# A MODIFIED RATE-DISTORTION OPTIMISATION STRATEGY FOR HYBRID WAVELET VIDEO CODING

*Thomas Davies*

BBC Research and Development

## ABSTRACT

Rate-Distortion Optimisation (RDO) techniques are applied in video coding in order to provide optimimum trade-offs between bit rate and quality, by minimising a metric combining measures of entropy and distortion. In hybrid wavelet video coding, inter as well as intra frames are wavelet transformed and encoded. Applying a single quantiser per subband does not allow for local adaption, and with the MSE metric for distortion, causes quality degradation as motion estimation failures are highly localised and poorly captured by the metric. This paper considers a fourth-power metric for measuring quantisation distortion within a subband, together with modifications to RDO motion estimation to reduce motion estimation failures. These techniques improve both subjective and objective performance without requiring multiple quantisers per subband. The approach has been applied to the open source Dirac video codec.

## 1. INTRODUCTION

Any transform (or none) may be used within a conventional hybrid video coding architecture. Block-based transforms remain the most popular choice, however, for a number of reasons. Partly this is because of their convenience for hardware design in permitting efficient pipe-lining (although this advantage is being eroded in modern software and hardware architectures, and by the requirements for complex motion compensation), and partly it is due to the existence of a very large body of research. However, an important factor is their capability for local adaptation which allows them to code motion compensated residual (MCR) data effectively.

A MCR frame will typically comprise a few isolated areas of high activity surrounded by regions of zero values where prediction has been highly accurate. Block-based coding allows blocks in these latter regions to be skipped whilst data is coded for blocks in the former regions. Furthermore, applying an RDO strategy as outlined in Section 2 to each block will tend to yield just such a result: the R-D curves for blocks in well-predicted areas being shallower than in poorly predicted areas, and hence resulting in higher quantisers being selected for the former. RDO therefore automatically allocates bit rate where it is subjectively needed, and the fine-grained localisation provides further savings.

By contrast, hybrid wavelet video codecs (as opposed to 3D wavelet codecs) such as the Dirac system [1] apply a transform to the whole MCR frame component. A subband may still have only a single quantiser for a whole subband spanning the entire MCR data. The slope of the R-D curve is necessarily shallow as the MCR data is mostly small or zero, and the resulting RDO quantisers will be too large for the poorly predicted areas. There will be an objective and subjective loss compared to a block-based codec.

Nevertheless, wavelet inter coding possesses two significant advantages. Firstly, decoupling prediction from coding allows improved prediction methods (such as [2]) to be adopted, which may adapt to the picture size and to the granularity of the motion. Secondly, in zerotree and similar coding techniques, arbitrarily-shaped areas of zero coefficients can be very efficiently coded. Areas of poor prediction in an inter frame are not naturally block-shaped, but dependent upon the shapes and motions of the objects within them, and wavelet coding techniques can exploit this.

This paper describes how RDO techniques have been refined in Dirac to improve subjective and objective performance.

## 2. RATE-DISTORTION OPTIMISATION

The RDO principle is one of optimisation subject to a constraint. Lagrangian multipliers allow distortion $D$ to be minimised with respect to a fixed rate constraint $R$ by instead minimising the *unconstrained* functional

$$J_\lambda = D + \lambda R$$

for a Lagrangian parameter $\lambda$ [3,4]. One may attain the target bit rate by varying $\lambda$, a higher value implying a lower rate and vice-versa.

### 2.1. Quantiser selection

In quantiser selection, $D=D(q)$ is a measure of the difference between quantised and unquantised values for a quantiser $q$, and $R=R(q)$ is a measure of the resulting rate for coding the quantised values, such as the entropy. the quantiser chosen is

$$q_{\min} = \arg\min(J_\lambda(q))$$

Applying this process across blocks or subbands will give the optimal combination of quantisers in rate-distortion terms, at the resulting point of the R-D graph. Note, however, that this will be in terms of the combined metrics for rate and distortion implied by those used in each block or subband.

### 2.2. Motion estimation

For RDO block matching based estimation, a Lagrangian parameter $\lambda_{\text{mot}}$ is used to control motion estimation via a matching metric

$$MJ_\lambda(\mathbf{v}) = BD(\mathbf{v}) + \lambda_{\text{mot}} MR(\mathbf{v})$$

combining a measure $BD(\mathbf{v})$ of the distortion implied by the choice of motion vector $\mathbf{v}$ for a block, and a measure $MR(\mathbf{v})$ of the motion vector bit rate implied. Typically, $BD$ is the sum of absolute differences (SAD) or the sum squared difference (SSD), and $MR(\mathbf{v})$ the $L^1$- or $L^2$-norm of $\mathbf{v}$-$\mathbf{v}^{\text{pred}}$, where $\mathbf{v}^{\text{pred}}$ is a prediction from neighbouring blocks.

This approach would be optimal if no residual coding were applied, the distortion and rate metrics were accurate, the combined measure was *jointly* minimised over all blocks, and if no inter frame was ever used as a reference.

In practice none of these conditions obtains, and the combined metric is effectively an ad hoc device for applying variable smoothing to motion vector fields. In addition, $\lambda_{\text{mot}}$ is typically derived as a monotone function of the quantisation Lagrangian parameter $\lambda_{\text{mot}}=f(\lambda)$, implying a fixed coupling between motion estimation and subsequent quantisation of the MCR data.

In quantiser selection, a rate metric can be assigned which indicates the rate of the whole data unit being quantised. In practical motion estimation, $MR(\mathbf{v})$ is a purely local measure of motion vector rate. Choosing a different, more expensive, motion vector may yield a lower overall value of $MJ_\lambda()$ averaged over all blocks.

It is worth noting that for wavelet coding, too, the distortion metric is also local with respect to the area that will subsequently be quantised and coded, whereas a block codec may match motion compensation and coding blocks.

## 3. DIRAC

Dirac is an open souce video codec, using a hybrid motion compensated architecture with Overlapped Block Motion Compensation (OBMC) and wavelet coding of both intra frame and MCR data. Full details of the Dirac system may be found at [5]. Dirac also has a specification [6], which is more functional than the current software, to which the software is being converged. One of the design principles of Dirac is to optimise subjective performance, if necessary at the expense of objective performance.

### 3.1. OBMC

OBMC is performed with basic blocks arranged into macroblocks consisting of a 4x4 array of blocks. Each macroblock may be split in one of three ways into prediction units consisting either of 16 individual blocks, or of an array of 4 mid-size blocks, termed sub-macrobocks, or of a single macroblock-sized block (Figure 1).

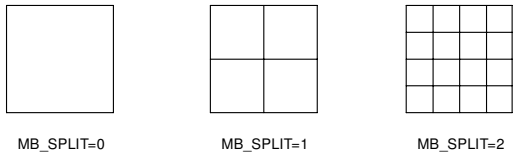


Figure 1. Macroblock splitting modes in Dirac

Each frame in Dirac may be predicted from up to two reference frames. Prediction modes can be varied by prediction unit, and there are four possibilities: Intra, Reference 1 only, Reference 2 only and Reference 1 and 2 (bi-directional prediction). Bi-directional prediction uses (½,½) filtering from the two references.

OBMC parameters may be changed frame-by-frame, but defaults exist based on frame sizes. The default for both streaming (CIF/SIF) and standard definition resolution is for $12\times12$ blocks which are overlapped at intervals of 8 pixels vertically and horizontally (the dimensions are scaled appropriately for chroma components of different resolutions). The OBMC overalapping function used is an integer approximation to the Raised-Cosine function.

### 3.2. Wavelet coding

Frame data, both for intra frames and MCRs, is wavelet transformed and coded subband by subband. Each subband may be divided into a number of rectangular code blocks, which may either be skipped (if all coefficients are zero) or quantised using the subband quantiser. The Dirac specification [6] also provides for the possibility of individual quantisers per code block, although this was not implemented in the tests. The ability to skip code blocks gives a degree of local adaptation.

Quantised coefficients are entropy coded using arithmetic coding, conditioned on neighbouring and parent values in the wavelet hierarchy.

### 3.3. Motion estimation

The Dirac example encoder software performs motion estimation in three distinct stages.

In the first stage, pixel accurate motion vectors are determined for each block and each reference frame by hierarchical block matching.

In the second stage, these pixel-accurate vectors are refined by searching sub-pixel values in the immediate neighbourhood.

In the final stage, mode decisions are made for each macroblock, determining the macroblock splitting level and the prediction mode used for each prediction unit. This last stage involves further block matching as block motion vectors are used as candidates for higher-level prediction units.

## 4. DIRAC RDO EXPERIMENTS

Two systems for RDO control of Dirac have been tested. The first system, System A, is a straightforward implementation of RDO using MSE metrics. The second system, System B, uses the more complex metrics employed in Dirac release 0.5.3 [5]. Both systems use the motion estimation scheme set out in Section 3.3.

### 4.1. System A RDO

In System A, the quantisation distortion metric $D(q)$ is the weighted mean square difference between quantised and original coefficients:

$$D(q) = \frac{1}{WN}\sum_{i,j}\left|c_{i,j} - q(c_{i,j})\right|^2$$

where $N$ is the number of coefficients in the subband and $W$ is a frequency-dependent weight derived for the subband from the ITU Rec. 959 function [7].

The rate metric $R(q)$ is the first-order entropy of the quantised coefficients.

For motion estimation, the distortion metric $MD$ is the SAD. The rate metric is the $L^1$ norm of the prediction error,

$$MR(\mathbf{v}) = \left\|\mathbf{v} - \mathbf{v}^{\text{pred}}\right\| = \left|v_x - v_x^{\text{pred}}\right| + \left|v_y - v_y^{\text{pred}}\right|$$

where $\mathbf{v}^{\text{pred}}$ is the median of vectors in blocks to the left, top-left and above the block currently being matched (the blocks being matched in conventional raster order, so that these blocks have already had motion vectors selected).

In the first, pixel-accurate, stage of motion compensation, $\lambda_{\text{mot}}$ is set to zero, the overlapping being sufficient to make the resulting field very smooth. For subsequent stages, $\lambda_{\text{mot}}$ is set by

$$\lambda_{\text{mot}} = \alpha\sqrt{\lambda}$$

for a fixed coefficient α.

### 4.2. System B RDO

In System B, $D(q)$ is defined using a fourth-power difference:

$$D(q) = \frac{1}{W}\left(\frac{1}{N}\sum_{i,j}\left|c_{i,j} - q(c_{i,j})\right|^4\right)^{1/2}$$

For a uniform quantisation differences, this metric is identical to System A's. In general, however, it greatly increases the contribution of localised areas of non-zero coefficients. The use of a higher power Minkowski summation for combining frequency error data is well-established in video quality metrics [8], and is a more accurate model of perceived errors than MSE, weighted or unweighted.

For motion estimation, $MD$ is SAD once again, but $MR$ is modified to be

$$MR(\mathbf{v}) = \min\left(\left\|\mathbf{v} - \mathbf{v}^{\text{pred}}\right\|, \left\|\mathbf{v}\right\|\right)$$

which means that transitions between foreground (near zero vectors) and background are less penalised.

#### *4.2.1. Transition detection*

For pixel-accurate matching, $\lambda_{\text{mot}}$ is set to zero. However, for sub-pixel matching and mode decisions $\lambda_{\text{mot}}$ may be non-uniform. The default value of $\lambda_{\text{mot}}$ is $\alpha\sqrt{\lambda}$ once again (although the coefficient α is different, resulting from the different quantisation distortion metric). However, after pixel-accurate motion estimation, blocks are partitioned into two classes, transition and non-transition blocks. Macroblocks are also partitioned, a transition macroblock being a macroblock containing a transition block. If a block is in a transition macroblock, $\lambda_{\text{mot}}$ is set to zero instead.

Transition blocks are determined by computing a smoothness field from the pixel-accurate motion vector field. For a block in position $(i,j)$, this is defined by:

$$S(i,j) = \max_{-1\le r\le 1,\,-1\le s\le 1}\left\|\mathbf{v}(i+r, j+s) - \mathbf{v}(i,j)\right\|$$

There is one field for each reference. The mean $\mu_S$ and standard deviation $\sigma_S$ are computed and a block at position $(i,j)$ is identified as a transition block if

$$S(i,j) > \mu_S + 3\sigma_S$$

For bi-directional prediction, the mean value of $\lambda_{\text{mot}}$ derived from the two transition fields is used.

This approach identifies motion transitions as exeptional relative to the inherent variation of the field itself, which is essential in order to reflect the wide variety of motion present in different real-world scenes, and in particular to control the motion vector bit rate in scenes with chaotic motion.

### 4.3. Test conditions

Five test pictures were used for the test. Their characteristics are summarised in Table 1.

These sequences represent a variety of motion type, speed and complexity. All sequences are progressive.

For SIF sequences, a 36 frame GOP was used, whereas for SD sequences a standard broadcast (12,3) GOP was used.

The target bit rates represent the average over the whole sequence: encoder control was purely by adjusting lambda parameters and no output buffering was employed.

| Sequence | Frame dimensions | Frame rate | Target bit rate |
|---|---|---|---|
| *football* | 352x240 | 15 | 512kb/s |
| *flower* | 352x240 | 15 | 512kb/s |
| *tennis* | 352x240 | 15 | 384kb/s |
| *mountain pass* | 720x576 | 25 | 1024kb/s |
| *tractor* | 720x576 | 25 | 1024 |

Table 1. Test picture characteristics

## 5. RESULTS

The resulting unweighted Peak Signal-to-Noise Ratio figures for the luminance component are shown in Table 2. Note that neither codec is designed to maximise PSNR, but employ frequency weighting to improve subjective performance.

| Sequence | System A PSNR(dB) | System B PSNR(dB) |
|---|---|---|
| *football* | 15.43 | 15.94 |
| *flower* | 14.81 | 15.04 |
| *tennis* | 20.67 | 21.32 |
| *mountain pass* | 24.70 | 26.94 |
| *tractor* | 28.74 | 29.32 |

Table 2. PSNR results for the two systems

Overall there is a minimum of 0.23dB advantage for System B over System A, rising to 2.24dB in the case of *mountain pass.*

Subjectively, the differences were very marked even where objective differences were small. System A showed motion compensation artefacts persisting uncorrected for considerable periods of time, causing double imaging and apparent disintegration of objects. Particular problems arose with revealed areas, which were not motion compensated and in System A were poorly corrected. *mountain pass* was especially challenging, consisting of a fast tracking helicopter shot with very rapidly moving background, with consequently a great deal of revealed area in each frame. Figure 2 illustrates the double imaging and disintegration produced by System A, which is visible in the areas to the left of the rock face.

## 6. CONCLUSIONS

RDO for hybrid wavelet codecs requires that local errors resulting from motion compensation failures are adequately captured in distortion metrics. It is possible to do this using a system of code blocks and multiple quantisers per subband, and the Dirac specification supports extensions to Dirac to allow this. However, good results can be obtained by choosing a perceptually-based fourth-power distortion metric that better reflects the greater sensitivity of the human visual system to large coefficient errors than MSE.

Figure 2. Artefacts of System A motion estimation

Rate-distortion control of motion estimation remains something of a black art, not least because of the inter-dependency of predicted frames which mean that motion compensation failures can persist over time. Rate-distortion optimisation can introduce motion compensation failures by over-smoothing fields across natural transitions. Transition-detection provides a simple and effective approach to mitigating these effects.

Future work will focus extend the quantisation distortion metric to incorporate multiple quantisers per subband, and the effect of alternative distortion metrics (such as a fourth-power difference) for motion estimation.

## 7. REFERENCES


[1] Borer, T., Davies, T., and Suraparaju, A., "Dirac video compression", *Proc. International Broadcasting Convention*, pp. Sept. 2005

[2] Servais, M., Vlachos, T., and Davies, T., "Motion-compensation using variable-size block matching with binary partition trees," ICIP 2005 [To appear]

[3] Gersho, A., and Shoham, Y., "Efficient bit allocation for an arbitrary set of quantizers," *IEEE Trans. Accoust., Speech, Signal Processing*, pp. 1445-1453, Sept. 1988

[4] Sullivan, G. J., and Weigand, T., "Rate-distortion optimization for video compression", *IEEE Signal Processing Magazine*, vol. 15, no.6, pp74-90, Nov. 1998.

[5] http://dirac.sourceforge.net/documentation.html

[6] http://dirac.sourceforge.net/specification.html

[7] "Experimental results relating picture quality to objective magnitude of impairment", ITU Rec. BT.959(2)

[8] Lai, Y.-K., "A Haar Wavelet Approach to Compressed Image Quality Measurement", J. Visual Comm. and Image Rep. **11**, pp 17–40, 2000