PERCEPTUALLY ADAPTIVE RATE-DISTORTION OPTIMIZATION FOR VARIABLE BLOCK SIZE MOTION ALIGNMENT IN 3D WAVELET CODING

Y. Sun[1], F. Pan[2] and A. A. Kassim[1]

[1] Department of Electrical and Computer Engineering
National University of Singapore, 10 Kent Ridge Crescent, Singapore 119260
[2] Institute for Infocomm Research
21 Heng Mui Keng Terrace, Singapore 119613

ABSTRACT

In this paper, a novel content adaptive rate-distortion optimization scheme has been proposed. The scheme could effectively distinguish texture region, edge region and flat region using directional field technique. Since the Human Vision System (HVS) perceives distortions more easily near edges and in flat regions, distortion reduction is more important in those regions than the bits it consumes to code the motion information. The adaptive rate-distortion optimization is carried out by adjusting the Lagrangian multiplier so that small values are assigned to edge and flat regions and large values to the random texture region. The proposed scheme has been tested in the scalable video coding (SVC) reference codec by Microsoft Research Asia (MSRA) [1]. Experimental results have shown that the accuracy of motion alignment in the visually important region is greatly improved in the temporal transform step of 3D wavelet coding and the scheme effectively preserves details in the most perceptually prominent regions for all bitstream layers with no loss in PSNR.

1. INTRODUCTION

The Human Vision System (HVS) is not able to acquire all the information that is presented to them with equal attention. It is assumed in vision research that human attention will always be drawn to certain pre-attentive visual features like orientation, curvature and motion [2]. Therefore, in regions containing these features, it would be easier to spot distortions. On the other hand, in the texture regions, distortions are less noticeable, and it is reasonable to emphasize on rate reduction and tolerate more distortions.

In most of the traditional video coders, temporal correlation is exploited by carrying out motion estimation based on minimal mean absolute error (MAE), which is generally not a good indication of the distortion as perceived by HVS. There have been several efforts in the past trying to include perceptual measures into video encoding. In [3], the focus was mainly on determination of proper quantization steps with sub-band *Just Noticeable Distortion* (JND). In recent H.264/MPEG-4 advanced video coding standard, highest coding efficiency is achieved by introducing the rate – distortion optimization (RDO) technique to give the best coding result by maximizing image quality and minimizing data rate at the same time. In [4], a new adaptive RDO scheme has been proposed which exploits motion and texture masking property to adjust the Lagrangian multiplier and achieves an overall bitrate reduction by allowing more distortion in the less noticeable background random texture region. The success of perceptual RDO scheme of this kind depends largely on the good estimation of visual features and thus a more accurate visual importance map is needed to be established.

The RDO technique is also an important component in wavelet-based scalable video coding (SVC) scheme that is currently under investigation by MPEG-21, Part 13 [5]. The SVC scheme requires an embedded bitstream to be formed from which bitstreams with different bitrate, resolution and frame rate could be extracted with reasonably good quality. In this paper, we propose an effective directional field based visual importance map which could successfully capture the features related to pre-attentive processing such as edges and curves. Based on the visual importance map, the regions with more pre-attentive features will get more distortion-reduction by assigning a smaller Lagrangian multiplier. Rate balance is achieved by assigning a relatively larger Lagrangian multiplier to the random texture region so that more distortion is allowed without noticeable visual degradation to the image. Since HVS is also sensitive to distortions in flat regions, we also assign a small Lagrangian multiplier to flat regions. Experimental results have shown that the new scheme could preserve most pre-attentive features and gives distortion reduction in the flat regions. There is virtually no loss in PSNR compared to the original MSRA codec.

The rest of the paper is organized as follows. Section 2 gives an overview of the motion alignment techniques under investigation in SVC. Section 3 discusses the techniques of getting visual importance map based on directional field estimations. Section 4 presents the content-based perceptually adaptive RDO scheme. Experimental results are summarized in Section 5, and conclusion is presented in Section 6.

2. LAYERED MOTION ALIGNMENT IN SVC

In SVC, scalable motion representation is necessary to achieve arbitrary rate-distortion optimized motion and texture at any bitrate range. Several ways of scalable motion coding have been proposed. In [6], motion information is represented in layers from coarse to fine resolution and accuracy. It consists of a base

layer and several enhancement layers. The base layer is generated using a relatively large Lagrangian multiplier λ, and the enhancement layers are generated by successively smaller λ. In the layered framework, the value of λ is fixed for the generation of one particular layer and thus the tradeoff between rate and distortion is fixed everywhere in that frame. Variable size block matching and selective motion vector coding is used within each layer of motion estimation [6]. The best matching result is selected from one of the seven block partition modes from which greater accuracy is obtained by a finer partition. With this scheme, it might be possible to have sub-optimal block-partition decisions. Coarser partitions might be selected for a detail region and an unnecessary fine partition might be selected for a random texture region because that λ is the same throughout the entire frame. If, for every macroblock *(MB),* λ can be made adaptive to importance of the local contents based on HVS, this kind of sub-optimal decisions could be avoided by assigning a small λ to detail regions to achieve a finer partition and assigning a large λ to texture regions and to stop the partition early. We have designed a mechanism to incorporate visual importance map in the motion coding process and the results are presented in subsequent sections.

3. DIRECTIONAL FIELD BASED VISUAL IMPORTANCE MAP ESTIMATION

There are many segmentation techniques, such as Edgeflow [7], Meanshift [8], and Bayesian estimations [9], which are capable of providing good segmentation results but are computational complexity intensive. In this paper, we propose an effective visual importance map based on directional field technique which is simple yet effective. Directional fields can be estimated by gradient-based methods [10]. In the gradient-based approach, the pixel gradient vectors in an image are described by the gradient vector $[G_x(x,y),\ G_y(x,y)]^T$, which can be simplified by the use of Sobel operators.

$$\begin{aligned} G_x(x,y) &= I(x-1,y+1)+2\times I(x,y+1)+I(x+1,y+1) \\ &\quad -I(x-1,y-1)-2\times I(x,y-1)-I(x+1,y-1) \\ G_y(x,y) &= I(x+1,y-1)+2\times I(x+1,y)+I(x+1,y+1) \\ &\quad -I(x-1,y-1)-2\times I(x-1,y)-I(x-1,y+1) \end{aligned} \quad (1)$$

where $I(x,y)$ represents the pixel intensity at location (x,y) in an image. The complex number representation of the gradient vectors is squared before averaging. After averaging, the gradient vectors have to be converted back to their single-angle representation. The squared vectors are found as,

$$(G_x + jG_y)^2 = G_x^2 - G_y^2 + j2G_xG_y \quad (2)$$

and the average squared gradient can be calculated by averaging in local neighborhood with a window size of W,

$$G_{xx} = \sum_W G_x^2,\ G_{yy} = \sum_W G_y^2,\ G_{xy} = \sum_W G_xG_y \quad (3)$$

With the use of above notations, the coherence (*Coh*) of the squared gradients can be expressed as below:

$$Coh = \frac{\sqrt{(G_{xx} - G_{yy})^2 + 4G_{xy}^2}}{G_{xx} + G_{yy}} \quad (4)$$

Fig. 1. 58th Frame of bus (top) and its importance map (bottom).

A coherence value of 1 refers to the extreme case where all squared gradient vectors are in the same direction. On the other hand, a coherence value of 0 indicates that the squared gradient vectors are equally distributed in all directions. The coherence value may vary between these two extremes. Therefore the coherence value of the directional field provides important information in classifying image into texture, edge and flat regions. A coherence of 1 implies that the neighborhood edges are consistently pointing in the same direction and this corresponds to regions with strong edges. A coherence of 0 implies that the neighborhood edges are scattered in all directions and there are no edges i.e., corresponding to a flat region. A coherence value near 0.5 implies that the neighborhoods are texture regions. An example is shown in Figure 1 where the white area denotes the edge region where the coherence value in the 4×4 neighborhood is greater than 0.9, and the gray area denotes random texture where the local coherence score is in between 0.6 and 0.9. It can be seen from the example that most of the important edges and textures are captured in this importance map.

4. ADAPTIVE RDO SCHEME

Since the motion estimation is performed on an *MB* by *MB* basis, the Lagrangian multiplier can also be adjusted at the *MB* level. Based on equations (1) to (4), we can compute the average coherence for an *MB*. A non-linear mapping is used to decide how much λ is to be adjusted based on the *MB*'s average coherence.

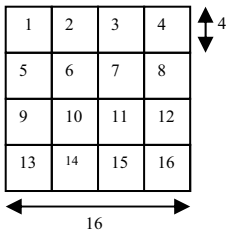

Fig. 2. An *MB* of size 16×16 is divided into 4×4 windows.

The current frame is divided into *MB*s of size 16×16 for motion estimation. To compute the coherence score of an *MB*, it is not suitable to directly set the window size to be the size of an *MB* as in Equation (3). The *MB* is too big to capture the local image characteristics accurately. To better estimate the image features, we divide the *MB* into small windows of size 4×4 and compute the coherence score for every 4×4 window in the *MB* using equations (1) to (4). A sample *MB* is shown in Figure 2. Every *MB* is divided into 16 4×4 windows. If the coherence of the *i*-th window in the *MB* is denoted as *Coh(i)*, the average coherence of the *MB* is obtained as in (5),

$$Coh_{MB} = \frac{1}{16}\sum_{i=1}^{16} Coh(i) \tag{5}$$

For the edges and flat regions with coherence 0 and 1, we would like to use a small λ in the RDO process, indicating that distortion reduction is more important than rate reduction. So we assign it to be 0.5λ, which is the lower bound of λ_{MB}. For other average coherence scores, λ_{MB} will be adjusted linearly. From Coh_{MB} = 0 to Coh_{MB} = 0.5, it represents a gradual change of image characteristics from a flat region to a random texture region and increasingly more distortion could be tolerated. At Coh_{MB} = 0.5, we select the maximum value for λ_{MB} to be 1.2λ. Our mapping function gives a linearly increasing λ_{MB} in that interval to give increasingly more rate reduction. From Coh_{MB} = 0.5 to Coh_{MB} = 1, the image characteristics changes from random texture to edges and less distortion could be tolerated. Our mapping function gives a linearly decreasing λ_{MB} to give increasingly more distortion reduction. The resultant mapping function is nonlinear and symmetrical in the line Coh_{MB} = 0.5 and is formulated in (6).

$$\lambda_{MB} = \begin{cases} (\alpha_1 \times Coh_{MB} + \beta_1)\lambda & \text{if } Coh_{MB} \le 0.5 \\ (\alpha_2 \times Coh_{MB} + \beta_2)\lambda & \text{otherwise} \end{cases} \tag{6}$$

where λ is predefined in the codec, $\alpha_1 = 1.4$, $\beta_1 = 0.5$ and $\alpha_2 = -1.4$, $\beta_2 = 1.9$.

5. EXPERIMENTAL RESULTS

The algorithm has been implemented into the MSRA reference software [11]. In our experiments, four levels of temporal decomposition have been performed during compression and several bitstreams are generated for a number of rate points as described in Table 1. The motion estimation used in this experiment is single layer for each of the temporal decomposition to better examine the effect of the proposed algorithm. Adaptation of the algorithm to the multiple layer frameworks for each level of temporal decomposition is straightforward. Coherence maps are generated using a 4×4 window for every target frame during motion search in the temporal decomposition. Since the resolution of the bitstream specified at high bitrate is CIF and at low bitrate is QCIF, the generation of motion information for the first two levels of the temporal decomposition is CIF and for the next two levels it is QCIF. λ is initialized with empirically determined value of 16 for the first two temporal levels and 32 for the next two levels. Five bitstream layers are generated for all the test sequences. The bitstream test points are listed in Table 1.

It has been observed that significant visual improvements were achieved in the reconstructed frames especially in the regions containing substantial amount of details. This is because all the important image features have been captured by the directional field based coherence map, and the value of λ has been adjusted accordingly in different image regions resulting in a better distortion reduction in the detailed region. As an example, in the frame number 12 of the foreman sequence shown in Figure 4, many details near the corner of the eye are preserved and less distortion is observed at the mouth. Similar effects can be seen in Figure 5 as well, the window of the bus is blurred in the reconstructed frame from the reference software, but it is well preserved in the new scheme.

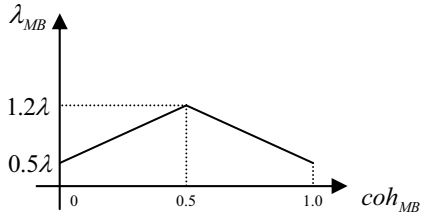

Fig. 3. Relationship of λ_{MB} against Coh_{MB}

The PSNR results are summarized in Table 1 for different sequences using the proposed algorithm. Figure 6 shows the PSNR comparison of the results from the reference software and the proposed algorithm. It has been observed that the performance of the proposed algorithm is comparable to the reference software in terms of PSNR. For all the sequences tested, the average PSNRs are almost the same. At some rate points, the proposed algorithm even outperforms the reference method. This is because the proposed scheme allows more local PSNR loss in random texture regions to same bit budget. The local loss is compensated by the gain in PSNR in the detail and flat regions. The suppression of distortion in the detail and flat regions lead to gain in perceptual quality. On the other hand, the increase in distortion in the random texture region is much less noticeable. This results in overall improvement of the visual quality.

6. CONCLUSION

A directional field based content adaptive rate-distortion optimization algorithm for scalable 3D wavelet video coding has been proposed in this paper. The algorithm is capable of distinguishing local characteristics of images in terms of edge, texture and flat region and it adaptively optimizes the R-D tradeoff at a *MB* level. The algorithm has a low complexity and could be easily implemented. Preliminary experiments have shown that the algorithm could increase the subjective quality of

the reconstructed frames in terms of detail preservation with no loss of PSNR at a wide range of rate points.

Fig. 4. Frame 12 of Foreman sequence at bitrate 256kbps, 30frame/s. Left: anchor image of the reference software; Right: proposed method.

Fig. 5. Frame 58 of bus sequence at bitrate 192kbps, 15 frame/s. Left: anchor image of the reference software; Right: proposed.

Table 1. PSNR results for different sequences

Sequence	Bitrates (kbit/s)	Spatial format	Frame Rate	PSNR (MSRA)	PSNR (proposed)
Foreman	32	QCIF	7.5	29.248	29.218
	48	QCIF	15.0	29.704	29.576
	96	CIF	15.0	30.886	30.877
	192	CIF	15.0	33.579	33.603
	256	CIF	30.0	34.329	34.333
Mobile	48	QCIF	7.5	22.791	22.791
	64	QCIF	15.0	23.267	23.234
	128	CIF	15.0	23.749	23.747
	256	CIF	15.0	26.908	26.912
	384	CIF	30.0	28.645	28.650
Bus	64	QCIF	7.5	26.002	25.986
	96	QCIF	15.0	26.295	26.289
	192	CIF	15.0	27.485	27.495
	384	CIF	15.0	30.466	30.478
	512	CIF	30.0	30.998	30.972
Football	128	QCIF	7.5	30.273	30.258
	192	QCIF	15.0	29.481	29.408
	384	CIF	15.0	31.074	31.116
	512	CIF	15.0	32.478	32.489
	1024	CIF	30.0	34.162	34.194

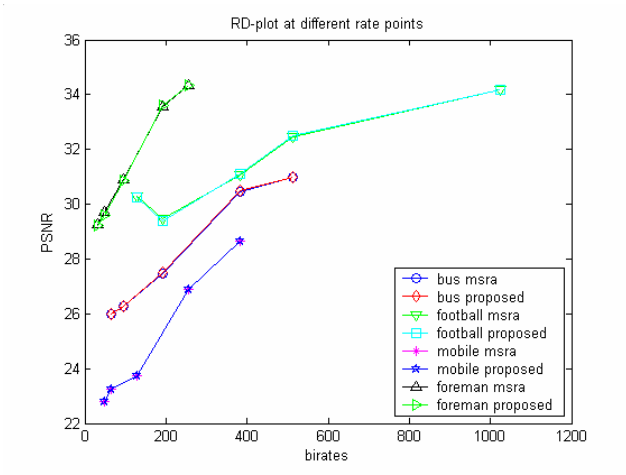

Fig. 6. RD-plot of different sequences.

7. REFERENCES

[1] J. Xu, et. al, "3D Sub-band Video Coding using Barbell lifting," ISO/IEC JTC/WG11 M10569, S05, March 2004.

[2] O. Le Meur, P. Le Callet, D. Barba, D. Thoreau, E. Francois, "From low level perception to high level perception, a coherent approach for visual attention modeling." *Proc. of SPIE-IS&T Electronic Imaging 2004*, SPIE Vol. 5292.

[3] C.-H. Chou and C.-W. Chen, "A perceptually optimized 3-d subband image codec for video communication over wireless channels," *IEEE Trans. Circuits Syst. Video Technol.* 6(2), pp. 143-156, 1996.

[4] Chun-Jen Tsai, Chih-Wei Tang, Ching-Ho Chen, and Ya-Hui Yu, "Adaptive Rate-distortion Optimization using perceptual Hints", *2004 IEEE International Conference on Multimedia and Expo (ICME'2004),* Taipei, Taiwan, June 27th – 30th, 2004.

[5] ISO/IEC JTC 1/SC 29/WG 11N6520, *Scalable Video Model 2.0*, July 2004, Redmond, WA, USA

[6] Ruiqin Xiong, Jizheng Xu, Feng Wu, Responses of CE1a in SVC: Scalable Motion. ISO/IEC JTC1/SC29/WG11 MPEG2004/M11128, Redmond, July 2004.

[7] Wei-Ying Ma and B. S. Manjunath, "EdgeFlow: A Techique for Boundary Detection and Image Segmentation", *IEEE Transactions on Image Processing*, vol. 9, No. 8 August 2000.

[8] Dorin Comaniciu and Peter Meer, "Mean Shift: A Robust Approach Toward Feature Space Analysis", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 24, No. 5, May 2002.

[9] Charles A Bouman and Michael Shapiro, "A Multi-scale Random Field Model for Bayesian Image Segmentation", *IEEE Transactions on Image Processing*, vol. 3, No.2, March 1994.

[10] A. M. Bazen and S. H. Gerez, "Systematic Methods for the Computation of the Directional Fields and Singular Points of Fingerprints," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 905-919. 2002.

[11] MSRA software, ftp://ftp.tnt.uni-hannover.de