I/P/B FRAME TYPE DECISION BY COLLINEARITY OF DISPLACEMENTS

Adriana Dumitraş and Barry G. Haskell

Apple Computer
2 Infinite Loop, MS:302–3KS, Cupertino, CA 95014, United States
Email: adrianad@ieee.org, b.haskell@ieee.org

ABSTRACT

State–of–art encoders in real life applications typically select the number of B frames to be coded between each I or P frame to be equal to one or two by default. The few research works that have addressed the problem of encoder optimization by frame type decision rely on measures of motion magnitude and amount to select the number of B frames. In contrast to such solutions, in this paper we advocate the idea of motion similarity in terms of speed and direction as the basis for deciding how many B frames to encode between any two stored frames. Such similarity is evaluated by the collinearity of the displacements in successive frames. Experimental results using the proposed decision method integrated in our H.264 compliant codec show that the bit rates achieved in the compressed bitstreams are optimal and near–optimal for the tested sequences, and are lower by up to 26% than those obtained using the default number of B frames of typical encoders. Furthermore, such bit rate savings are obtained with no subjective loss in the visual quality of the decoded sequences. Last but not least, our method provides a simple form of temporal scalability and can be employed in any coding system that makes use of I, P and B frames.

1. INTRODUCTION

Selection of the number of Bidirectional Motion Compensated (B) frames to be coded between each Intra (I) or Unidirectional Motion Compensated (P) frames is an encoder decision that affects significantly the bit rate of the compressed bitstreams. State–of–art encoders in real life applications typically select the number of B frames to be coded between each I or P frame to be equal to one or two. This decision is motivated by experimental work, which shows that for most video sequences, such a decision reduces the bit rate without affecting negatively the visual quality of the decoded sequences. The opportunity exists, however, to reduce the bit rate much more for sequences that exhibit slow motion or camera pans by increasing the number of B frames. Moreover, the presence of a higher number of coded B frames makes possible achieving an inherent simple form of temporal/computational scalability in MPEG–1, MPEG–2, H.263+ and H.264 standards by selective decoding of frames at no extra cost in complexity [1].

To the best of our knowledge, the difficulty of the I/P/B decision is one reason that prevents current systems from taking full advantage of coding with a variable number of B frames. Indeed, the appropriate number of B frames to be coded for each sequence not only depends on both the temporal and spatial characteristics of the sequence, but it may also vary across the sequence as the motion characteristics often change. Therefore, different numbers of B frames for different parts of the sequence may be required. Another reason is the increase in the encoder's computational complexity that the implementation of the frame type decision would determine. A brute force approach would simply code every combination of B frames and pick the combination that minimized the bit rate. However, this method is usually far too complex. It also requires a very large number of trial–and–error operations, most of which must be discarded once a final decision is made.

In [2], a method for rate–distortion (RD) optimized frame type selection in MPEG encoding was proposed. The frame type is optimized in the sense that the distortion is minimized under a bit budget constraint. The quantization for each group of pictures is also optimized in the RD sense. The complexity and memory requirements of this method reduce dramatically its chances to be included in a practical (even non–real time) encoder. In [3], the same authors make use of I, P and B frames that have either full or reduced bit allocation. The reduced allocation P frames are inserted when the difference between reference frames increases above a predefined threshold, thereby exploiting the temporal masking performed by the human visual system. Clearly, images with high detail and contrast can easily trigger the placement of a reference frame even when the motion magnitude is small [4]. In [4] and [5], the frame type is selected as P or B depending on the motion characteristics evaluated using the perceived motion energy, and the accumulated motion magnitude and amount, respectively. The former work requires computation of both forward and backward motion vectors, some of which are discarded after the frame type decision. The latter work does not provide bit rate results for the encoded test sequences.

The above works make use of global and local measures of motion and have varying complexity. However, the characteristics of motion that are being exploited are not necessarily those that always motivate the use of several B frames during coding. By contrast, in this paper we advocate the idea of similarity in terms of *motion speed and direction*, as the basis for deciding how many B frames should be coded. We also propose a low complexity algorithm to evaluate such similarity.

2. PROPOSED METHOD

We propose a method for frame type (I/P/B) decision during encoding that yields a variable number of B frames between each pair of stored frames. Our method consists of five steps. These are illustrated in Fig. 1 and are described next.

The first frame of the sequence is encoded as an I frame. For each frame starting with the second frame of the sequence, we detect if there is a scene cut present between the current frame n and the previous frame $n-1$. We identify the scene cuts using the correlation coefficient of each two adjacent frames, given by

$$C = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} x_n(i,\, j)\; x_{n+1}(i,\, j)}{\sum_{i=1}^{M} \sum_{j=1}^{N} x_n^2(i,\, j)\; \sum_{i=1}^{M} \sum_{j=1}^{N} x_{n+1}^2(i,\, j)} \qquad (1)$$

where x_n, x_{n+1} are sample values in two adjacent frames n and $n+1$, and $(i,\, j)$ are the locations of the samples within each frame and $1 \le i \le M$, $1 \le j \le N$. Small values of C indicate that

0-7803-8554-3/04/$20.00 ©2004 IEEE.

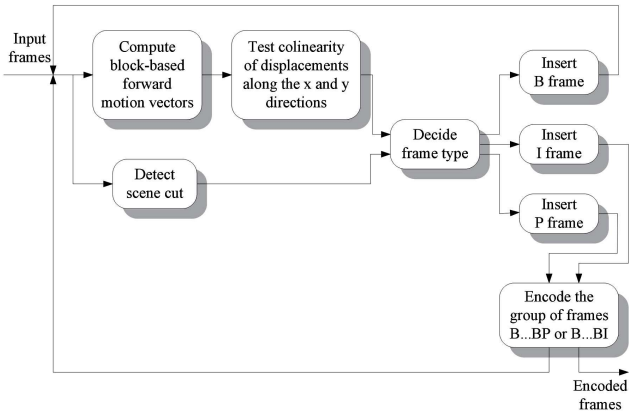


Fig. 1. Block diagram of the proposed encoding method using a variable number of B frames.

two adjacent frames have content that is sufficiently different to be classified as a scene change.

Second, we compute block–based forward motion vectors as if the final coding were to be carried out using all P frames, each referencing the reference frame of the previous group of frames. Let us denote throughout this work by "group of frames" (GOF) any group of B frames and the following P frame in display order (B...BP). Note that a GOF is different than the more commonly used term "group of pictures" (GOP), which includes all (B and P) frames between two successive I frames.

Third, we test the collinearity of the displacements as follows. Three or more points $p_i = (x_i, y_i, z_i)$ are collinear if the distance between one point and the line determined by the other two is zero [6], that is, $|(p_2 - p_1) \times (p_1 - p_3)| = 0$, where $\times$ denotes the cross product operator. Alternatively, three points are collinear if and only if the ratios of their distances satisfy the condition

$$x_2 - x_1 : y_2 - y_1 : z_2 - z_1 = x_3 - x_1 : y_3 - y_1 : z_3 - z_1 \quad (2)$$

Let d_x and d_y be the components of a motion vector (displacements) along the x and y directions. As illustrated in Fig. 2, the displacement points belong to the same plane. Moreover, we assume that the time difference between each pair of successive frames is the same. Under these constraints, we simplify the collinearity condition as discussed next. If a scene cut does not exist between the first frame of a GOF and the preceding frame, we assume that the first frame of the GOF is a B frame (B_1). Starting with frame B_1 for each frame of the GOF at time n we compute the motion speed. As also illustrated in Fig. 2, the motion speed of the block b along the x and y directions is measured by the slopes $S_x(n, b)$, $S_y(n, b)$ and $S(n, b)$, respectively:

$$S_x(n, b) = \frac{d_x(n, b)}{n}, \qquad S_y(n, b) = \frac{d_y(n, b)}{n} \quad (3)$$

$$S(n, b) = S_{x+y}(n, b) = \frac{d_x(n, b) + d_y(n, b)}{n} \quad (4)$$

Starting with frame B_2, for each frame in the GOF we also compute the motion speed error per block b with respect to the motion speed at n = 1 in the GOF:

$$e(n, b) = e_{x+y}(n, b) = S(n, b) - S(1, b) \quad (5)$$

The above steps yield one error value for each motion vector/image block. The mean absolute speed error per frame $E(n)$ at time n is then computed as the mean of all absolute block speed errors within the frame and is given by

$$E(n) = \sum_{1}^{N_{blocks}} \frac{|e(n, b)|}{N_{blocks}} \quad (6)$$

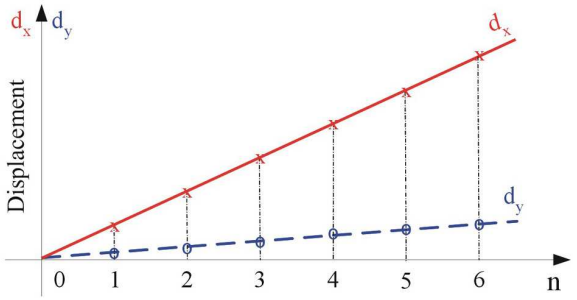


Fig. 2. Collinearity of the displacement points along x and y coordinates for ideal pan sequences.

where N_{blocks} is the number of blocks per frame and is equal to $(M \times N)/b_{size}$ with M and N the frame dimensions, b_{size} the size of the block used for the computation of the motion vectors and $|\ |$ the absolute value operator.

Fourth, we decide the frame type according to the following rules, each of which may contravene the preceding ones:

• IF the error $E(n) < \epsilon_2$, THEN the frame will be coded as a B frame.

• IF the error does not satisfy the above error condition, OR if a scene cut has been detected between the current frame and the previous frame, OR if the maximum number of B frames (N_B) in the group of frames has been exceeded, THEN the current frame will be encoded as a P frame.

• IF a scene cut has been detected between the current frame n and the previous frame $n - 1$, AND the current frame is not the first frame of the GOF, THEN one of the following frame type selections is made: (a) The frame before the scene cut is coded as a P frame and the frame after the scene cut is coded as an I frame: .....P $\|$ I....., where $\|$ marks a scene cut. Each of the I and P frames may be coded at full quality or low quality; (b) The frames before and after the scene cut are coded as P frames: .....P $\|$ P.....Each of the P frames may be coded at full quality or low quality; (c) A few frames before and after the scene cut are B frames, preceded and succeeded (respectively) by at least one P frame: .....P B B $\|$ B B P.....The P frames are coded at full quality and the B frames are coded at low quality.

• IF a periodicity condition for random access applications is met, THEN the current frame is coded as an I frame. We note that the periodicity condition for such applications typically requires frequent insertion of I frames, the frequency of which is determined prior to encoding.

• IF the current frame is the last frame[1] in display order, THEN it is encoded as a P frame.

Fifth, all the frames B...BP in the GOF are encoded with full re–use of the all of the motion vectors computed earlier[2]. Then, the algorithm continues with step two for the next GOF. The maximum number of look–ahead frames that are processed prior to encoding a GOF is equal to $N_B + 1$, since there are at most N_B frames that are B frames (with N_B selected prior to encoding), and at least one P frame in the GOF.

[1]Note that the last frame is known by the encoder for stored sequences. However, it is signaled to the encoder from a client for real–time applications such as videoconferencing.

[2]This is important in terms of computational efficiency, because in most modern coders, the process of motion estimation consumes far more computational resources than any other operation.

Table 1. Test sequences. Notations P, C, Z, X, stand for pans, scene cuts, zooms, and cross fades.

No.	Sequence name	Frame size	No. frames
P1	Synthetic sequence	704×464	36
P2	Discovering Egypt, ch. 6	720×480	99
P3	Discovering Egypt, ch. 1	704×464	99
P4	Discovering Egypt, ch. 2	704×464	99
P5	Golden Eye	704×352	99
P6	The English Patient, ch. 6	704×464	99
P7	Any Given Sunday	704×352	65
C1	Lord of the Rings	480×208	99
C2	Lilo and Stitch	640×384	99
C3	The Lion King	720×384	99
Z1	Lord of the Rings	480×208	23
Z2	Discovering Egypt, ch. 2	704×464	99
Z3	Kono	720×480	99
X1	Lord of the Rings	480×208	99
X2	Lilo and Stitch	640×384	99
X3	Christina Aguillera	320×240	1176

To summarize, the main idea of the proposed method is the evaluation of the motion speed error over successive frames. When the motion speed error is very small, the speed is almost constant, and therefore a higher number of B frames can be inserted. Motion speed need not be constant throughout the entire sequence for our method to be effective, but rather piecewise constant over segments of the video sequence.

3. EXPERIMENTAL RESULTS

The video test set employed in our experiments consists of the high quality, color movie sequences in Table 1. The frame rate is equal to 24 frames per second (fps) for all sequences. The block size employed for the computation of the motion vectors is equal to 16×16 pixels. The quantization parameter values for the I, P and B frames are equal to 25, 26 and 28. One P frame is encoded at full quality after each scene change. We evaluate the effectiveness of the proposed method using the bit rate of the compressed video sequences, the subjective and objective visual quality of the decoded sequences evaluated by visual inspection of the videos and the peak signal–to–noise ratio (PSNR) values (respectively).

For each frame, the frame type is determined during encoding using the method proposed in Section 2 and integrated in our H.264 compliant encoder. Displacements, speed and speed errors for a test sequence are illustrated in Fig. 3. In general, the speed errors are much smaller for pan sequences than for all other types of sequences tested. The ratio of the bit rate obtained by encoding with a variable number of B frames and the bit rate obtained using one B frame for various types of sequences is illustrated in Fig. 4. This figure shows that, for all sequences tested, the former bit rates are lower than or equal to the latter bit rates. In particular, bit rate reductions of up to 26% have been obtained using our method and pan sequences. Subjective evaluation of the decoded sequences obtained by encoding with a variable number of B frames indicates no subjective loss as compared to the visual quality of the sequences encoded using a fixed number of B frames. As illustrated in Fig. 5, the former encoding scenario yields a reduction[3] of up to 0.55 dB in the PSNR as compared to the latter. Examples of frame type decisions using our method are illustrated in Fig. 6.

[3]Most likely due to the fact that more B frames are typically encoded using our method, all of which have a quantization parameter value that is higher than that of the P frames.

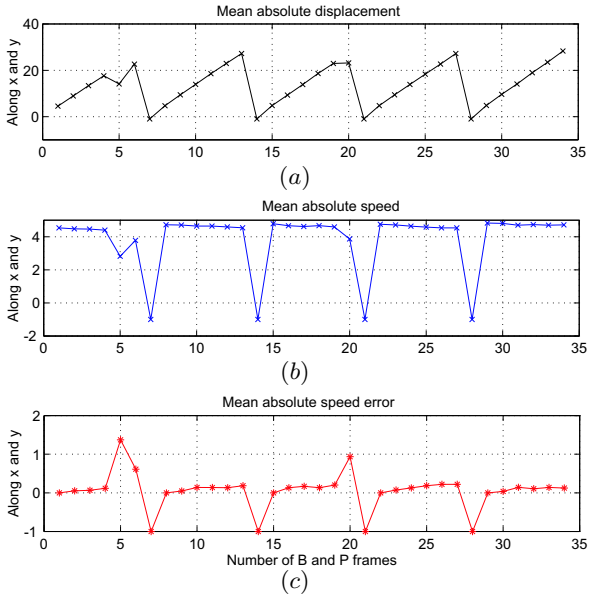


Fig. 3. (a) Displacements, (b) speed and (c) speed errors along coordinates x and y in sequence P1. P frames are marked using the value -1.

4. CONCLUSIONS

We have proposed an effective method for the frame type (I/P/B) selection during encoding. The proposed method makes use of motion similarity in terms of speed and direction for frame type decision and yields a variable number of B frames to be coded between any two stored frames. In the framework of an H.264 compliant codec, we have shown that the correct (and variable) number of B frames is selected by our method such that the bit rate is reduced with no loss in the subjective visual quality of the decoded frames. This makes the proposed method suitable for any coding system that makes use of I, P and B frames.

5. REFERENCES

[1] Barry G. Haskell, Atul Puri, and Arun N. Netravali, *Digital Video: An Introduction to MPEG–2*, Chapman and Hall, USA, 1997.

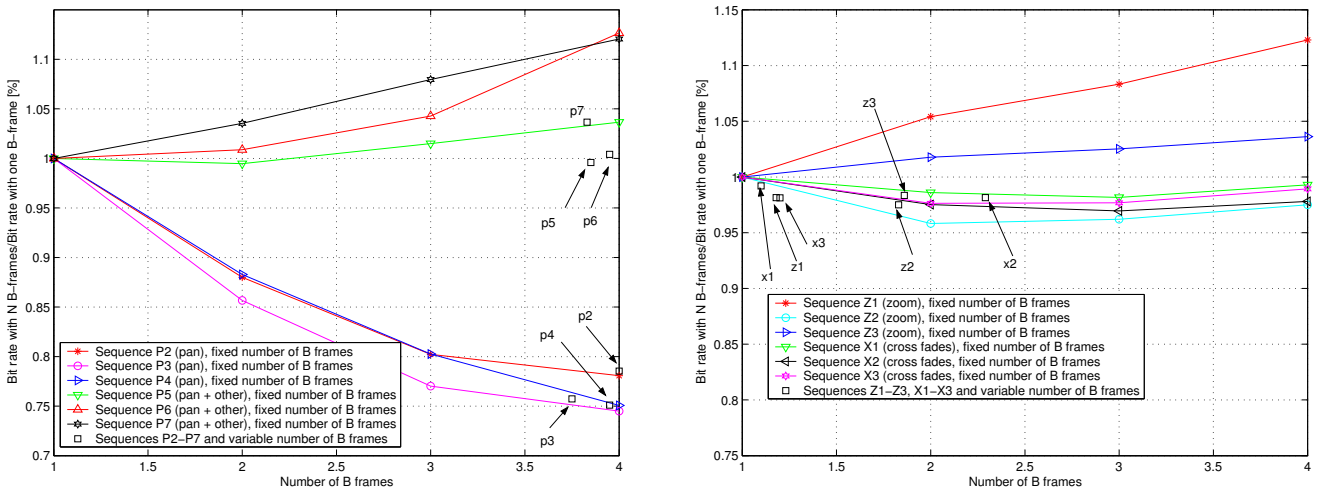
[2] Jungwoo Lee and Bradley W. Dickinson, "Rate-distortion optimized frame type selection for MPEG encoding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 7, no. 3, pp. 501–510, June 1997.

[3] Jungwoo Lee and Bradley W. Dickinson, "Temporally adaptive motion interpolation exploiting temporal masking in visual perception," *IEEE Transactions on Image Processing*, vol. 3, no. 5, pp. 513–526, Sept. 1994.

[4] Anthony Y. Lan, A.G. Nguyen, and J-N Hwang, "Scene-context-dependent reference-frame placement for MPEG video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 3, pp. 478–489, Apr. 1999.

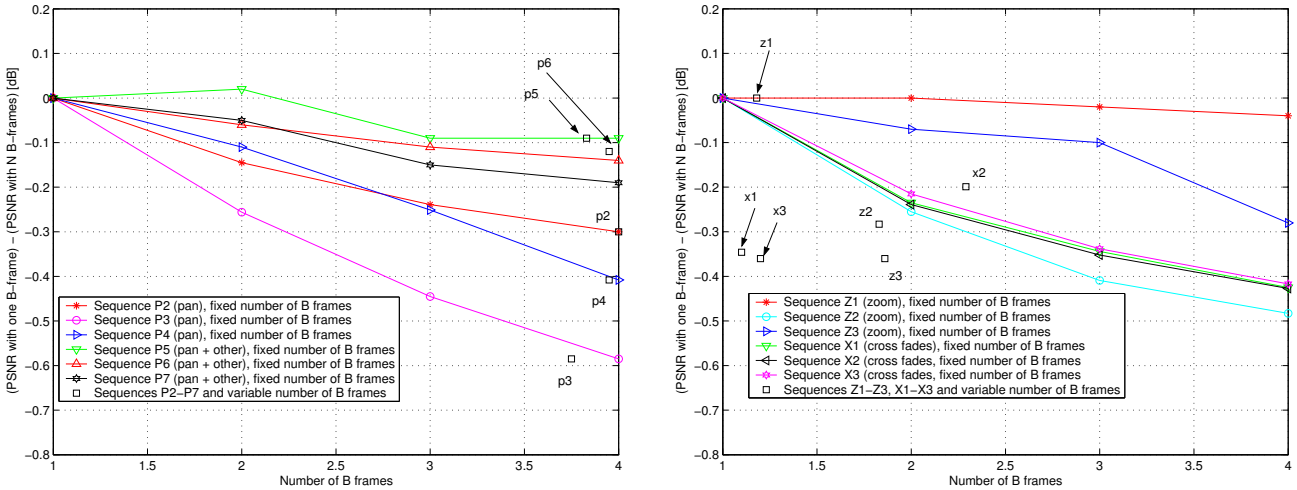
[5] Xiaodong Gu and Hongjiang Zhang, "Implementing dynamic GOP in video encoding," in *IEEE Intl. Conference on Multimedia and Expo (ICME), Baltimore*, 2003, vol. 1, pp. 349–352.

[6] Eric Weisstein, *CRC Concise Encyclopedia of Mathematics*, CRC Press, USA, 2002.



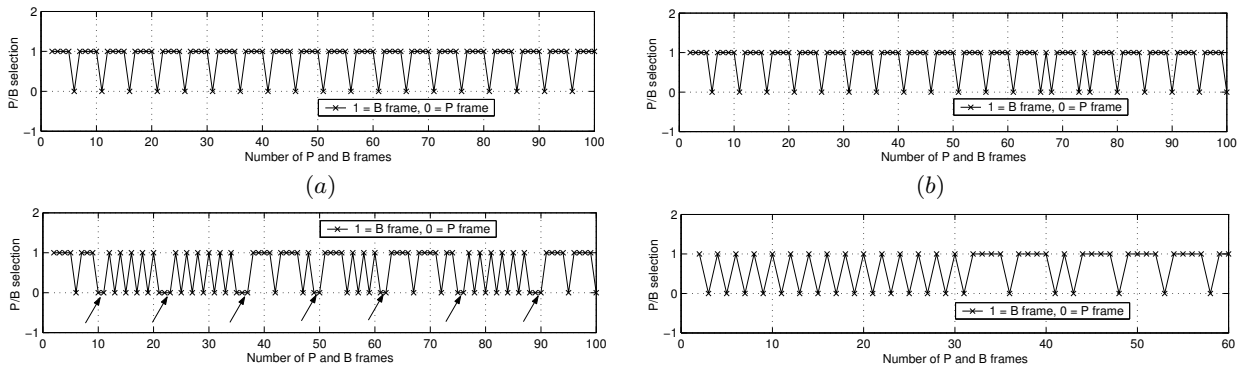
(a) (b)

Fig. 4. Ratio of the bit rate using a variable number of B frames and the bit rate using one B frame for (a) pans and (b) zooms and cross fades. The points marked with lower cases indicate the bit rate obtained using a variable number of B frames for the corresponding sequence in upper case (e.g., p1 is a point that refers to sequence P1).



(a) (b)

Fig. 5. Difference between the peak signal–to–noise ratio (PSNR) using one B frame and the PSNR using a variable number of B frames for (a) pans and (b) zooms and cross fades. The points marked with lower cases indicate the PSNR obtained using a variable number of B frames for the corresponding sequence in upper case.



(a) (b) (c) (d)

Fig. 6. Frame type selection using a variable number of B frames for pan sequences (a), P2, (b) P4, (c) C2, and (d) Z3. Each arrow in (c) indicates the frame after a scene cut. B and P frames are marked using the values 1 and 0, respectively.