PREDICTION AND SEARCH TECHNIQUES FOR RD-OPTIMIZED MOTION ESTIMATION IN A VERY LOW BIT RATE VIDEO CODING FRAMEWORK*

Yuen-Wen Lee[1] *Faouzi Kossentini*[1] *Mark J. T. Smith*[2] *Rabab Ward*[1]

[1]Department of EE, University of British Columbia, Vancouver BC V6T 1Z4, Canada
[2]School of ECE, Georgia Institute of Technology, Atlanta, GA 30332-0250

ABSTRACT

Prediction and search techniques are introduced for efficient rate-distortion optimized motion estimation in a very low bit rate video coding framework. For prediction, three types of predictors are considered: mean, weighted mean, and median. Prediction allows us to constrain the motion vector search to a small diamond-shaped area whose center is the predicted motion vector. The size of the search area is further constrained by employing a probabilistic model. We evaluate two models, both of which permit the contraction or the expansion of the search area as a function of the local statistics of the motion flow. The proposed techniques are analyzed in the context of a very low bit rate DCT-based video coding framework, where a rate-distortion criterion is used for motion estimation as well as for 8×8 block coding mode selection. A particular resulting very low bit rate video coder is shown experimentally to outperform the H.263 TMN5 simulation model in terms of encoding speed and compression performance, simultaneously.

1. INTRODUCTION

Motion estimation is usually used to exploit temporal redundancies in very low bit rate video coding systems [1, 2]. Among the many motion estimation algorithms [1] that have been developed, those that are based on the conventional block matching algorithm (BMA) stand as the most popular and the simplest in concept, design, and implementation. The most notable example is the two-step BMA-based algorithm, which is adopted in many H.263-based video coders such as Telenor's TMN5 simulation model. The first step is an integer-pel accuracy full-search BMA (FS-BMA). The second step improves estimation accuracy by producing $\frac{1}{2}$-pel motion vector estimates.

There are many problems associated with the above two-step motion estimation algorithm. Besides its high computational complexity, the FS-BMA performs poorly during non-translational motion activities. This, coupled with the FS-BMA's sensitivity to video input noise, produces a non-smooth motion field that costs many precious bits at very low bit rates. Moreover, producing motion vector estimates with $\frac{1}{2}$-pel accuracy increases the complexity and the bit rate while normally providing a small performance advantage.

In this paper, we present a predictive rate-distortion (RD) optimized motion estimation algorithm, where several prediction and search methods are analyzed and compared. The work is presented in the context of a simple DCT-based video coding framework, where 8×8 blocks are used during both motion estimation/coding and DCT residual coding.

Predictive motion estimation has recently become an important research area [3, 4, 5, 6, 7]. We here study the complexity and performance of linear prediction (mean and weighted mean) and non-linear prediction (median) when applied to motion estimation/coding in the context of very low bit rate video coding. We also introduce two probabilistic models that allow the expansion or contraction of the predicted search area based on the local statistics of the motion flow.

To reduce the motion bit rate, which occupies as much as 50% of the total bit rate, several RD-constrained [8, 9, 10, 11] motion estimation algorithms have recently been introduced. In this paper, we build on our earlier work [8, 9], where we use a Lagrangian-based criterion both to locate the best motion vector and to alternate between three modes of operation: motion-only coding, motion-compensated predictive coding, and intra coding.

The proposed motion estimation algorithm is computationally efficient, yet its estimation performance is comparable to that of the FS-BMA. Experimental results show that a resulting very low bit rate video coder outperforms Telenor's TMN5 simulation model in terms of both computational complexity and compression performance. Our video coder also has the additional advantage that quality, bit rate, and complexity are easily controllable. Next, we present our motion estimation algorithm. This is followed with a description of the overall video coding framework. Section 4 presents our experimental results.

2. PROPOSED MOTION ESTIMATION

Suppose that the video frame to be predicted is partitioned into 8×8 blocks. For each block, a vector $\mathbf{d} = (x, y) \in \mathcal{S}$, where $\mathcal{S}$ is the set of all possible vectors in the search area, is sought that minimizes the Lagrangian $J_\lambda^M(\mathbf{d}) = \sum_{\mathbf{r} \in \mathcal{W}} \left(I(\mathbf{r}, n) - I(\mathbf{r} + \mathbf{d}, n-1)\right)^2 + \lambda \; R^M(\mathbf{d})$, where $\mathbf{r}$ is the spatial index of the image pixels, n is the time index, $I(\mathbf{r}, n)$ is the image intensity of the candidate block in the current frame, $I(\mathbf{r} + \mathbf{d}, n-1)$ is the image intensity of the matching block in the previous frame, $\mathcal{W}$ is the size of the matching window, and $R^M(\mathbf{d})$ is the motion vector bit rate. Minimizing $J_\lambda^M(\mathbf{d})$ is guaranteed only by considering all the possible candidate motion vectors in the search area, which involves a large number of search operations. The cost of each search operation is reduced in this work by using the partial Lagrangian computation technique, which is a generalization of the well-known partial distortion computation technique. Such a technique reduces the number of computations by as much as 80%. Unfortunately, the com-

*THIS WORK WAS SUPPORTED IN PART BY NSERC UNDER GRANT # OGP-0187668 AND ALSO BY NASA.

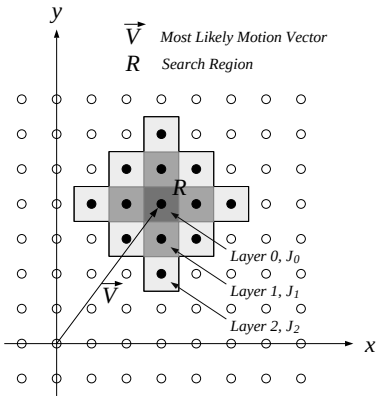

Figure 1. A diamond-shaped vector search area.

putational load can still be very high. This suggests that the size of the search area be substantially reduced. The fact that low resolution video sequences generally exhibit limited motion activity allows us to limit the search area to, for example, a small square of size ± 8. Even then, 289 search operations must still be performed. Therefore, many techniques such as three-step search and hierarchical search have been suggested [1], reducing substantially the set of possible motion vectors. The cost, however, is a significant decrease in estimation performance.

2.1. Prediction

By accurately predicting the location of the best motion vector candidate, one can limit the searching to a relatively small area in the neighborhood of the predicted motion vector, while still likely locating the "optimal" motion vector. This is indeed possible, thanks to the large amount of motion field redundancies within the same frame as well as between consecutive frames. Figure 1 shows an example of a search area whose center is the most likely integer motion vector $\mathbf{v} = (x_i, y_i)$ given a prediction model. The components of $\mathbf{v}$ are the closest integers to the corresponding components of the *real* predicted motion vector. Next, we describe the three different predictors (mean, weighted mean, and median) considered in this work. The design of the linear predictors is simplified by computing correlation values of motion vectors within a sufficiently large three-dimensional region of support (ROS) with the current motion vector. Such a region includes previously coded motion vectors representing blocks that are close spatially and/or temporally. Figure 2 shows the average correlations (c_x, c_y) of the x and y components of the ROS motion vectors with the current one. Note that these values decrease rapidly as we go away from the current block along the spatio-temporal axis, and that spatial dependencies are stronger than temporal ones.

The mean predicted motion vector is given by $\tilde{\mathbf{v}} = \frac{1}{K}\sum_{k=1}^{K} \mathbf{v}_k$. Based on Figure 2, only motion vector values representing blocks A, B, and C are averaged to produce the mean prediction. More accurate prediction is obtained by the weighted mean, given by

$$\tilde{\mathbf{v}} = \sum_{k=1}^{K} \alpha_k \mathbf{v}_k,$$

where the α_k's are proportional to the correlation values shown in Figure 2. In this work, the α_k's are computed

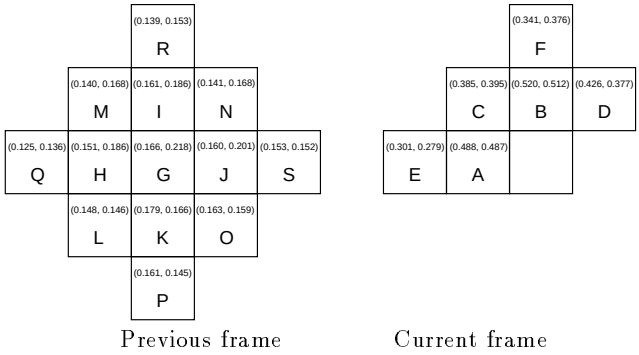

Figure 2. Average correlation values (c_x, c_y) of the ROS motion vectors with the current one.

off-line following conventional linear prediction techniques. Based on Figure 2, only the motion vectors representing the blocks A, B, C, D, and F are significantly correlated with the current one. Thus, a fifth order predictor is used.

Like those of the mean and weighted mean vectors, the two components of the median vector are computed independently in the same way. Two different median vectors are computed: one based on the H.263 3-block ROS (blocks A, B, and D) and another based on a 5-block ROS (blocks A, B, C, D, and G). As will be shown later, the 3-block median is better in terms of prediction performance. However, the 5-block median potentially provides more robustness to channel errors.

2.2. Search Area Size

The size of the diamond-shaped search area can be expressed in terms of layers, as shown in Figure 1. A conceptually simple searching technique is to search an experimentally pre-determined fixed number of layers. However, this technique can be inefficient, as if the prediction is accurate, the motion vector located at the center is likely the best candidate. We next introduce an alternative technique, where we employ a probabilistic model that places a soft constraint on the size of the diamond-shaped search area. First, let us assume that the layers $0, 1, 2, \ldots$ are searched sequentially in the same order as we go away from the center of the search area. Moreover, let $J_0, J_1, \ldots, J_n, \ldots$ be the minimum Lagrangian values associated with the layers $0, 1, \ldots, n, \ldots$, respectively. The two proposed probabilistic models are based on the following two hypothesis:

- Hypothesis I: If $J_n > J_{n-1}$, then it is unlikely that we will find a better motion vector by continuing the search outward. Thus, searching more layers is not necessary. This hypothesis is violated when the Lagrangian surface is not convex.

- Hypothesis II: Only if $J_{n-1} < J_n < J_{n+1}$, will we be confident that searching more layers is wasteful. This almost guarantees optimality, but at the expense of a much larger computational load.

Either model can be better than the other in terms of complexity-performance tradeoffs, depending on the particular video sequence being coded. As expected, both models would fail in areas where many non-motion changes occur.

3. THE VIDEO CODING FRAMEWORK

The proposed motion estimation algorithm is studied, analyzed, and tested within a simple RD-constrained DCT-based framework, where the first video frame is intra coded, and the following frames are forward inter coded. Inter coding of each 8×8 block involves motion-only coding, motion-compensated predictive coding, or intra coding.

After determining and coding[1] the motion vector $\mathbf{d}^*$ leading to the minimum value of the Lagrangian $J_\lambda^M(\mathbf{d})$, the corresponding prediction error block is DCT coded. The quantizer $q_R \in Q_R$ is chosen to minimize the Lagrangian

$$J_\lambda^R(q_R) = D_{dct}(q_R) + \lambda\, R_{dct}(q_R),$$

where $R_{dct}(q_R)$ and $D_{dct}(q_R)$ are the average bit rates and distortions, respectively, associated with quantizer $q_R \in Q_R$. Moreover, the current original 8×8 block is intra coded by minimizing the Lagrangian

$$J_\lambda^I(q_I) = D_I(q_I) + \lambda R_I(q_I),$$

where $R_I(q_I)$ and $D_I(q_I)$ are similarly the bit rates and distortions, respectively, associated with quantizer $q_I \in Q_I$. Finally, let the quantity R_m be the rate associated with specifying the mode m of operation. By incorporating such information, and other types of side information (e.g., quantizer number), the three Lagrangian values are computed, and the mode of operation leading to the smallest value should be selected for the current block.

Unfortunately, achieving the best rate-distortion performance can only be guaranteed by comparing the three Lagrangian values, which typically requires a relatively large number of computations. One computation-reduction method involves simplifying the DCT intra and residual coding procedures. A better method involves avoiding altogether the DCT coding process once the Lagrangian $J^M(\lambda)$ is relatively small, which can reduce the video coding process to mostly motion vector estimation and coding.

Finally, an important problem is how to determine a value for the Lagrangian parameter λ. In this work, λ is updated recursively as described in [12] to meet a bit rate and/or a quality constraint. For example, consider video coding in a fixed-rate communication system that is governed by $s(t+1) = s(t) + R(t) - B$, where $s(t)$ is the size of the buffer at time t, $R(t)$ is the variable output bit rate of the encoder, and B is the fixed output bit rate of the buffer. Assuming S_{max} is the maximum size of the physical buffer, we would like to maintain a buffer size equal to $s^* = \frac{S_{max}}{2}$. This can be closely achieved by recursively computing $\lambda(t)$ using the formula [12]

$$\lambda(t) = \lambda(t-1)\ \frac{s(t)}{s^*}.$$

Although simple, this method is found experimentally to perform quite well. In fact, using a buffer of size 10 kilobits, overflow and underflow problems were never encountered during our coding simulations.

4. EXPERIMENTAL RESULTS

The target bit rates for our experiments are in the range between 4 and 10 kbps for color sequences. The MISS AMERICA and CAR PHONE sequences in QCIF format at 10 frames per second are selected for testing. Only integer-pel accuracy motion vector estimates are obtained, and B-frames are not used.

[1] Each motion vector offset (with respect to the predicted motion vector) is coded using a Huffman-like VLC table.

PSNR	MEAN	WM	MED-A	MED-B
35.9	0.566	0.559	0.561	0.582
36.8	0.620	0.602	0.612	0.628
37.8	0.689	0.670	0.659	0.679
31.2	1.686	1.598	1.553	1.602
33.0	2.008	1.930	1.860	1.928
34.9	2.318	2.190	2.109	2.194

Table 1. Average entropy for MISS AMERICA (top three rows) and CAR PHONE (bottom three rows) of the motion vector x components in the search area centered at the mean, weighted mean (wm), 3-block median (med-a), and 5-block median (med-b).

Table 1 shows the average entropy for MISS AMERICA and CAR PHONE of the motion vector x component in the search area centered at the mean, weighted mean, 3-block median, and 5-block median. Notice that, as a predictor, the weighted mean consistently outperforms the uniform mean. Moreover, the 3-block median outperforms the linear predictors, especially for the more active sequence CAR PHONE. With the exception of its potentially higher robustness to channel errors, the 5-block median does not seem to be a good choice. Finally, note that the entropy does not depend significantly on the type of prediction being used. Taking prediction performance, complexity, and robustness into consideration, the 3-block median stands on top.

Figure 3 suggests that both of the search models (represented by Hypothesis I and Hypothesis II) discussed above lead to more than one order of magnitude reduction in number of computations, while sacrificing only an average of 0.15 dB loss in PSNR performance. Even for the relatively active video sequence CAR PHONE, the reduction in number of computations can be as large as $100:1$.

Finally, Figure 4 shows a comparison in terms of average PSNR between a resulting video coder and Telenor's TMN5 simulation model (using all advanced options) for the Y component of 150 frames in the bit rate range of interest. Clearly, our coder performs significantly better than the H.263 TMN5, especially at the very low bit rates. Moreover, note that the new fast searching technique results in a small loss in PSNR performance. Similar experiments using CAR PHONE reveals that the our coder performs consistently better than Telenor's H.263 TMN5 simulation model, yet it is also more computationally efficient.

REFERENCES

[1] V. Bhaskaran and K. Konstantinides, *Image and Video Compression Standards: Algorithms and Architecture*. Boston: Kluwer Academic Publishers, 1995.

[2] K.-H. Tzou, H. G. Musmann, and K. Aizawa, "Special Issue on Very Low Bit Rate Video Coding," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 4, pp. 213–367, June 1994.

[3] C. Hsieh, P. Lu, and J. Shyn, "Motion estimation using interblock correlation," in *Proc. IEEE Int. Sym. Circuits and Systems*, vol. 2, pp. 995–998, 1990.

[4] J. Chalidabhongse, S. Kim, and C. Kuo, "Fast multiresolution motion vector estimation for video coding by using spatial correlation," in *SPIE Proc. Vi-*

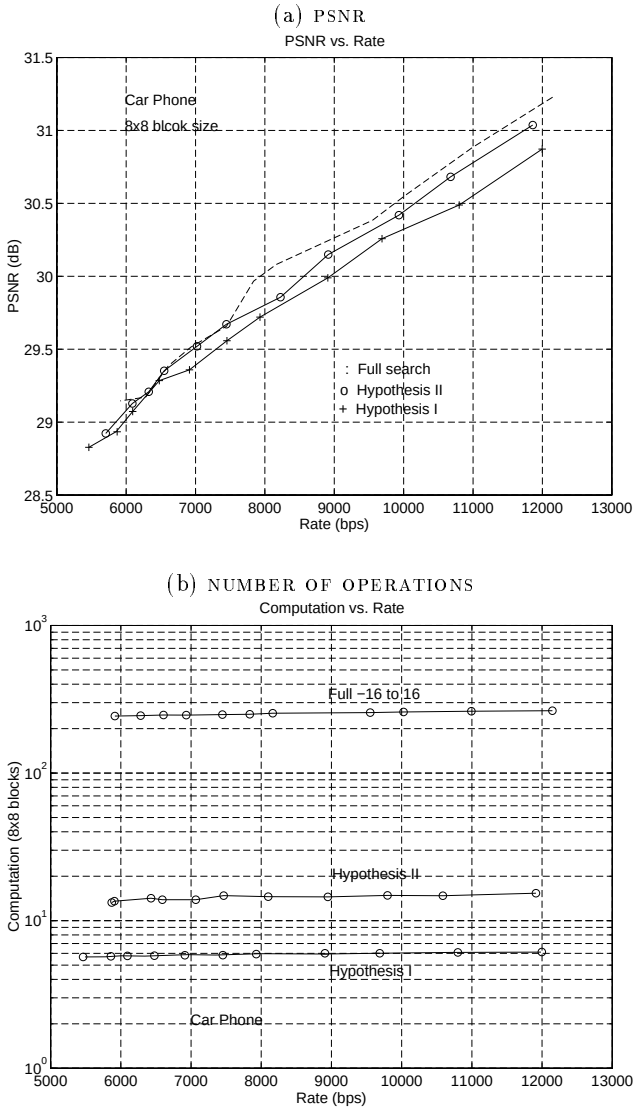

Figure 3. Performance of the probabilistic models in terms of PSNR and number of computations relative to the FS-BMA: CAR PHONE.

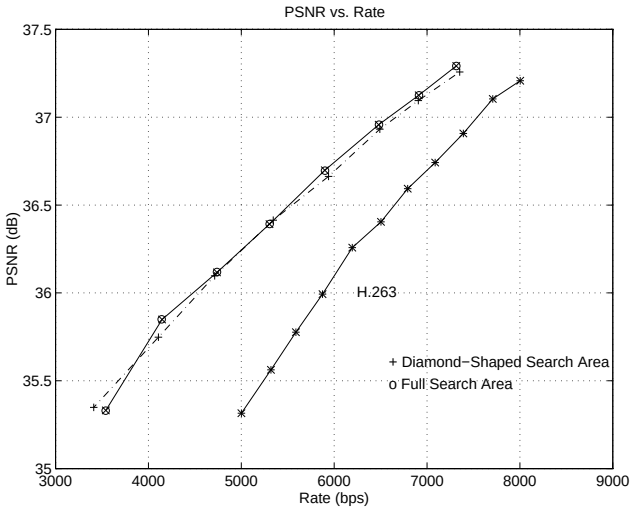

Figure 4. Performance comparison between the proposed video coder and Telenor's TMN5 (using all advanced options) for the test sequence MISS AMERICA **at very low bit rates.**

sual Communications and Image Processing, vol. 2464, pp. 76–87, 1995.

[5] R. Arminato, R. Schafer, F. Kitson, and V. Bhaskaran, "Linear predictive coding of motion vectors," in *Proceedings of the IS&T/SPIE EI'96*, Jan. 1996.

[6] Y. Lee, F. Kossentini, R. Ward, and M. Smith, "Towards MPEG4: An improved H.263-based video coder," *Accepted for publication in the MPEG4 Special Issue on Image Communication*, Aug. 1996.

[7] Y. W. Lee, R. K. Ward, F. Kossentini, and M. J. T. Smith, "Very low rate DCT-based video coding using dynamic VQ," in *ICIP96*, (Lauzanne, Switzerland), Sept. 1996.

[8] W. Chung, F. Kossentini, and M. Smith, "A new approach to scalable video coding," in *IEEE Data Compression Conference*, (Snowbird, UT, USA), pp. 381–390, Mar. 1995.

[9] W. Chung, F. Kossentini, and M. Smith, "An efficient motion estimation technique based on a rate-distortion criterion," in *ICASSP96*, (Atlanta, GA), May 1996.

[10] T. Wiegand, M. Lightstone, D. Mukherjee, T. Campbell, and S. Mitra, "Rate-Distortion Optimized Mode Selection for Very Low Bit Rate Video Coding and the Emerging H.263 Standard," *IEEE Trans. on Circuits and Systems for Video Technology*, pp. 182–190, Apr. 1996.

[11] D. Hoang, P. Long, and J. Vitter, "Efficient cost measures for motion compensation at low bit rates," in *IEEE Data Compression Conference*, (Snowbird, UT, USA), pp. 102–111, Apr. 1996.

[12] Y. Lee, F. Kossentini, M. Smith, and R. Ward, "Predictive RD-constrained motion estimation for very low bit rate video coding," *Submitted to the Special Issue of the IEEE Transactions on Selected Areas in Communications*, Aug. 1996.