# SCENE CONTEXT DEPENDENT REFERENCE FRAME PLACEMENT FOR MPEG VIDEO CODING

*Austin Y. Lan,* *Jenq-Neng Hwang*

*Information Processing Laboratory*
*Department of Electrical Engr., Box # 352500*
*University of Washington, Seattle, WA 98195*
*(206) 685-1603, hwang@ee.washington.edu*

## ABSTRACT

The MPEG video compression standard effectively exploits spatial, temporal, and coding redundancies in the algorithm. In its generic form, however, only a minimal amount of scene adaptation is performed. Video can be further compressed by taking advantage of scenes where the temporal statistics allow larger inter-reference frame distances. This paper proposes the use of motion analysis (MA) to adapt to scene content. The actual picture type (I, P, or B) decision is made by examining the accumulation of motion measurements since the last reference frame was labeled. Depending on the video content, this proposed algorithm can achieve from 2% to 13.9% savings in bits while maintaining similar quality.

## 1. INTRODUCTION

The MPEG video compression standard addresses important issues relating to the exploitation of redundancies in video. Spatial, temporal, and coding redundancies are all taken advantage of in the algorithm [4]. In its generic form, however, only a minimal amount of scene adaptation is performed. Indeed, a fixed arrangement of picture types, or temporal decorrelation structures, is the most widely used encoding method. Encoding in this manner reduces complexity, but does not allow for changing temporal statistics. Video can be further compressed by taking advantage of scenes where the statistics allow larger inter-reference frame distances. The additional compression over the fixed picture arrangement style comes from the reduction in the number of reference frames spread throughout the coded sequence.

In order to take advantage of scene content, one needs to perform sequence decomposition based on identifying: 1) scene changes, 2) significant changes within a scene that would require reference frame placement. Lee and Dickinson [2] developed several distance metrics to measure the change in video content, which include: 1) Histogram of Difference Image (HOD) 2) Block Histogram Difference 3) Block Variance Difference. An adaptive size group of pictures (GOP) is examined in [3]. The extreme case of allowing only one independent 'I' reference frame for every GOP was considered.

This paper proposes the use of motion analysis (MA) to adapt to scene content. The motion analysis computes the motion information between every set of adjacent frames. The simple full search block matching algorithm (BMA) is used because it can generate useful statistics for scene change detection. To take care of scene changes, the "mean square prediction error (MSPE)" method, which is similar to that proposed in [2], is adopted. The detection of significant changes within a scene in our algorithm differs from [2] in that we actually measure and accumulate the *amount* and *magnitude* of global/local motion between adjacent frames, which cannot be achieved by the histogramming analysis [2]. A benefit of using MA is the ability to perform local motion tracking. Since motion information is obtained from every set of adjacent frames, one can easily trace the path of motion from the current frame to its reference, which could be many frames away.

The scene adaptation process begins by performing MA on the current frame. Statistics are gained that can lead to easy scene change detection, in addition to the flow vectors from current frame to its neighbor. A picture type decision is made by first examining the scene change statistics for abnormality. Scene changes are handled using 'I' references. If the current frame is not the first frame of a new scene, the process continues and a motion measurement is taken on the flow vectors that is sensitive to the amount of motion. The accumulated motion measurements since the last reference frame is used as the criterion for labeling the current frame a 'P' reference. If all conditions fail, the current frame is labeled a 'B' picture [1].

## 2. CONTEXT DEPENDENT ANALYSIS

**Design Considerations:** This frame type selection algorithm was developed to take into account several different guidelines:

1. A scene change is modeled as a total change in content. An 'I' frame should be used.
2. Global motion (including zooms, pans, and rotations) affects the whole image.
3. Local motion should also be considered when a "significant" portion of the image is affected.
4. Segments of video with significant global motion should yield frequent placement of 'P' reference frames, and vice versa.
5. Local motion requires tracking in order to facilitate ME/MC.

**Motion Analysis:** The BMA is used to estimate motion using a small search range, since movement between consecutive frames is minimal. The MPEG ME/MC step requires both a past and future prediction for 'B' pictures, so motion vectors are obtained from the current frame to the immediate past frame and also immediate future frame.

Scene change statistics are also calculated in this step. The mean square prediction error (MSPE) method is adopted [2]. MSPE is an indication of how well a frame can be predicted from the previous frame.

$$MSPE = \frac{1}{N \cdot M} \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} [O(i,j) - P(i,j)]^2, \quad (1)$$

where $O$ and $P$ represent the original and predicted (compensated) frames respectively. $N$ and $M$ are the frame dimensions.

**Picture Type Decision:** The MSPE gathered in the MA stage is thresholded to determine if a change exists. If scene change is detected, the current frame is the first frame of a new scene and is encoded as an 'I' picture, while the previous frame a 'P' picture to reset the prediction dependencies.

Significant changes within a scene that require placement of a 'P' reference frame are detected by using the forward motion vectors from the MA stage. The first step involves locating the "motion areas" of the image. The "motion areas" are the macroblocks which have a motion vector magnitude greater than a threshold. The number of "motion areas" must be greater than a user defined activity threshold for the algorithm to continue. The "motion measurement" for the frame is calculated as the average motion vector magnitude from all the motion areas. The decision to code the frame as 'P' or 'B' is based on the accumulated "motion measurements" since the 'P' frame. If the accumulated "motion measurement" exceeds a predefined value, then the current frame is labeled a 'P' reference.

**Motion Vector Interpolation:** In addition to picture type labeling, the motion information that is obtained for every set of adjacent frames is useful for the tracking of object and background motion. Normal MPEG motion estimation defines the search center to be the original macroblock spatial coordinates, and a search range surrounding this location. This implies that the search range must be increased as the current frame is moved farther away from its reference. The typical strategy is to multiply each search range dimension for adjacent frames by the frame distance. This motion tracking scheme does not suffer from the same increase in complexity, but it requires repeated interpolations which may incur some error. Interpolation is required because only the motion between adjacent frames is known [1].

**Adaptive Quantization:** The quantization of 'B' pictures is adaptively changed according to the inter-reference frame distance, so that a bit rate reduction can still be attained at a minimal reduction of SNR. Long inter-reference frame distances cause a double coding penalty for 'B' pictures: the number of bits increase while SNR decreases. This is due to the degraded prediction performance and reduction in correlation due to increased distance from the reference frames. For these reasons, an alternate quantization parameter for 'B' pictures is used to keep the SNR at adequate levels when reference frame distances extend farther than a certain length. Otherwise, the original quantization parameter is used.

## 3. SIMULATION RESULTS

The principle objective of the proposed algorithm is to compress raw video more than a fixed GOP arrangement while maintaining quality, as measured by SNR and human perception. SNR is defined as the signal variance divided by the variance in the quantization noise. The standard fixed arrangement of picture types is defined to be 'IBBPBBPBBP...' In areas of high motion, the encoder is defaulted to the fixed arrangement to prevent an unnecessary increase in bit rate. Variable bit rate coding is used since the goal is compression. A different constant quantization parameter is defined for each picture type. 'B' pictures also require an alternate value for video segments that have a large inter-reference frame distance.

### 3.1. Performance Measures

Since the frame selection algorithm operates only based upon the luminance (Y) data, SNR is taken from the luminance component for comparison purposes. When comparing the performance of the scene adaptive encoder to the fixed arrangement of picture types, three measures are scrutinized:

1. Total bits spent on coding all video frames (luminance and chroma).
2. Average 'P' and 'B' picture SNR (dB).
3. Average local 'B'-to-reference SNR ratio, which is defined to be:

$$\frac{B_{SNR}}{(R_{1,SNR} + R_{2,SNR})/2} \quad (2)$$

where $B_{SNR}$ represents either the minimum 'B' picture SNR or the average SNR between two references. $R_1$ and $R_2$ are the past and future references for the 'B' pictures. These can either be 'I' or 'P'. A local measure represents information that is lost in global average measures, such as measure 2.

### 3.2. Parameter Adjustment

The motion threshold and alternate quantization parameter for 'B' pictures must be adjusted to tradeoff bit savings with SNR. The user defined maximum magnitude of motion is thresholded against the accumulated motion since the last reference frame, as explained in the algorithm description. Adjusting this parameter yields different inter-reference frame distances. Long distances mean fewer 'P' reference frames, and therefore higher compression. As the distance increases, though, the number of bits for the 'B' pictures in between the references increases while the SNR drops. The alternate quantization parameter can be used to counter the degradation in quality at the cost of increasing the number of bits.

Long distances between references can also cause difficulties due to object occlusion, acquisition noise, and object/background dissimilarities. Therefore, the inter-reference frame distance should be kept as short as possible, which will facilitate motion vector interpolation and MPEG motion estimation. The alternate quantization parameter should be kept as close to the normal value as possible to prevent over compensation in video segments with small inter-reference frame distances.

### 3.3. Test Video Description

The video used to test the algorithm comes from the public domain. The three sequences are Claire, Garden, and Tennis. Claire and Garden represent video from the two extreme ends of motion. Claire is a newscast, and so it has no global motion, a still background, and low foreground motion. Garden on the other hand has global panning, moderate object motion, object occlusion/exposure, and high detail. Tennis is a table-tennis sequence composed of four segments with differing characteristics. A sample of each is shown in Figure 01. The first segment has moderate local motion, stationary background, and high detail. The second segment has global zooming out, moderate to major local motion, and diminishing resolution due to the zooming. The third segment has a scene change, moderate local motion, stationary background, and low detail. The fourth segment also has a scene change, but it has little local motion, stationary background, and high detail. The first, third, and fourth segments will achieve better compression than the second segment since they lack global motion.

**Scene Change Results:** To test the performance of the adaptive encoder under the condition of a total change in content, a test sequence was created using short segments of Claire, Garden, and Tennis. The full Tennis sequence was also used as a test since it has two scene changes (segments 3 and 4). The MSPE measure for scene change detection successfully identify the two scene changes in each of the experiment sequences by showing strong peaks. From these sequences, it would be easy to determine a fixed threshold for autonomous scene change detection.

**Scene Adaptive Encoder Results:** The adaptive encoder was compared to the fixed arrangement of picture types based on SNR and the savings on the total number of bits. Only one 'I' picture was used (first frame) for each scene unless otherwise specified. In the Claire sequence, the motion threshold was set to 5.0, and the activity threshold was set to 4 macroblocks. The activity threshold was set low in order to capture the local activity of the head and facial features. The bit rate savings is 13.9% over the entire 46 frames. The measure that shows the most degradation in the adaptive method is the minimum 'B'-to-reference SNR ratio. The adaptive encoder is penalized by a 0.1 dB drop, which is very insignificant to human perception when considering the viewing distance and expected frame rate.

The Garden sequence was tested using 20.0 for the motion threshold, and 16 macroblocks for the activity threshold. The bit rate savings is only 2.0% due to the significant global motion involved. The high image detail also works against the adaptive system. The details contribute to high variance in the prediction residuals, and therefore a higher yield of bits. The number of 'P' pictures cannot be reduced as much as in Claire. The frame bitrates increase more dramatically than in Claire as the reference frames are spread farther apart. This can be attributed to the global motion, which affects the entire image rather than a small local region.

In the Tennis sequence, the motion threshold was set to 1.4, while the activity threshold was set to 100 macroblocks. The activity threshold at this level measures global motion. Four 'I' pictures were used in this sequence at frames 1, 23, 68, and 98. Frame 23 is the beginning frame of the global zooming action (the 2nd segment) and is deliberately placed an 'I' frame. Frame 68 and 98 are the first frames of new scenes. This arrangement of 'I' pictures temporally decomposes the Tennis sequence into four segments of different characteristics. Encoding in this fashion yields the result shown in Table 1 and Figure 02. The circled picture types represent how scene changes are coded: 'I' picture for the first frame of the new scene, and 'P' picture for the last frame of the old scene. The bit rate savings is 4.88% over the entire sequence with a 'B' picture SNR difference of about 0.2 dB. The degradation in the 'B' pictures seen as small variations around the moving edges. Overall perceptual quality remains similar to the fixed arrangement, however.

## 4. CONCLUSION

The proposed adaptive encoding scheme is able to satisfy an intuitive coding model: low motion means less reference frames are needed since frame correlation extends over a large span of consecutive images; high motion requires the frequent placement of references to uphold the quality. Simulations on Claire and Garden, sequences from the two extremes ends of motion, demonstrate this principle. Claire, a newscast sequence with little motion, saves up to 13.9%, while Garden, a sequence with considerable global camera panning, saves only 2.0%.

## REFERENCES

[1] A. Y. Lan, *Scene Context Dependent Reference Frame Placement for MPEG Video Coding*, M.S. Thesis, University of Washington, Seattle, March 1996.

[2] J. Lee and B. W. Dickinson, "Temporally Adaptive Motion Interpolation Exploiting Temporal Masking in Visual Perception," IEEE Trans. on Image Processing, 3(5):513-526, September 1994.

[3] H. C. Liu and G. Zick, "Automatic Determination of Scene Changes in MPEG Compressed Video," in Proceedings IEEE Symposium on Circuits and Systems, pp. 764-7 vol.1., Seattle, WA, 1995.

[4] Committee Draft of the Standard: ISO 11172-2, "Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to about 1.5 Mbits/s," Nov., 1991.

|  | Frames | Ave. P SNR (dB) | Ave. B SNR (dB) | Ave. Local B/P (min) | Ave. Local B/P (ave) | Ave. P Size (bits) | Ave. B Size (bits) | Total Size (bits) |
|---|---|---|---|---|---|---|---|---|
| IBBPBBP... | 112 | 19.62 | 17.70 | 0.896 | 0.902 | 81,287 | 14,061 | 4,823,376 |
| Adaptive | 112 | 19.60 | 17.71 | 0.886 | 0.893 | 86,512 | 15,971 | 4,587,976 |

Table 1: Tennis sequence results for fixed and adaptive frame placement.

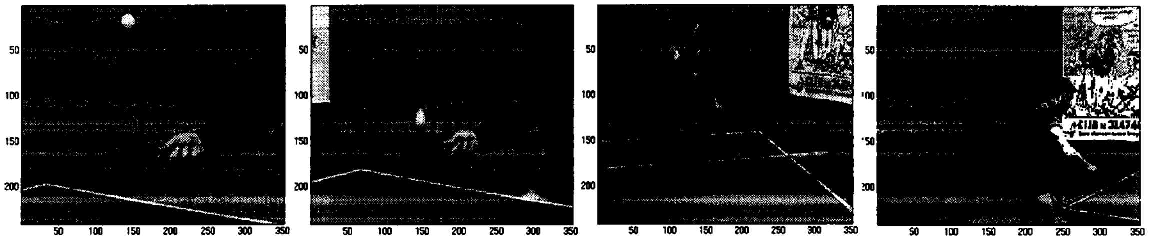

Figure 01: Four segments of the Tennis sequence.

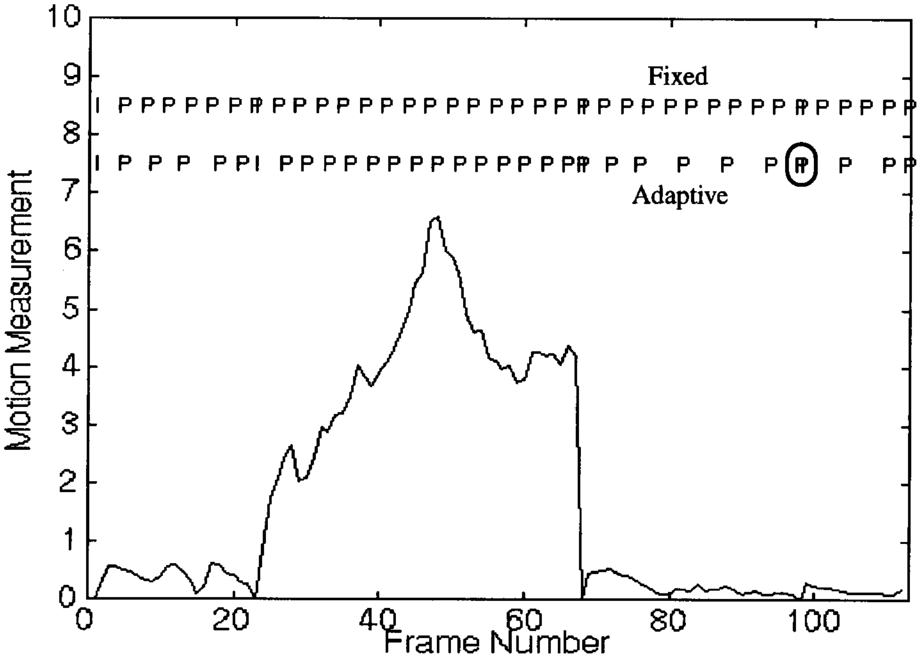

Figure 02: The fixed and adaptive frame type arrangements for Tennis sequence, along with the motion measurements that are made using the motion analysis based algorithm.