ESTIMATION OF FADE AND DISSOLVE PARAMETERS FOR WEIGHTED PREDICTION IN H.264/AVC

Fatih Kamisli *David M. Baylon*

Massachusetts Institute of Technology, Cambridge, MA Motorola, Inc. San Diego, CA

ABSTRACT

Weighted prediction (WP) is one way of overcoming the limitations of simple motion compensation for scenes with gradual transitions such as fades or dissolves. In WP, the predictions for inter coded blocks are obtained from scaled versions of the reference frames. H.264/AVC is the first video coding standard that has incorporated WP tools. In this paper, we focus on the estimation of weights for WP in dissolves. Our findings indicate that the estimation approaches for dissolves should be different from the estimation approaches for fades. Specifically, estimating the weights jointly for the two reference frames of B-frames gives better performance for dissolves under most circumstances.

Index Terms— Video coding, motion compensation, parameter estimation

1. INTRODUCTION

Video compression is heavily dependent on exploiting the temporal correlation between adjacent frames. Motion compensation (MC) is the most widely used technique for exploiting this correlation. In some gradual scene transitions such as fades or dissolves, however, the regular motion compensation technique is not very efficient as it cannot model such scenes well. Dissolves are gradual transitions where the transition occurs from one scene to a different second scene. In fades, one of the two scenes in the transition is a constant intensity frame such as black. To overcome the shortcomings of MC for such scenes, weighted prediction (WP) has been proposed [1,2]. In WP, scaled versions (by a multiplicative constant and an additive offset) of reference frames are used for MC.

H.264/AVC is the first standard to include WP tools [2]. The determination of the weights for WP is performed at the encoder and is open to encoder design. It is desirable to have precise and efficient estimation methods. Methods have been proposed to estimate weights [1-4]. However, the distinction of the estimation approaches in fades versus dissolves has not been emphasized. In this paper, we report on our findings which indicate that the estimation approaches of weights could be different for fades and dissolves. Specifically, estimating the weights jointly for the two reference frames of B-frames gives better performance for dissolves under most circumstances. For fades, however, estimating the weights separately gives better performance.

The remainder of the paper is organized as follows. In Section 2, we discuss methods to estimate weights for WP. In Section 3, we briefly review the WP tool in H.264/AVC. Our simulation results and related discussions are presented in Section 4. We conclude the paper in Section 5.

2. METHODS TO ESTIMATE WEIGHTS

2.1. Method 1: Ratio of Luminance DC's

In this technique [2], used in the H.264/AVC reference software, it is assumed that the luminance DC of the original non-faded frames does not change over small temporal distances (e.g. from I_0 to P_3 in Figure 1) and therefore the luminance DC change between the faded frames is due to the fade. Hence, the weight that scales I_0 (Figure 1), which is then to be used as a reference for P_3, is estimated as follows:

$$w_1 = DC(P_3)/DC(I_0) \qquad (1)$$

Similarly, the weight w_1 (w_2) that scales P_6 (P_9), which is then to be used as a reference for B_8 (B_8), is estimated by dividing the luminance DC of B_8 (B_8) by the luminance DC of P_6 (P_9). Note that in this latter case, w_1 and w_2 are estimated separately. In other words, w_1 is computed using information only from P_6 and B_8, and w_2 is computed using information only from B_8 and P_9, and therefore w_1 is independent of P_9 whereas w_2 is independent of P_6.

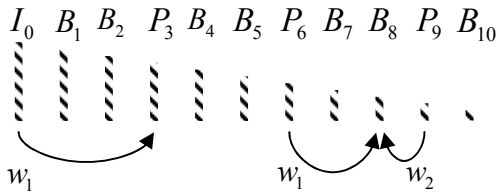

Figure 1- Frames in fade. Diminishing bars indicate the fade.

2.2. Method 2: Minimization of Mean-Square-Error

In this technique [1,3], weights are obtained by minimizing the MSE between the frame that is to be predicted (P_3) and the prediction for it (w_1*I_0), as shown in Figure 1. Ideally, it is desirable to use the motion-compensated I_0. However, this may require performing MC twice or developing a joint weight and motion vector estimation algorithm. Therefore, for simplicity, we use directly I_0 and not the motion compensated I_0. The MSE expression is then as follows:

$$MSE = \sum_{\text{i: all pixels}} \left(P_3(i) - w_1 \cdot I_0(i)\right)^2 \qquad (2)$$

The above expression is minimized by the following value of w_1:

$$w_1 = \sum_{\text{i: all pixels}} \left(P_3(i) \cdot I_0(i)\right) \Big/ \sum_{\text{j: all pixels}} I_0^2(j) \qquad (3)$$

For the estimation of weights for frame B_8, there are two options. The first option is to consider P_6 and B_8 to estimate w_1 and then P_9

and B_8 to estimate w_2. In other words, in this option two different MSE expressions (both in the form of (2)) are formed as follows:

$$MSE_1 = \sum_{\text{i: all pixels}} \left(B_8(i) - w_1 \cdot P_6(i)\right)^2 \quad (4)$$

$$MSE_2 = \sum_{\text{i: all pixels}} \left(B_8(i) - w_2 \cdot P_9(i)\right)^2 \quad (5)$$

Values of w_1 and w_2 that minimize these expressions are obtained similar to (3). The estimation of w_1 is independent of P_9, and the estimation of w_2 is independent of P_6. We refer to the estimation method in this first option as separate estimation of weights.

The second option for estimating the weights for frame B_8 is to form one MSE expression, which includes both w_1 and w_2:

$$MSE = \sum_{\text{i: all pixels}} \left(B_8(i) - w_1 \cdot P_6(i) - w_2 \cdot P_9(i)\right)^2 \quad (6)$$

The values of w_1 and w_2 minimizing the above expression are computed as shown below in (7) and (8):

$$w_1 = \frac{\left(\sum B_8(i)\cdot P_9(i)\right)\cdot\left(\sum P_6(i)\cdot P_9(i)\right) - \left(\sum P_9^2(i)\right)\cdot\left(\sum B_8(i)\cdot P_6(i)\right)}{\left(\sum P_6(i)\cdot P_9(i)\right)^2 - \left(\sum P_9^2(i)\right)\cdot\left(\sum P_6^2(i)\right)} \quad (7)$$

$$w_2 = \frac{\left(\sum B_8(i)\cdot P_6(i)\right)\cdot\left(\sum P_6(i)\cdot P_9(i)\right) - \left(\sum P_6^2(i)\right)\cdot\left(\sum B_8(i)\cdot P_9(i)\right)}{\left(\sum P_6(i)\cdot P_9(i)\right)^2 - \left(\sum P_9^2(i)\right)\cdot\left(\sum P_6^2(i)\right)} \quad (8)$$

Note that in this option, both w_1 and w_2 depend on B_8, P_6 and P_9. Therefore, we refer to this estimation method as joint estimation of weights.

To provide an intuitive reasoning for the superiority of joint estimation of weights for dissolves, consider Figure 2, where a segment of a dissolve is shown. The dissolve goes from the sequence G to the sequence R. Each frame in the dissolve contains contributions from both sequences. The contribution from the sequence G decreases over time while the contribution from the sequence R increases over time. The contributions of each sequence in each frame are shown below these frames; for example for frame B_5, we have 0.4*G and 0.6*R. Now let us assume that a prediction for frame B_5 is made using frames P_3 and P_6. The correct weights w_1 and w_2, which should be used to weight P_3 and P_6, are 1/3 and 2/3 respectively. If we employ separate weight estimation (either Method 1 or separate weight estimation technique of Method 2), then it is unlikely that we can obtain 1/3 for w_1 and 2/3 for w_2. With joint estimation of weights, however, weights that are very close to 1/3 and 2/3 can be obtained.

Note that an analogy between separate and joint estimation of motion vectors (MV) and separate and joint estimation of weights for WP can be drawn. To estimate the MV's for frame B_5, the conventional way is to estimate the backward MV using B_5 and P_3, and the forward MV using P_6 and B_5. This is separate MV estimation. However, it is also possible to estimate the forward and backward MV's in a joint manner [6]; i.e. one can use P_3, B_5 and P_6 to estimate both the forward and backward MV's. The estimation of weights for WP in a separate or joint manner is a similar problem.

3. WEIGHTED PREDICTION TOOL IN H.264/AVC

A nice overview of the WP tool in H.264/AVC can be found in [2]. Here, we provide a brief discussion of the relevant aspects. There are two available modes for WP in H.264/AVC: explicit mode and implicit mode. In explicit mode, the weights that scale the reference frames (and also additive offsets to shift the DC's of the reference frames) are transmitted in the bitstream. In implicit mode, the weights are not transmitted; they are derived from the picture-order-counts (POC) of the reference frames. For P-frames, only explicit mode can be used. For B-frames, either explicit or implicit mode can be used. The weights derived from the POC's in B-frames are inversely proportional to the POC differences of the reference frames and the frame to be predicted. For example, in Figure 2, if we assume that the subscripts of the frames denote the POC's, then w_1 and w_2 can be derived as shown below in (9). Note that, for linear dissolves, the weights obtained from POC's turn out to be the correct weights, which were discussed in Figure 2.

$$w_1 = \frac{6-5}{6-3} = \frac{1}{3} \qquad w_2 = \frac{5-3}{6-3} = \frac{2}{3} \quad (9)$$

For P-slices, a multiplicative weight and additive offset for each color component may be coded for each of the allowable reference pictures in list0. For B slices, each of the allowable reference pictures in list0 and list1 may have one weight and an additive offset coded. Coding only one weight for each reference picture may be sufficient for fades. For dissolves, however, the weight to use for one reference picture depends also on which reference picture is used from the other list, as was discussed in Section 2. Therefore, allowing to code more than one weight for each allowable reference picture in a list could be useful for dissolves. However, there is also the additional overhead due to coding more weights. The flexibility of changing the weight of a reference picture, depending on which reference picture is used from the other list, is not available in the explicit mode, but it is present in the implicit mode. In the implicit mode, since the weights are derived from the POC's of the reference pictures, the weight used for a reference picture changes depending on which picture is used from the other list. Therefore, the weights in implicit mode are more adaptive (in the sense that there can be as many weight combinations as there are reference picture combinations from list0 and list1) when compared to explicit mode.

For B-frames, the encoder has also the option to make predictions using only one reference picture (as in P-frames), either from list0 or from list1. In this case, there is a difference in the use of weights in explicit mode and implicit mode. In explicit mode, the H.264/AVC specification requires to use the weight that is coded for that particular reference picture. In the implicit mode, however, the specification requires that the default weight of one

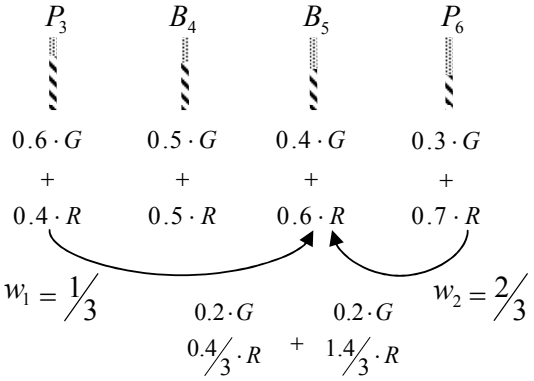

Figure 2– Frames in dissolve. Diagonal hatched portion of the bars indicate the contribution of the sequence G and the dotted portion of the bars indicate the contribution of the sequence R.

is used; this is equivalent to using no weight at all. For dissolves, using the default weight may be a better choice, we believe, because if the weight coded in explicit mode is calculated assuming that the prediction is to be formed from two reference pictures, but the prediction is formed from only one reference picture, then this weight is not reasonable anymore. The AVC specification does not allow for a default weight in this case of prediction from only one reference picture in the explicit mode.

4. SIMULATION RESULTS AND DISCUSSIONS

We present our simulation results in three sub-sections. In Sub-sections 4.1 and 4.2, we report on simulations performed on dissolves. First, we impose some constraints in order to evaluate directly the quality of the computed weights for B-frames (Sub-section 4.1), and then remove these constraints (Sub-section 4.2). In Sub-section 4.3, we report on simulations for fades.

A traditional PBB picture pattern, two reference frames, RD-optimization and a frame rate of 30 fps are used in the experiments. We use fixed quantization and run each of the experiments on four QP's: 28,30,32,34. Using the resulting PSNR and bit-rates, we obtain the Bjontegaard delta bit-rate (BD-bitrate) savings [5], which we use throughout as a quantitative measure for making comparisons. The sequences we use are dissolves and fades of foreman (CIF) and coastguard (CIF). The dissolves and fades are of different lengths and are either linear or exponential.

4.1. Forced prediction mode for B-frames in dissolves

We want to evaluate how good the weights are for each of the estimation techniques for B-frames in dissolves. Therefore, we force the mode of each macro-block in B-frames to be B_Bi_16x16. This means that each of these macro-blocks will be predicted from two reference frames using 16x16 blocks; no intra or any other inter mode is used. Without forcing the mode, the encoder may decide to use intra coding or other type of coding that is more useful from a rate-distortion point of view and therefore the comparisons may not reflect well the direct performances of the weights the estimation techniques produce.

Figure 3 provides the BD-bitrate savings compared to the case where WP is turned off for a linear dissolve of 60 frames. Since we want to focus on the performance of weight estimation for B-frames, we turn off WP for P-frames. Method 1, which is a separate estimation technique, performs worst. The joint estimation technique of Method 2 performs best, while the implicit mode, which is also a joint estimation technique, performs intermediate. These numbers are generated using the PSNR and bit-rates obtained from the entire sequence (I-frame, P-frames and B-frames). However, as mentioned, we use WP only for B-frames. Therefore, in Figure 4, we provide the BD-bitrate savings obtained only from the PSNR and bitrates of B-frames. In this case, the gains obtained from WP increase by a factor of about 3 and can go up to 12%. These results indicate that joint weight estimation techniques can provide better weights for B-frames in dissolves.

4.2. Weighted prediction in dissolves

We now remove the forcing of modes and let the encoder decide on the mode in an RD-optimized manner. Figure 5 shows result for a 30-frame-long exponential dissolve and Figure 6 shows results for a 16-frame-long linear dissolve. Looking at these two figures, two observations can be made. First, the tallest three bars in both figures use joint weight estimation techniques (implicit mode or joint technique of Method 2) for B-frames. The other shorter bars use either separate weight estimation techniques (Method 1 or 2) or no WP at all for B-frames. Second, the gains achievable depend on the dissolve length. Gains increase for shorter dissolves. This is expected since longer dissolves have slower transitions which do not hurt regular MC as much. These two figures reflect PSNR and bitrate results obtained from all frames in the dissolves, i.e. I-frame, P-frames and B-frames. Considering the fact that WP is much more effective for B-frames than P-frames (due to the two-sided estimation opportunity), we also show in Figure 7 the BD bitrate savings for only B-frames of the dissolve of Figure 6. The gains in this case increase by roughly 3 and can go up to 50%.

Macroblocks in B-frames may also be coded in modes that use prediction from only one reference frame. In such cases, as discussed in Section 3, the implicit mode uses the default weight of one, while the explicit mode uses the transmitted weight. However, if this weight is computed assuming two-sided prediction, then it is not a reasonable weight if it is used in prediction from a single frame. This is the reason why among the three tallest bars (in Figures 5, 6 and 7) the explicit mode (P:m2 B:m2j) performs slightly worse than the implicit modes. This reasoning is supported by the number of modes selected in B-frames for each experiment. In the explicit mode experiment, the number of single-directional prediction modes in B-frames is by far less than in the implicit mode experiments (P:m1 B:imp, P:m2 B:imp).

4.3. Weighted prediction in fades

Figure 8 shows results for a 30-frame-long exponential fade. We have two observations. First, the gains achievable are much larger for fades than for dissolves. This is, first, due to the fact that the dynamic range of frames decrease towards zero during fades, and second, due to the fact that there is only one sequence in fades while there are actually two superimposed sequences in dissolves, making dissolves more difficult to compress. Our second observation is that for fades, separate weight estimation techniques perform better than joint techniques. The reason behind this is that joint estimation is not as important for fades as for dissolves since there is only one sequence in fades while there are two superimposed sequences in dissolves. In addition, a difference exists between the uses of weights in case of prediction from one frame in B-frames. In those cases, the use of the weights in the explicit mode is preferable.

5. CONCLUSIONS AND FUTURE WORK

Weighted prediction is a very useful tool for coding fades and dissolves. Our results indicate that estimation could be different in fades and dissolves. While separate weight estimation techniques perform better for fades, joint weight estimation techniques perform better for dissolves under most circumstances.

One future research direction could be to combine joint weight estimation with joint motion vector estimation. In the experiments, even though weights were estimated jointly, forward and backward motion vectors were estimated separately. If forward and backward motion vectors and weights are all estimated jointly, further improvements could be achieved.

6. REFERENCES

[1] K. Kamikura, H. Watanabe, H. Jozawa, H. Kotera, S. Ichinose, "Global brightness-variation compensation for video coding," *IEEE Trans. CSVT*, Vol. 8. No. 8, p. 988-1000, Dec 1998.

[2] J. Boyce, "Weighted prediction in the H.264/MPEG4 AVC video coding standard", *ISCAS*, pp. 789-792, May 2004

[3] Kato, H.; Nakajima, Y., "Weighting factor determination algorithm for H.264/MPEG-4 AVC weighted prediction," *Multimedia Signal Processing, 2004 IEEE 6th Workshop on* , pp. 27- 30, 29 Sept.-1 Oct. 2004

[4] Koto, S.; Chujoh, T.; Kikuchi, Y., "Adaptive bi-predictive video coding using temporal extrapolation," *Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on* , vol.3, pp. III- 829-32 vol.2, 14-17 Sept. 2003

[5] G. Bjontegaard, Calculation of average PSNR Differences between RD-curves*, VCEG-M33*, Austin, TX, March 26, 2001

[6] Siu-Wai Wu Gersho, A "Joint estimation of forward and backward motion vectors for interpolative prediction of video," *IEEE Trans. Image Processing*, Vol.3, Issue 5, p.684-687 Sep. 1994

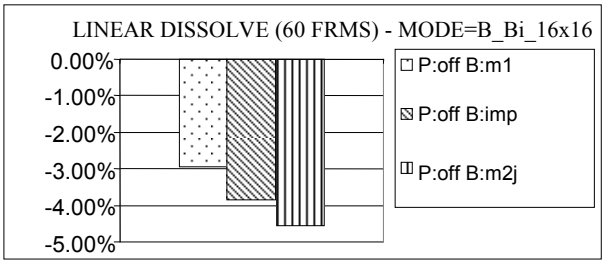

Figure 3* – BD-bitrate savings with respect to no WP. All macroblocks in B-frames have mode=B_BI_16x16. Dissolve includes first 60 frames of foreman and coastguard and goes from foreman to coastguard.

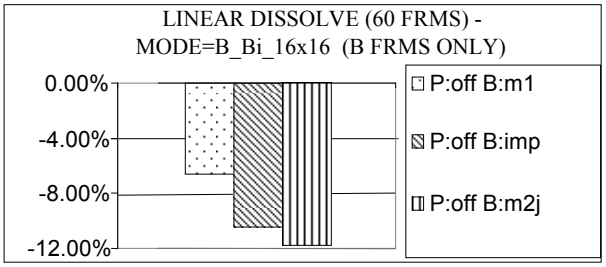

Figure 4 – BD-bitrate savings with respect to no WP, computed only from B-frames. All macroblocks in B-frames have mode= B_BI_16x16. Dissolve includes first 60 frames of foreman and coastguard and goes from foreman to coastguard.

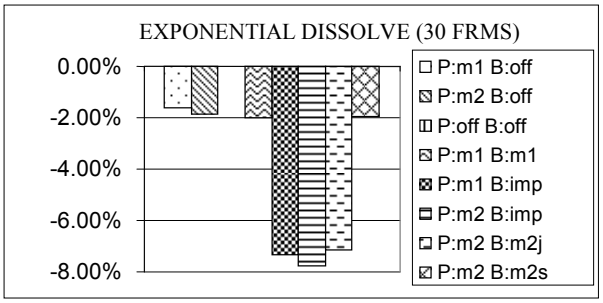

Figure 5– BD-bitrate savings with respect to no WP. Dissolve includes first 30 frames of foreman and coastguard and goes from foreman to coastguard.

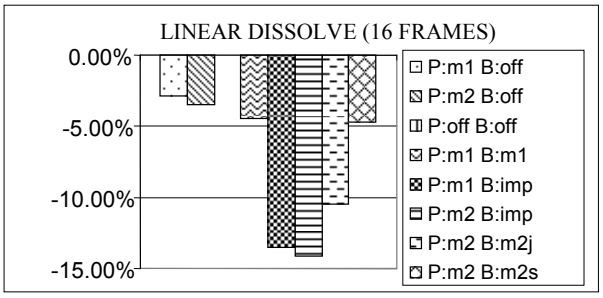

Figure 6 – BD-bitrate savings with respect to no WP. Dissolve includes first 16 frames of foreman and coastguard and goes from foreman to coastguard.

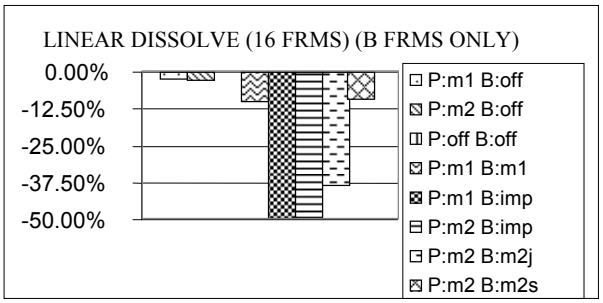

Figure 7 – BD-bitrate savings with respect to no WP, computed only from B-frames. Dissolve includes first 16 frames of foreman and coastguard and goes from foreman to coastguard.

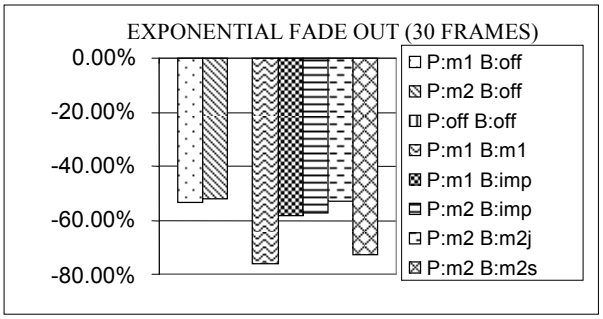

Figure 8 – BD-bitrate savings with respect to no WP. Fade includes first 30 frames of foreman and goes to black.

* In the legends of all figures, "P:" and "B:" describe the WP modes of P- and B-frames. 'off', 'imp', 'm1', 'm2', 'm2j', 'm2s' refer to no WP, implicit mode, Method 1, Method 2, joint technique of Method 2 and separate technique of Method 2, respectively.