14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

# H.264 ENCODING OF VIDEOS WITH LARGE NUMBER OF SHOT TRANSITIONS USING LONG-TERM REFERENCE PICTURES

*Nukhet Ozbek[1] and A. Murat Tekalp[2]*

[1]International Computer Institute, Ege University
Bornova, 35100, Izmir, Turkey
phone: + 90-232-3423232 (116), fax: + 90-232-3887230, email: nukhet.ozbek@.ege.edu.tr
[2]College of Engineering, Koc University
Sariyer, 34450, Istanbul, Turkey

## ABSTRACT

*Long-term reference prediction is an important feature of the H.264/AVC standard, which provides a trade-off between gain and complexity. A simple long-term reference selection method is presented for videos with frequent shot/view transitions in order to optimize compression efficiency at the shot boundaries. Experimental results show up to 50% reduction in the number of bits, at the same PSNR, for frames at the border of transitions.*

## 1. INTRODUCTION

The H.264/AVC standard relies on several new features added into the classical block-based hybrid video coding scheme, such as intra prediction, variable block size motion compensation, multiple reference frames, content adaptive entropy coding, and in-loop de-blocking filter. There are two distinct reference picture lists, one is, ***list 0***, dedicated to inter prediction of a P, B or SP whereas the other one, ***list 1***, is only for a B slice. The Memory Management Control Operation (MMCO) commands are designed for adding, removing, and indexing short-term (ST) and long-term (LT) references into the Decoded Picture Buffer (DPB). Reference picture selection can be managed by using Reference Picture List Reordering (RPLR) commands. Thus, MMCO and RPLR commands enable reference list selection and management in an H.264/AVC video encoder. This paper focuses on the use of long term reference pictures for efficient encoding of video with a large number of shot/view transitions.

Early approaches for long-term memory motion compensated prediction have been presented in the MPEG-4 standardization group. In the "short-term frame memory/long-term frame memory" (STFM/LTFM) prediction approach, STFM stores the most recently decoded frame, while the LTFM stores a frame which has been decoded earlier and the encoder is enabled to use both to improve prediction efficiency [1]. Another approach was to include frames into the LTFM which are generated by background memory prediction techniques for the layered video coding [1].

In the H.264/AVC standard, LT references are designed to make long-term prediction from far frames and supported with multiple frame references. When a new reference is to be placed into the DPB, which can hold up to 16 pictures, sliding window memory management removes a ST reference. Only LT references can be kept in the DPB continuously and removed explicitly via MMCO commands.

Rate Distortion Optimization (RDO) is utilized in H.264/AVC for both motion estimation and mode decision processes. After motion estimation for all block types, the decision of reference selection is performed by jointly minimizing the distortion, which is Sum of Absolute Differences (SAD), and rate, which consists of motion vector and reference index bits [2]. It is a fact that the temporal correlation between two distinct frames reduces as the distance increases. Therefore, decision of a LT for prediction strictly depends upon the correlation between the two pictures.

In [3], LT references have been utilized to encode key frames, which are the best representative frames for each shot, in H.264/AVC video coding with multiple frame references. In this study, they have been employed to obtain additional gain in coding efficiency for typical movie or music videos composed of multiple shots/views with frequent transitions. If the first frame after a view transition is encoded as a P-frame, most macro-blocks are encoded in intra mode due to low correlation between pictures belonging to different scenes. We propose to employ LT references in order to reduce the cost of such P frames.

The paper is organized as follows: Section 2 describes the proposed LT selection and the brute-force methods. Section 3 provides comparative results and conclusions are drawn in Section 4.

## 2. LONG-TERM REFERENCE SELECTION

In the following, we propose a simple LT reference selection method for videos with large number of transitions between a fixed number of views. A brute-force approach is discussed in Section 2.2 as a benchmark for comparison.

### 2.1 Proposed LT Reference Selection

The proposed method for selecting LT references is developed for video which is a mixture of a fixed number of (N) continuous scenes captured by N cameras that are interleaved with frequent transitions between the scenes by means of cuts. Hence, the first frame after each cut is likely to be highly correlated to the last frame of one of the previous shots. Then, we propose to mark the last reference frame, P or reference B according to GOP structure, of each shot as LT and keep in the DPB. It is added into reference

list 0, when the successor shot of that scene starts again. The corresponding LT reference should exist in list 0 through the duration of that shot in order to observe where and how the LT is decided by the RDO module. To that end, list 0 size is set to two including one ST reference and one LT reference. At the end of a shot, the LT of the corresponding scene is renewed by the last reference of the shot.
In this study, two types of videos are examined; music clips and motion pictures. It is assumed that offline encoding is performed, and shot transitions are known *a priori*.

*2.2.1 Music Clip: Sandal Sequence*
"Sandal", the first test sequence, is a music clip, which consists of 14 shots and 453 frames. Sandal sequence is in CIF resolution and composed by images from four different scenes. Shot transitions of the video are depicted in Figure 1 where frame numbers indicate the last frame of a shot. Display orders, namely POC (Picture Order Count) values, of the frames in LT reference set for Sandal sequence are as follows: 12, 26, 94, 110, 158, 210, 254, 270, 324.

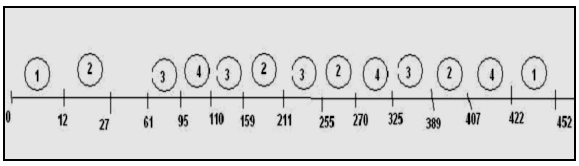


Figure 1 - Scene and shot settling in Sandal sequence.

*2.2.2 Motion Picture: Troy Sequence*
"Troy", the second test sequence, is a movie clip, which consists of 9 shots and 549 frames. Troy sequence is 640x272 and composed of switching images of two different cameras at the same scene. Shot transitions of the video are depicted in Figure 2. POC values of the frames in Troy's LT reference set are as follows: 98, 136, 226, 276, 432, 474.

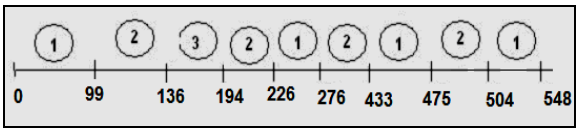


Figure 2 - Scene and shot settling in Troy sequence.

## 2.2 Brute-Force LT Reference Selection

In order to be able to evaluate the performance of the proposed method, there needs to be a benchmark method for comparison. We believe that brute-force selection method is the best benchmark for comparison.
In the brute-force approach, every shot is encoded by using every frame, P or B, within the previous (occurrence) shot of the same scene as a possible LT reference option, one by one at a time, and total bits of frames encoded within the current shot are saved for every trial. The trial that yields the minimum value of total bits determines the optimum LT reference picture for the current shot.
With the brute-force approach, the LT reference sets for test sequences turn out to be different than the ones used in our method; i.e., they are not necessarily the last frame of the previous occurrence of that shot. The following frames are selected as LT references using the brute-force method:

Sandal LT set : {6, 14, 64, 102, 132, 160, 254, 256, 312}
Troy LT set : {66, 102, 224, 274, 294, 440}

## 2.3 Selection of Other Encoder Parameters

In [4], it is reported with a wide range of test sequences that hierarchical B pictures can improve compression efficiency up to 1.5 dB compared to classical "IBBP..." structure. Note that, hierarchical B pictures, supported by the H.264/AVC syntax, correspond to MCTF (Motion Compensated Temporal Filtering) coding (without the update step).
Figure 3 shows a typical hierarchical prediction structure with GOP size 16 (5 levels). In this structure, every GOP has a key picture which is encoded first and all other pictures in the GOP are encoded after it. A key picture can be coded as intra or inter by using other key pictures. The non-key pictures (B pictures) are predicted using only the nearest past and future pictures of the lower temporal level. So, all B frames have to be stored B pictures except the ones in the last level.
LT references are utilized in [4] to code the key pictures when GOP size is more than 32. The only disadvantage of exclusive use of long-terms could be non-negligible overhead of MMCO and RPLR commands at low bit-rate videos. Simulation results show that the maximum coding efficiency was reached with GOP size 4 or 8 for high motion videos, whereas the GOP size is 16 or 32 for sequences with low or regular motion. Taking into consideration this result and in order to use for both input videos, we selected GOP size as 4. Another reason is that a smaller GOP size is better for the videos with frequent shot transitions so that the prediction loop interrupted by a new shot should be smaller.

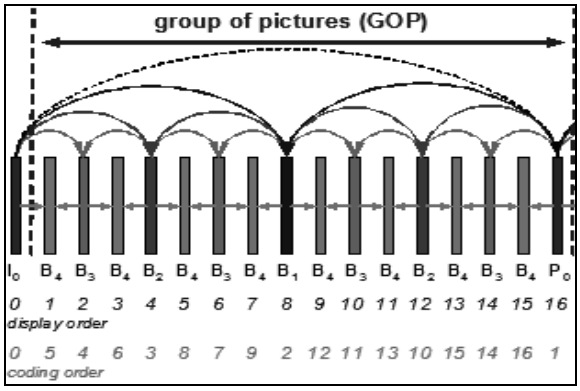


Figure 3 - Hierarchical coding structure GOP = 16 [4].

List 0 and list 1 have 2 and 1 elements, respectively. It is recommended to cascade QP (Quantization Parameter) values in [4]. Hence, QP values of pictures are determined according to hierarchy levels such as

$$QP(RB) = QP(P) + 1. \quad (1)$$

$$QP(B) = QP(RB) + 1. \quad (2)$$

where RB (Reference B) means stored-B picture. The QP values for P, RB, and B are 30, 31 and 32, respectively. The search range value is set as 32 for Sandal and 16 for Troy.

## 3. RESULTS

We have obtained results using the proposed method, the brute-force method and the JM reference software v9.2 [5] on two input videos with the same coding parameters. Tables 1, 2, 4 and 5 show the POC numbers of the pictures at shot transitions, which are key/P frames and firstly coded pictures of the shots, bits and PSNR values, and bit saving over JM v9.2. For how many block or sub-macroblock in that P picture, the corresponding LT is selected as reference is given in 4x4 block units in those tables, as well. The bitrates and bitrate savings achieved for the whole videos are presented in Table 3 and 6.

In order to observe how many times the LT references are selected in each frame, the reference histograms are constructed. They are depicted in Figure 4, 5, 6, and 7. When the figures are examined, it is observed that LT references are chosen at the first P picture of each shot at around 50% of the MBs. The rest of the MBs, where LT reference is not selected, are most likely coded as intra. For the second and third encoded frames of a shot, LT reference decision ratio is very low since they prefer the first P frame as reference, which is already predicted from the LT.

Tables 1 and 2 show that the bit savings for the first P frame after transitions vary between 0% to 50% for the Sandal sequence. For instance, in frames 112 and 272, the bit savings are over 40% at the same PSNR. However, in frames 160, 212 and 408 LT reference usage did not cause any bit saving. We also notice that for the frames in which bit saving is higher, the number of blocks selected LT reference is also higher. When Table 1 and 2 compared, it is shown that generally the same pictures are coded with lower bits and higher bit saving is achieved in Table 2 than that of Table 1. This results in two important observations:

- The optimum LT reference found by brute-force approach has outperformed the LT as last reference of the shot and has been chosen more by the RDO reference selection module.
- Providing LT references should be selected/used more for motion compensated prediction, the bit saving in shot leader P frame has been increased.

While searching the minimum number of bits to find optimum LT, the total bits for all frames within the shot are taken into account instead of the shot leader P frame bits only. Therefore, the optimum LT case has not to guarantee optimum bit saving for the shot leader P frame, as it can be seen from Table 1 and 2 in the instance picture number "112".

Tables 4 and 5 show that similar results are also obtained for the Troy sequence. For example, in frame 436 bit saving reaches 47%. For both Sandal and Troy sequences, although the bit savings in the first P frames after shot transitions are around 30% - 40%, the bitrate savings over the whole video is low. As seen in Tables 3 and 6, the bitrate saving over the whole sequence does not even reach 3% for both videos.

There are two reasons for this fact:

- The bit savings obtained at shot transitions are averaged out over the whole video, by dividing the bit saving into the number of frames within the shot, which varies between 15 and 157.
- The first shot, in which the first LT reference of that scene is to be found, brings no gain.

Consequently, the bit saving at a shot transition depends on how similar the selected long-term reference is to the first P picture of the shot. The overall bitrate saving also depends on the frequency of shot transitions.

| picture no | # of bits | psnr (dB) | bit saving % | # of 4x4 blocks |
|---|---|---|---|---|
| 112 | 30920 | 35.754 | 48.5 | 3612 |
| 160 | 100680 | 33.342 | 1.0 | 352 |
| 212 | 56144 | 35.863 | 0.0 | 332 |
| 256 | 89952 | 33.951 | 3.5 | 852 |
| 272 | 53296 | 33.286 | 43.0 | 3552 |
| 328 | 41488 | 35.685 | 31.0 | 3092 |
| 392 | 83736 | 33.905 | 3.2 | 652 |
| 408 | 66768 | 34.403 | 0.0 | 160 |
| 424 | 81712 | 33.539 | 6.3 | 1252 |

Table 1: Transition frames w/ proposed LTs for Sandal seq.

| picture no | # of bits | psnr (dB) | bit saving % | # of 4x4 blocks |
|---|---|---|---|---|
| 112 | 31704 | 35.419 | 47.2 | 3900 |
| 160 | 101208 | 33.349 | 0.4 | 272 |
| 212 | 55624 | 35.816 | 1.0 | 336 |
| 256 | 80360 | 33.760 | 13.8 | 1944 |
| 272 | 49832 | 33.208 | 46.7 | 3528 |
| 328 | 39904 | 35.654 | 33.6 | 3248 |
| 392 | 76840 | 33.813 | 11.2 | 1420 |
| 408 | 66608 | 34.376 | 0.2 | 140 |
| 424 | 70176 | 33.533 | 19.5 | 2464 |

Table 2: Transition frames w/ optimum LTs for Sandal seq.

| software | bitrate (kbps) | psnr (dB) | saving % |
|---|---|---|---|
| JM v9.2 | 411.61 | 33.74 | - |
| Proposed LTs | 404.19 | 33.74 | 1.8 |
| Optimum LTs | 401.59 | 33.72 | 2.4 |

Table 3: Performance results for Sandal sequence

| picture no | # of bits | psnr (dB) | bit saving % | # of 4x4 blocks |
|---|---|---|---|---|
| 196 | 61944 | 37.564 | 4.0 | 952 |
| 228 | 55512 | 37.194 | 25.5 | 4684 |
| 280 | 46896 | 37.748 | 20.4 | 3088 |
| 436 | 42624 | 36.845 | 46.2 | 7968 |
| 476 | 50408 | 38.534 | 4.4 | 1252 |
| 508 | 59200 | 36.610 | 27.8 | 5584 |

Table 4: Transition frames w/ proposed LTs for Troy seq.

| Picture no | # of bits | psnr (dB) | bit saving % | # of 4x4 blocks |
|---|---|---|---|---|
| 196 | 60144 | 37.524 | 6.8 | 1536 |
| 228 | 52168 | 37.115 | 30.0 | 6012 |

| | | | | |
|---|---|---|---|---|
| 280 | 46416 | 37.743 | 21.2 | 3312 |
| 436 | 42088 | 36.840 | 46.9 | 7880 |
| 476 | 46504 | 38.493 | 11.8 | 2240 |
| 508 | 56480 | 36.590 | 31.1 | 6104 |

Table 5: Transition frames w/ optimum LTs for Troy seq.

| software | bitrate (kbps) | psnr (dB) | saving % |
|---|---|---|---|
| JM v9.2 | 218.53 | 36.96 | - |
| Proposed LTs | 214.55 | 36.95 | 1.8 |
| Optimum LTs | 213.54 | 36.96 | 2.3 |

Table 6: Performance results for Troy sequence

## 4. CONCLUSION

In this study, H.264/AVC long-term reference selection for videos with frequent camera/shot transitions is investigated in order to increase compression efficiency. The proposed method for LT selection and optimal LT selection methods are compared with the original H.264/AVC reference encoder, bit savings and LT reference decisions have been analysed at both transition and total video basis. For a multi-scene multi-shot music video and a single-scene multi-shot movie video, test results are presented in this paper.

The proposed LT reference selection technique gives very close results when compared to the brute-force method, without any increase in the computational complexity.

The results show that for pictures around a camera/shot transition, the possibility of the LT reference selection is much more than the possibility of the ST reference for prediction. The possibility of selecting LT as reference strictly depends on the correlation between the LT and the picture encoded, and affects the reduction in bit costs of transition frames. Up to 50% reduction in the number of bits is achieved at the same PSNR for frames at the border of camera transitions whereas the overall bitrate reduction can not exceed 2.4%. In order to get higher gains similar videos with more frequent transitions should be used.

## REFERENCES

[1] T. Wiegand, X. Zhang, and B. Girod, "Long-Term Memory Motion-Compensated Prediction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, pp. 70-84, 1999.

[2] A.M. Tourapis, F. Wu, and S. Li, "Direct Mode Coding for Bipredictive Slices in the H.264 Standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, pp. 119–126, 2005.

[3] N. Ozbek and A.M. Tekalp, "Fast H.264/AVC Video Encoding with Multiple Frame References", in *Proc. ICIP 2005*, Genova, Italy, Sept. 11-14, vol. 1, pp. 597-600, 2005.

[4] H. Schwarz, D. Marpe, and T. Wiegand, "Hierarchical B Pictures," *Document JVT-P014*, 2005.

[5] Joint Video Team Reference Software, Version 9.2 (JM9.2), http://iphome.hhi.de/suehring/tml/download/

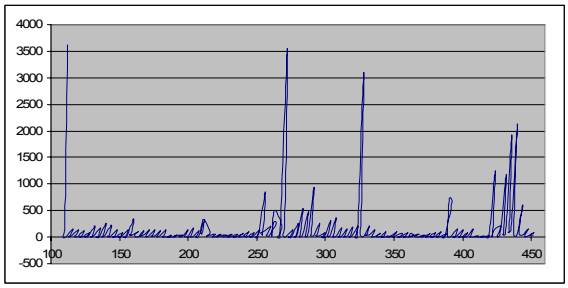


Figure 4: Reference histogram w/ proposed LTs for Sandal seq.

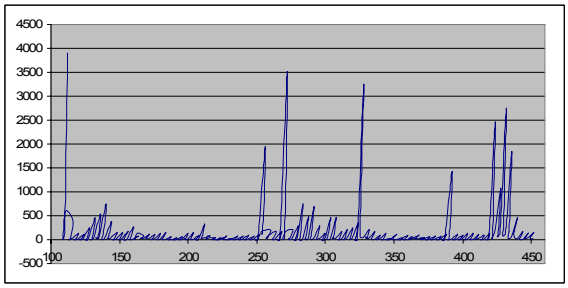


Figure 5: Reference histogram w/ optimum LTs for Sandal seq.

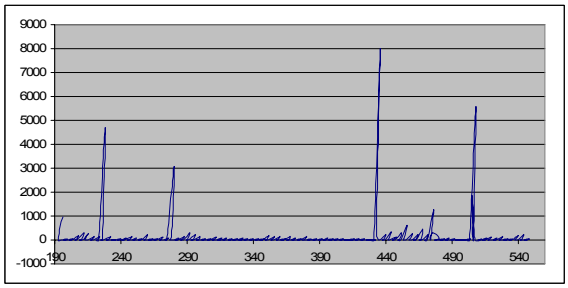


Figure 6: Reference histogram w/ proposed LTs for Troy seq.

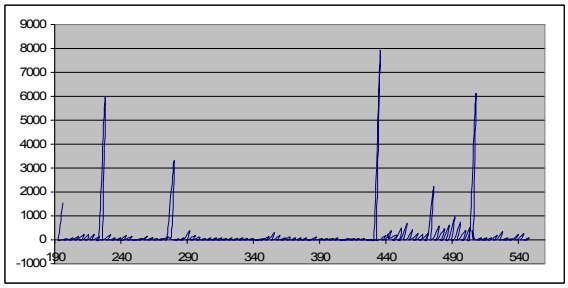


Figure 7: Reference histogram w/ optimum LTs for Troy seq.