LOCALLY OPTIMAL, BUFFER-CONSTRAINED MOTION ESTIMATION AND MODE SELECTION FOR VIDEO SEQUENCES

C. B. Peel, S. E. Budge

Electrical and Computer Engineering Dept.
Utah State University
Logan, UT 84322-4120
chris.peel@ieee.org, scott@ece.usu.edu

K. M. Liang and C.-M. Huang

Sorenson Vision, Inc.
1011 West 400 North
Logan, UT 84321
{kyminhl,chien-min}@s-vision.com

ABSTRACT

We describe a method of using a Lagrange multiplier to make a locally optimal trade off between rate and distortion in the motion search for video sequences, while maintaining a constant bit rate channel. Simulation of this method shows that it gives up to 3.5 dB PSNR improvement in a high motion sequence. A locally rate-distortion (R-D) optimal mode selection mechanism is also described. This method also gives significant quality benefit over the nominal method. Though the benefit of these techniques is significant when used separately, when the optimal mode selection is combined with the R-D optimal motion search, it does not perform much better than the codec does with only the R-D optimal motion search.

1. INTRODUCTION

A fact that is readily apparent to researchers in low bit-rate video compression is that as the bit-rate decreases, the percentage of motion offsets in the compressed bitstream increases. We propose a motion estimation method which makes the motion displacement coding more efficient, allowing more bits for coding of motion residuals, thus improving the quality. Because this efficiency has more of an impact when the motion displacements constitute a high percentage of the bitstream, a more significant gain will be observed at low bit-rates. We also propose a method for making a locally optimal block mode decision, and show how it interacts with the more efficient motion coding method.

Simulations were carried out with a vector quantization (VQ) based video codec. Constant bit rate output was maintained with the use of buffer-constrained bit allocation. The motion coding system is similar to that used for the H.263 [1] video compression standard. The general scheme for the system is that of an input block passing to a quantizer, and quantization bits passing through a buffer before reaching the channel (see Figure 1).

1.1. Hierarchical VQ

The theoretical advantage of VQ over scalar quantization is well documented [2]. A practical advantage is not as obvious; finite computational and storage resources constrain implementable vector quantizers. We used hierarchical residual VQ for the coding core of our studies [3]. This technique provides a good trade-off between computational complexity, storage, and quality. Residual VQ requires less storage and computational resources than full-search VQ [4]. Hierarchical VQ allows a wide range of bit-rates to be achieved [5].

Macroblocks are quantized at several different sub-block sizes and with several stages of residual VQ. A rate-distortion cost (described in the next section) is used to determine the best permutation of sizes and stages for a macroblock. A Huffman-coded header is transmitted, indicating the decisions made, followed by the appropriate codebook indices.

1.2. Lambda Feedback

Buffer-constrained bit allocation [6, 7] is used to enable a constant bit-rate channel. Figure 1 shows how the buffer fullness is fed back to control the quantization. A Lagrange multiplier λ is created as a function of the buffer fullness. This is used to make the best (per macroblock) decision by minimizing the cost equation $D + \lambda R$ for each quantization option, given a distortion D and rate R.

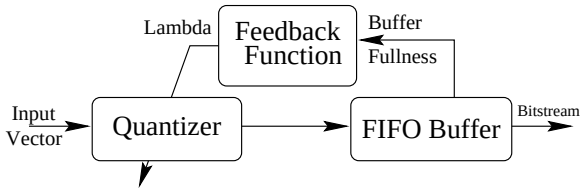

Figure 1: Buffer fullness feedback.

A more formal exposition follows. Suppose that we have a sequence of blocks to be coded $\mathbf{X}_n, n = 1, 2, ..., N$, and possible quantization options $Q_l(\cdot), l = 1, 2, ..., L$. Distortion induced for coding $\mathbf{X}_n$ with quantization level $Q_l(\cdot)$ is indicated by $D_{n,l} = d(\mathbf{X}_n, Q_l(\mathbf{X}_n))$ and bits used by $R_{n,l}$. A locally (per block) rate-distortion optimal coding decision C_n is made by minimization of the cost $J_{VQ}(n, l)$ over possible coding options l. The Lagrange minimization parameter λ is a function of the buffer fullness B_{n-1} obtained after coding the previous block, introducing a buffer constraint into the minimization:

$$D_{n,l} + \lambda(B_{n-1}) R_{n,l}. \quad (1)$$

This "lambda feedback" is the mechanism by which a constant bit-rate channel is achieved [8]. All simulations use lambda feed-

This work was supported by Sorenson Vision, Inc.

Figure 2: Luminance of decoded `foreman` image when a distortion cost is used in the motion estimation (2484 total bits used).

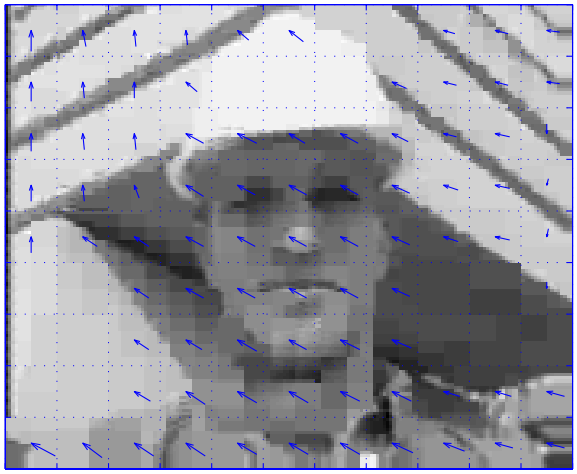

Figure 3: Luminance of decoded `foreman` image when a R-D cost is used in the motion estimation (899 total bits used).

back for the VQ coding. We will use the terms "lambda feedback" and "rate-distortion decision" interchangeably.

1.3. Macroblock Modes

A macroblock is coded with one of several modes. These modes include a no-update mode, where a block may be copied from the previous frame, with no further bits transmitted, and intra coding of a macroblock with VQ. Two motion modes are also available: motion on 16x16 blocks, and motion on 8x8 blocks. For the 16x16 case, one entropy-coded motion displacement is transmitted, while for the 8x8 case, four displacements are encoded. In each motion case, the residual block may be further refined by VQ.

2. BUFFER-CONSTRAINED MOTION COMPENSATED PREDICTION

It has been common to do motion estimation by minimizing an absolute error or squared error distortion measure over possible displacements [9]. This method works well, but has several problems. The most noticeable is that of poor performance at low bit rates or with sequences with large motion displacements. Figure 2 shows the noisy motion fields produced when using an absolute error criterion.[1]

Many researchers have investigated methods of incorporating rate as well as distortion into the motion estimation problem [10]. We are interested in the slightly different problem of minimizing distortion as we maintain a constant bit rate. In other words, we desire to apply buffer-constrained bit allocation to motion estimation.

[1]The images shown in Figure 2 and Figure 3 are the second output frames of simulations done at 20 Kbps, with QCIF data at 10 fps. An AVI file can be found under "Papers" at `http://www.engineering.usu.edu/ece/staff/peel/`.

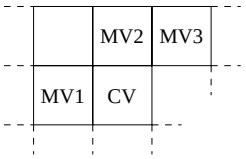

Figure 4: The H.263 median predictor.

2.1. Lambda Feedback in Motion Estimation

The H.263 median predictor [1] is used to code motion displacements in all simulations. This predictor is formed from three blocks in the causal context of the block in question. Figure 4 shows the motion vectors $MV1$, $MV2$, and $MV3$, which are in the causal neighborhood of the current vector CV. A predictor $PV = Median(MV1, MV2, MV3)$ is formed, and $CV - PV$ is the value transmitted if a motion mode is chosen. An intuitive explanation of the benefit of this method is that PV provides a way to indicate that we want the motion vectors for an object to be "similar" to surrounding motion vectors. This predictor can also be used during motion estimation to indicate how many bits would be used to code the motion vector at a particular displacement.

$$D_{n,v} + \lambda(B_{n-1})R_{n,v}. \tag{2}$$

The motion displacement which minimizes equation (2) is chosen. This is the same as described above for the VQ quantizer, with the substitution that the quantizer level l is now a possible motion displacement v. Note that $R_{n,v} = T_{Huff}(CV - PV)$, where T_{Huff} is a table of bits used to code a predicted motion offset. It is significant that this table lookup is very simple, and is overshadowed by the computational cost of calculating $D_{n,v}$. With this method, the decision between 16x16 and 8x8 motion modes, as well as the motion vector determination, is made with a R-D cost.

Figure 3 shows the decoded image and motion field when using lambda feedback. There were 480 bits used for motion displacements in the lambda feedback case, while 2119 were used in the distortion minimization case shown in Figure 2. Thus, less

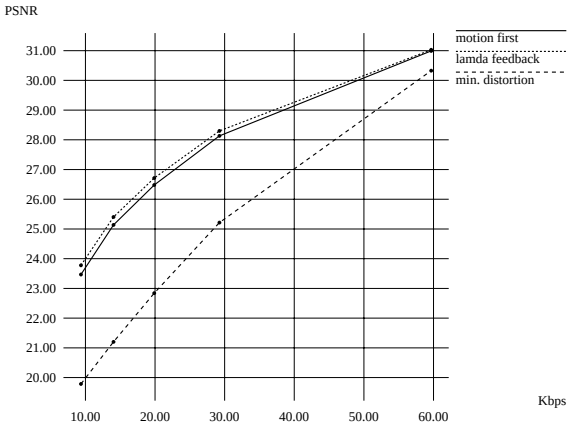

Figure 5: Rate-distortion curves for `foreman` sequence.

than one fourth of the bits were used in the lambda feedback case than in the distortion minimizing case for coding motion displacements, while producing a more reasonable motion field and better quality decoded data.

Figure 5 shows rate-distortion curves for the QCIF `foreman`[2] sequence compressed at five rates, from ten through sixty kilobits per second. The lambda feedback method gives a significant gain over the distortion minimization method, ranging from 3.5dB at 10 Kbps to 0.67dB at 60 Kbps.[3] This range corresponds to the decreasing percentage of the motion displacements in the bit stream as the rate increases.

2.2. Motion estimation before residual coding

The simulations above were done using lambda feedback on a macroblock by macroblock basis; that is to say that the motion search for a macroblock was done after the motion estimation and residual coding of macroblocks in the causal context were finished. This was done because the motion displacement median predictor depends upon the decisions made in the causal context.

An important question for system designers is whether or not the motion searches for an image of a sequence can be done before residual coding. If a system is to use lambda feedback in the motion search, must the motion searches be done on a block by block basis, or can they be done all at once, before residual coding of a group of blocks?

To answer this question, we must search for ways to estimate the number of bits that would be required to encode a particular displacement. Estimating the bits for the first block is simple, because the predictor is zero. One simple method for subsequent blocks is to assume that the previously estimated motion displacements are actually used in the bitstream, and calculate the displacement predictor accordingly. Note that according to the H.263 method, blocks which aren't coded with motion have their motion displacements set to zero when used in the displacement predictor calculation. This is the reason that the method outlined above is merely an estimation of the number of bits that a particular displacement will use, rather than the true value.

Because the buffer fullness doesn't change during the motion search for a color plane, the current buffer fullness is used to calculate a fixed lambda for all the motion searches for a color plane. Simulations with lambda feedback and this bit estimation scheme show that it does work well. The trace labeled "motion first" in Figure 5 shows performance when all the motion searches are done before residual coding of a color plane. There is a maximum of 0.3dB of PSNR performance decrease from the full lambda feedback case.

3. LOCALLY OPTIMAL MODE SELECTION

For inter-frame coding, as done with the two motion modes as described above, there are several methods of making the decisions in the motion estimation and the residual coding. The fully optimal method is to code the residual at each possible motion displacement, and choose the one which best minimizes our cost equation. Because of the immense number of possible motion offsets, this is clearly impractical. A simpler method is to VQ code a subset of the possible residuals, thus simplifying the problem. Note that our aim here is not to provide a frame-optimal mode decision as other research has focused on [11].

The four possible modes for coding of a macroblock require three decisions. Before any decision is made, the motion estimation is done. The first decision (which assumes motion is chosen) is between coding of four 8x8 residuals or one 16x16 residual. For lambda feedback, this is a rate-distortion decision, otherwise the decision is made with distortion and a bias which prefers 16x16 motion.

The second is the decision between motion residual coding and intra coding. This decision is based on the energies of the input block and motion residual block. A decision involving rate cannot be made, because there is no information available yet as to how many bits each option would use for VQ coding. After this decision is made, the motion residual or input block, as determined by the decision, is vector quantized. The third decision is between the quantized block, and the "no-update" option. This decision is always made on a rate-distortion basis.

The question arises after reviewing these decisions: "How much gain could be obtained by VQ coding both the motion residual *and* the input block" *before* the decision is made? Because we would be able to use rate and distortion from the quantization in our decision, we would expect some gain. We could make a locally optimal decision between the two modes.

Other options are also available. The decision between 8x8 motion, and 16x16 motion is made without the information of how these residuals would be quantized. We could quantize these two residuals, and use the resulting distortion and rate information in our decision. Another possible option for quantization is the residual at the location determined by the motion displacement median predictor.

We ran simulations using all four of these options. Thus there are four blocks that were vector quantized for every input block. These are: the input block, the residual using four motion vectors on 8x8 blocks, the residual from using a single motion vector on a 16x16 block, and the motion residual at the predicted position. A locally optimal rate-distortion decision was made between the

[2]The `foreman` sequence has high amounts of motion. Sequences with less motion show smaller gains.

[3]In general, the H.263 TMN 3.0 codec from UBC, with default options, seems to produce better PSNR than our coder with lambda feedback in the motion search. However, the visual quality of our coder with lambda feedback is better than that of the H.263 results.

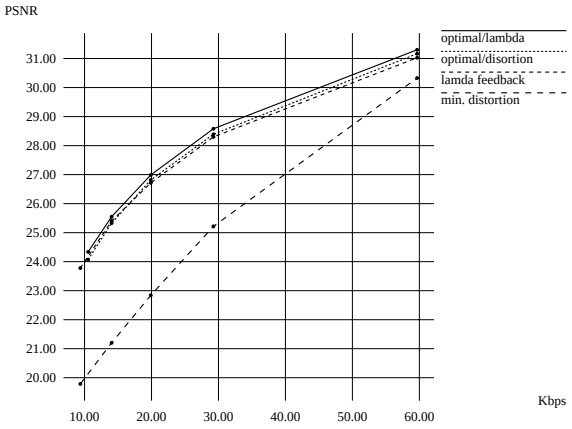

Figure 6: RD curves for optimal mode selection with the `foreman` sequence.

options. In other words, the mode which minimizes the cost is chosen:

$$D_m + \lambda(R_m^{motion} + R_m^{VQ} + R_m^{mode}). \quad (3)$$

Note that we are using several stages of decisions. First the motion estimation is done, which (for the lambda feedback case) uses a R-D cost (see equation (2)) to determine the best motion offset. Then VQ searches are done on the four vectors described above. Finally, a decision is made as to which mode to use using equation (3). Note that R_m^{mode} is the bits used for the Huffman-coded mode m, and the values R_m^{VQ} and R_m^{motion} are as used in equations (1) and (2), respectively.

Figure 6 shows the results of these simulations. The trace labeled "optimal/distortion" used the optimal mode decisions, but a minimum distortion decision was made in the motion search. Note that it gives a significant gain over the case without optimal mode decisions , labeled "min. distortion." The trace labeled "optimal/lambda" used the optimal mode decision, and lambda feedback in the motion search. It is surprising that this performs only slightly better than the lambda feedback without R_m^{VQ} or R_m^{mode} used in the cost equation. The fact that the two methods give about the same results seems to indicate that using these bit numbers (R_m^{VQ} and R_m^{mode}) in the cost equation does not bring us much closer to the globally optimal solution. It also may indicate that the 8x8 vs. 16x16 motion decision that is made in both cases is a very significant decision, and making this decision with a R-D cost of any sort gives significant gain. As well as no significant performance gains, the extra computations that are required for this method of mode selection also make it less attractive for implementation.

4. SUMMARY

We presented the use of a Lagrange multiplier as a function of buffer fullness to make a trade off between rate and distortion. We described how lambda feedback could be used in a motion estimation scheme. Use of this method provided up to 3.5 dB of improvement over the distortion minimizing search. This gain was for a high motion sequence at a low bit-rate; smaller improvements were observed for higher data rates, and low motion sequences. We also used a bit estimation method with the lambda feedback, to allow all of the motion searches to be done before residual coding of a frame. This estimation method worked well, requiring only a 0.3 dB decrease from simulations using the normal lambda feedback in the motion search.

We described a locally optimal mode selection scheme. Though this also gave a significant improvement over the nominal when used with the distortion minimizing motion search, the use of lambda feedback in the motion search nearly obviates the benefit of the new mode selection scheme. Because of its high computation cost, simplification of this method would be required to make it of use in a video compression system.

5. REFERENCES

[1] *ITU–T Recommendation H.263 "Video coding for low bitrate communication"*, June 1996.

[2] T. D. Lookabaugh and R. M. Gray, "High-resolution quantization theory and the vector quantizer advantage," *IEEE Transactions on Information Theory*, vol. 35, pp. 1020–1033, September 1989.

[3] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Dordrecht, Netherlands: Kluwer Academic Publishers, 1992.

[4] V. Swaminathan, "Residual vector quantization," Master's thesis, Utah State University, Logan, UT, 1996.

[5] K. M. Liang, C. M. Huang, and R. W. Harris, "Comparison between adaptive search and bit allocation algorithms for image compression using vector quantization," *IEEE Transactions on Image Processing*, vol. 4, pp. 1020–1023, July 1995.

[6] A. Ortega, K. Ramchandran, and M. Vetterli, "Optimal trellis-based buffered compression and fast approximations," *IEEE Transactions on Image Processing*, vol. 3, pp. 26–40, January 1994.

[7] C. B. Peel, "Buffer-constrained bit allocation," Master's thesis, Utah State University, 1997.

[8] J. Choi and D. Park, "A stable feedback control of the buffer state using the controlled Lagrange multiplier method," *IEEE Transactions on Image Processing*, vol. 3, pp. 546–558, September 1994.

[9] *ITU–T Study Group 16, Video Expert Group, Document Q15–A–59, "Video codec test model, near–term, version 8 (TMN8), release 0*, June 1997.

[10] M. C. Chen and J. Alan N. Willson, "Rate–distortion optimal motion estimation algorithms for motion–compensated transform video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, pp. 147–158, April 1998.

[11] T. Wiegand, M. Lightstone, D. Mukherjee, T. G. Campbell, and S. K. Mitra, "Rate-distortion optimized mode selection for very low bit rate video coding and the emerging H.263 standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 6, pp. 182–190, April 1996.