theoretical prediction.

4. EXPERIMENTAL RESULTS

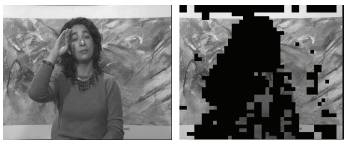
Overall experimental results are performed using a number of CIF&QCIF standard video sequences. All sequences are encoded at 25 frames per second. Full-search fractional ME with ±15 as the search length is used. For comparison, we have selected Ding's algorithms and the H.264 with fixed 32 GOP size using five reference frames. We have selected Ding's algorithm as we have found that Ding's algorithm is the best in terms of rate-distortion performance, SCD, and AGOP through the literature survey. For the completeness we have also selected the H.264 using fixed GOP and five reference frames. For the Ding's algorithms we have used two references (i.e., the immediate and 2nd immediate previous frames as the references). For the proposed algorithm the immediate previous frame and the McFIS are used as references.

Fig 2(b) shows the average percentages of reference using McFIS and 2nd frame for the proposed and Ding's methods respectively. The figure demonstrates that 26% and 11% references are selected by the proposed and Ding's algorithms respectively. This large number of referencing indicates rate-distortion improvement using McFIS as a reference frame over Ding's method. Fig 3 shows reference mapping by the proposed scheme. A large number of areas (normal regions in Fig 3(b)) are referenced by the McFIS, which indicates the effectiveness of the McFIS for improving coding performance.



(a) (b)

Fig 2: Computational time comparison between the proposed and Ding's algorithms against the H.264 5 MRFs technique (a); and percentages of references comparison between Ding's and the proposed algorithms (b).



(a) (b)

Fig 3: Referenced regions by the proposed method; (a) decoded 24th frame of Silent sequence, (b) black and other regions are referenced from the immediate previous frame and the McFIS respectively.

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*, *Waterfall*, *Coastguard*, *Paris*, *Hall Monitor*, *Bridge far*, *Highway*, *Football*, *Bridgeclose*, 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 MRFs by 0.5~2.0 dB. It is mainly due to the large number of cases the McFIS is used as a reference frame for the background areas (see Fig 2(b) and Fig 3(b)). The proposed scheme also successfully inserts I-frame based on the SCD to improve coding efficiency.

5. CONCLUSIONS

In this paper, we proposed a new video coding technique using dynamic background frame (termed as a McFIS) as the second reference frame to improve coding efficiency for uncovered background, repetitive motion, non-integer motion displacement, light change, etc. Unlike the sprite/ multiple-sprite coding, 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. The proposed video coding technique outperforms not only the state of the art algorithm but also 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 successfully detects scene change more accurately compared to the state of the art algorithm and outperforms it by at least by 0.5 dB with comparable computational time. The proposed algorithm also outperforms the H.264 with fixed GOP and 5 reference frames by 0.5~2.0 dB, and saves 60% of computation time.

6. REFERENCES

[1] T. Wiegand, G. J. Sullivan, G. Bjøntegaard, and A. Luthra, "Overview of the H.264/AVC Video Coding Standard," *IEEE Transaction on Circuits and Sys. for Video Techn.*, vol. 13, no. 7, pp. 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*, vol 2, no. 2, pp. 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 Transaction on Circuits and System for Video Technology*, vol. 16, no. 4, pp. 507-522, 2006,

[4] L. Shen, Z. Liu, Z. Zhang, and G. Wang, "An Adaptive and Fast Multiframe 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*, 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. Computer Vision and Pattern Recognition*, vol. 2, pp. 246–252, 1999.

[8] D.-S. Lee, "Effective Gaussian mixture learning for video background subtraction," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 27, no. 5, pp. 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 International Conference on Pattern Recognition pp. 1-4, 2008.

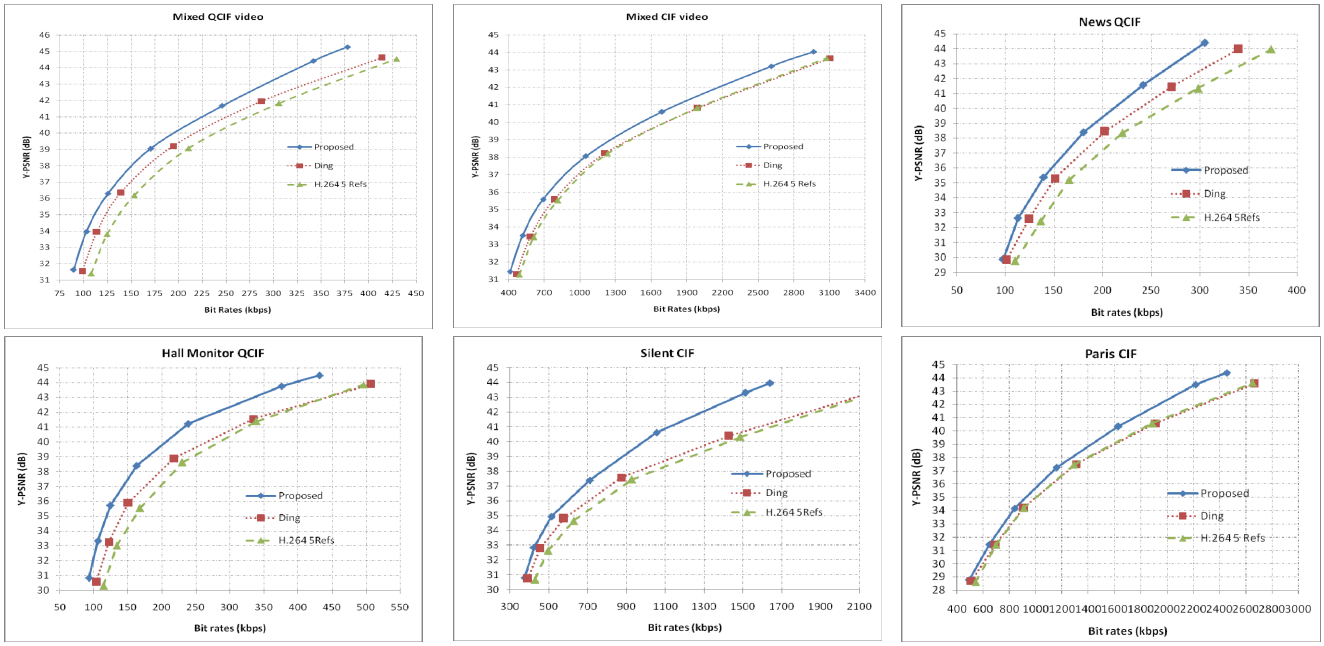


Fig 4: Rate-distortion performance by the proposed, Ding's and the H.264 with 5 MRFs algorithms for mixed and other 4 CIF& QCIF videos.