RATE-DISTORTION CONSTRAINED QUADTREE SEGMENTATION FOR STEREOSCOPIC VIDEO CODING

Chia-Wen Lin
Computer and Communication Research Labs
Industrial Technology Research Institute
Hsinchu, Taiwan 310, ROC

Eryin Fei, and Yung-Chang Chen
Department of Electrical Engineering
National Tsing Hua University
Hsinchu, Taiwan 300, ROC

ABSTRACT

In this paper, a rate-distortion framework is proposed to define a jointly optimal displacement vector field estimation (DVFE) and quadtree segmentation technique for stereo video coding. This technique achieves maximum reconstructed image quality under the constraint of a target bit rate for the coding of the displacement vector field, quadtree segmentation information, and the residual signal. A fast hierarchical motion/disparity estimation scheme as well as a conditional optimization strategy is proposed to drastically reduce the large computation cost required in the R-D optimization process. The simulation results show that the proposed method can achieve more than 1 dB PSNR quality improvement over the H.263 TMN5 video codec at an acceptable extra computation cost.

1. INTRODUCTION

Rate-distortion optimization technique has recently been widely investigated for the applications in image and video coding, because it can achieve maximum reconstructed image quality under the constraint of a target bit rate [4-5]. The effect of rate-distortion optimization would be more obvious when dealing with low bit rate video coding, since the ratio of bits needed for encoding the displacement vector field and segmentation information will become relatively large in such cases. Therefore, it is desirable for low bit-rate video coding to reduce as much as possible the bit rate needed to transmit the displacement vector field and segmentation information, provided that this reduction does not produce intolerable distortion in the reconstructed image.

Quadtree segmentation has been widely adopted in variable block-size video coding to effectively reduce the number of transmitted motion vectors as well as maintain the motion integrity in large uniform regions which are often segmented into smaller blocks in fixed block-size video coding schemes. The gain of such kind of variable-block size schemes over the traditional fixed block-size schemes is especially obvious in the application of low-motion video (e.g., head-and-shoulder images in video phone applications).

Combing quadtree segmentation scheme with R-D optimization concept can further improve the coding performance as shown in [3]. The optimization of segmentation in rate-distortion sense is, however, very computationally intensive thereby making it impractical in real-time applications. In this paper, we develop a framework of R-D constrained quadtree segmentation for stereoscopic video coding. A fast hierarchical motion/disparity estimation scheme as well as a conditional optimization strategy is proposed to drastically reduce the computation load so as to meet real-time requirement.

2. R-D CONSTRAINED STEREOSCOPIC VIDEO CODING

Let $v_i \in V$ be the displacement vector corresponding to the block i of the image, where V is the set of all displacement vectors determined by the proposed search algorithm. The purpose of the R-D constrained video coding is to minimize the distortion D of the reconstructed image sequence, under the constraint of the target rate R_{target} for transmitting the displacement vector field, the segmentation information and the error image. This corresponds to the following constrained optimization problem:

$$\min_{v_i \in V} \sum_{i=1}^{N} D(v_i, s_i) \tag{1}$$

subject to

$$\sum_{i=1}^{N} R(v_i, s_i) \le R_{\text{target}} \tag{2}$$

where N is the total number of blocks in the image, $D(v_i,s_i)$ is the contribution of the jointly considered pair (v_i,s_i) to the overall distortion, and $R(v_i,s_i)$ is the contribution of (v_i,s_i) to the total rate. In the proposed R-D constrained coding method, the rate part is composed of three components: one is the bit rate for transmitting the displacement vector field, another one is the bit rate for sending the quadtree segmentation information, and the rest is for coding the error image. On the other hand, the distortion part is determined by means of Displaced Frame Difference (DFD). From the methodology shown in [4-5], the above problem can be transformed into an unconstrained optimization problem by adopting the Lagrange multiplier λ. Thus the solution $\{ (v_i^*(\lambda), s_i^*(\lambda)), i = 1, \ldots, N \}$ of the unconstrained minimization of the cost function $C(\lambda)$:

$$C(\lambda) = \sum_{i=1}^{N} C_i = D(\lambda) + \lambda R(\lambda)$$
$$= \sum_{i=1}^{N} D[v_i(\lambda), s_i(\lambda)] + \lambda \sum_{i=1}^{N} R[v_i(\lambda), s_i(\lambda)] \tag{3}$$

is also a solution of (1) if:

$$R_{\text{target}} = \sum_{i=1}^{N} R[v_i^*(\lambda), s_i^*(\lambda)] \tag{4}$$

It was shown in [4] that $D(\lambda)$ and $R(\lambda)$ are monotonic functions of the Lagrange multiplier λ, with values ranging from zero (highest rate, lowest distortion) to ∞ (lowest rate, highest distortion). A value of λ corresponds to a (R, D) operating point. Since the relationship between $D(\lambda)$ and $R(\lambda)$ is nearly one-to-one, all we have to do is to find an optimal Lagrange

0-7803-5471-0/99/$10.00©1999 IEEE

multiplier λ^* which makes $R(\lambda^*)$ close to R_{target}. The corresponding solution $\{ (v_i^*(\lambda), s_i^*(\lambda)), i = 1, \ldots, N \}$ constitutes the optimal displacement vector field under the target rate constraint. Each (R, D) point represents a jointly considered pair of DVFE and segmentation. The optimal DVFE and segmentation pair will lie on the surface of convex hull of all (R, D) points. Thus, the optimal Lagrange multiplier λ^* can be traced on the surface of the convex hull for a given rate or distortion constraint [3]. A fast method called convex hull bisection searching algorithm has been introduced in [3,5]. The algorithm can efficiently find the optimal Lagrange multiplier λ^* and the optimal pair of DVFE and segmentation on the surface of convex hull of all (R, D) points by decreasing the interval of possible λ in an iterative manner.

3. THE PROPOSED R-D CONSTRAINED QUADTREE SEGMENTATION

The proposed architecture to realize the aforementioned framework of R-D constrained stereoscopic video coding is depicted in Figure 1. The R-D constrained quadtree segmentation is performed only on the left-view image sequence. A hierarchical split-and-merge scheme based on motion and disparity information is used to segment an image into variable-size blocks with uniform motion or depth as described in Sec. 3.2. The R-D optimization process requires performing quadtree segmentation and motion estimation in an iterative fashion to find the best match. This process for finding the optimal (R, D) operational point is very computationally intensive, thus making it difficult to be directly adopted in real-time video communication. In order to reduce the computation load, some strategies are adopted to speed up the optimization process without introducing severe degradation. Firstly, a fast hierarchical estimation scheme similar to our previously proposed approach in [2] is adopted to drastically reduce the computation load in motion/disparity estimation. Furthermore, in our proposed method, the rate-distortion optimization is not applied to the whole image, instead it is only applied to the moving regions. This conditional optimization strategy can also save much computation time while still maintaining good video quality. The right-view image sequence is encoded by using a hybrid motion/disparity compensated coding scheme. Adopting the suppression theory, the spatial resolution is halved without significantly affecting the stereo perception of human visual system.

3.1 Hierarchical Motion/Disparity Estimation with Spatial Predictors

Block-wise structure is widely used for the currently existing coding techniques and for disparity estimation. For the block-based techniques of disparity estimation, it is important to choose a proper block size. Using large-size blocks will lead to inaccurate disparity estimation results. On the other hand, small block size decreases the reliability of disparity estimation. Using hierarchical block matching (HBM) method [1] with pyramid structure can solve this problem. In our proposed hierarchical estimation scheme, a constant block size of $b_x \times b_y$ is used in all levels of the pyramid. In this pyramid structure, one block at level N of the pyramid corresponds to 4 blocks at level N-1. That is,

$$v^j_{N-1}(i,j) = 2v_N(i/2, j/2) \quad (5)$$

$$v^j_{N-1}(i+b_x, j) = 2v_N(i/2, j/2) \quad (6)$$

$$v^j_{N-1}(i, j+b_y) = 2v_N(i/2, j/2) \quad (7)$$

$$v^j_{N-1}(i+b_x, j+b_y) = 2v_N(i/2, j/2) \quad (8)$$

Due to the epipolar constraint, the search range of disparity estimation in horizontal direction is much larger than in vertical direction. For lacking of epipolar constraint, the search range of motion estimation in horizontal direction is equal to vertical one. The search range for the proposed 2-level HBM for motion and disparity estimation is specified in table 1, where the coarse estimation with full-pixel precision is performed at level 1 and fine estimation with half-pixel precision is performed at level 0. Reconstruction of the total vector field is performed by using:

$$v_{total} = 2^{L-1} v_L - 1 + \sum_{P=0}^{L-2} 2^P dv_P \quad (9)$$

where L is the total level number of the pyramid and dv_P is the difference value between the estimated vector and the initial estimation vector.

By considering the spatial continuity of motion and disparity, we can use the adjacent estimated motion/disparity vectors as the predictors for the current motion/disparity vector in the estimation process. This is under the assumption of a large object with a smooth surface and motion shown in the stereoscopic image pair. In the proposed method, we utilize the three adjacent estimated vectors as the predictors to estimate the current vector as depicted in Figure 2.

The predictors are only utilized at the highest level of the pyramid. The one with the least MAD among the three predictors is taken as the initial estimation of the current vector with a relatively small search area. The main effect of predictors is significantly reducing the computation load in the estimation process while degrading the quality of the estimation in an acceptable range. In addition, the whole displacement vector field is becoming smoother and it brings benefits to the segmentation process.

3.2 Quadtree Segmentation of Stereoscopic Video Using Conditionally Hierarchical Split-and Merge

As shown in Figure 1, a change detector is used to detect the moving regions and the static ones. Firstly, the difference image between the current image and the previous image is determined. If an absolute pixel value of the difference image is lower than a threshold, the pixel is classified as static, otherwise it is classified as moving. If over fifty percent of pixels in a region are moving, then the region is regarded as a moving region, otherwise it is a static region. Meanwhile, postprocessing is performed to fill the holes and eliminate the isolated regions.

The quadtree segmentation is performed by splitting blocks of a predefined large block size into smaller blocks with uniform motion in a top-down manner. Before processing the split phase, a split criterion should be determined. That is, if the cost of

splitting a block is smaller than no splitting, then the block is split, otherwise, not split. For the purpose of rate-distortion optimization, the cost function is the summation of rate and distortion with the Lagrange multiplier λ as the weighting coefficient as defined in (3). Only those blocks which are classified as "moving blocks" are quadtree segmented.

The purpose of the merge phase is to refine the segmentation result of the split phase. The split phase is performed only in the left view of the stereoscopic image sequence, and the right view is used to determine the disparity vector field. The depth information can be derived directly from the disparity vector field. As depicted in [6], under the assumptions that the convergence angle is approaching zero and the length of viewing line is much larger than the base line, the depth of object i can be approximated by:

$$depth_i = d_{i,x} / \left(d_{i,x}^2 + d_{i,y}^2 \right) \quad (10)$$

where $(d_{i,x}, d_{i,y})$ is the disparity vector of object i. If the difference of depth between two adjacent objects is below a threshold, then the two objects are merged into one object and the new motion vector is the one with smaller distortion chosen from the motion vectors of the two original objects.

4. EXPERIMENTAL RESULTS

Table 2 compares the simulation results of average PSNR, bits/pixel and relative computation time with the test image sequence "Tunnel" encoded at 384 kbits/sec using the proposed jointly R-D constrained quadtree segmentation and motion estimation method, the R-D constrained motion estimation method, and the traditional TMN5 H.263 coder respectively. The proposed scheme is evidently the most efficient one among the three coding schemes in Table 2 by achieving more than 1 dB PSNR improvement on the left-view sequence over the other two schemes. The computation cost required for the proposed scheme is, however, a bit higher than the TMN5 coder. Figure 3 shows the PSNR performance comparison of the first 10 frames of "Tunnel" for the three methods mentioned above. Table 2 also indicates the percentage of the number of bits required for encoding the motion vectors, the segmentation information, the prediction errors, and I frames. It is shown that the number of bits required for encoding the motion vectors is significantly reduced thus more data bits can be assigned to encode the residuals thus leading to performance improvement. This coding strategy will be especially advantageous in low bit rate applications. Figure 4 illustrates the quadtree segmentation results at each stage. Only about 15% regions are classified as moving regions as shown in Figure 4, thus the R-D computation only needs to be performed on a small portion of an image using the proposed method.

5. REFERENCES

[1] D. Tzovaras, M.G. Strintzis, and H. Sahinoglou, "Evaluation of multiresolution block matching techniques for motion and disparity estimation," *Signal Processing: Image Communication*, vol. 6, no. 1, pp. 59-67, March 1994.

[2] Chia-Wen Lin, Eryin Fei, and Yung-Chang Chen "Hierarchical disparity estimation using spatial correlation," *IEEE Trans. Consumer Electronics,* vol. 44, no. 3, pp. 630-637, August 1998.

[3] G. J. Sullivan and R. L. Baker, "Efficient quadtree coding of image and video," *IEEE Trans. Image Processing*, vol. 3, no. 4, pp. 327-331, May 1994.

[4] Y. Shoham and A. Gersho, "Efficient bit allocation for an arbitrary set of quantizers," *IEEE Trans. Acoustic, Speech, and Signal Proc.*, vol. 36, pp. 1445-1453, Sept. 1988.

[5] K. Ramchandran and M. Vetterli, "Best wavelet packet bases in a rate-distortion sense," *IEEE Trans. Signal Proc.*, vol. 2, pp. 160-175, April 1993.

[6] D. Tzovaras, N. Grammalidis, and M.G. Strintzis, "Object-based coding of stereo image sequence using joint 3-D motion/disparity compensation," *IEEE Trans. on Circuit and System For Video Tech.*, vol. 7, no. 2, pp. 312-327, April 1997.

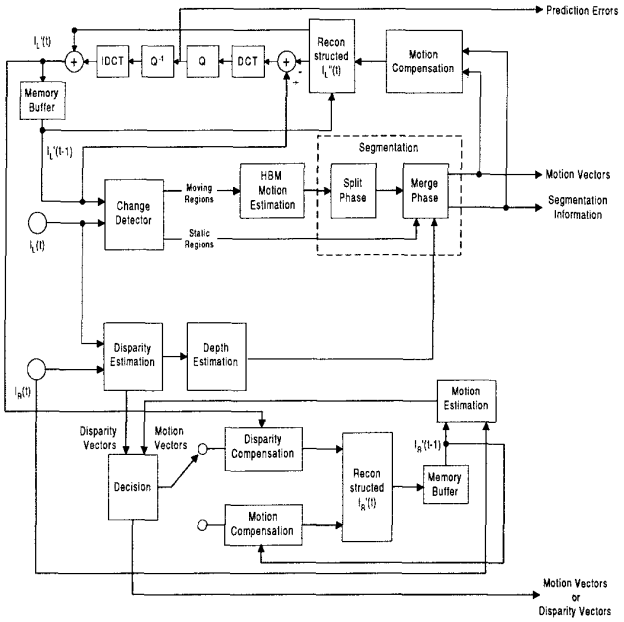


Figure 1. The proposed stereoscopic video encoder

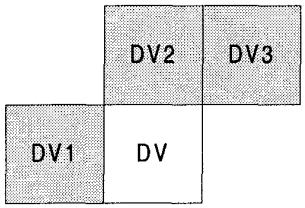


Figure 2. DV: Current displacement vector; DV1, DV2, DV3: Predictors

	Motion search range		Disparity search range	
Level	0	1	0	1
X	±5 (half)	±7 (full)	±5 (half)	±7 (full)
Y	±5 (half)	±7 (full)	±3 (half)	±1 (full)

Table 1. Motion and disparity estimation search range when using 2-level pyramid

	Tunnel (384 Kb/s)		
	PSNR (dB)	bits/pixel	Time
H.263 TMN5	31.664 (left)	M: 10%, E: 62.5%, I: 27.5%	100%
RDO for ME / MC	31.719 (left)	M: 5.2%, E: 67.6%, I: 27.2%	149.27%
Proposed Scheme	32.736 (left, split)	M: 1.6%, S: 0.6%, E: 70.1%, I: 27.1%	128.64%
	32.912 (left, split + merge)	M: 1.2%, S:0.6%, E: 71.6%, I: 26.6%	155.63%
	25.612 (right)		35.72%

Table 2. Performance of the proposed stereoscopic coding scheme and two other methods. M, S, E, and I in the item of bits/pixel denote the percentage of the number of bits needed for coding motion vectors, segmentation information, prediction errors, and the intra frame respectively.

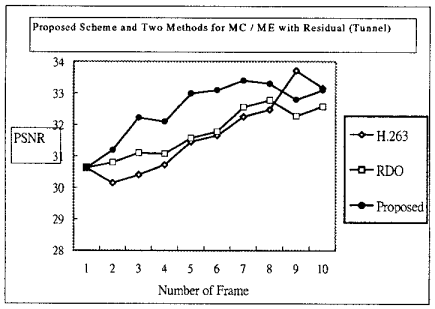
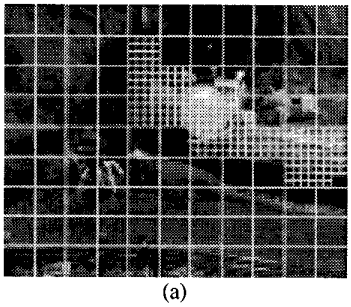
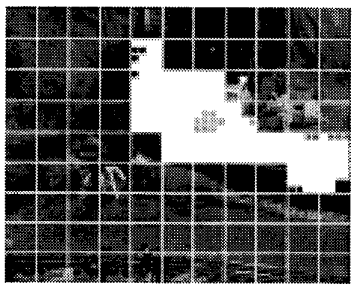


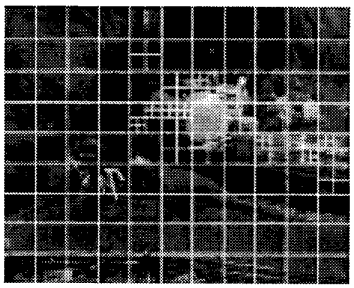
Figure 3. PSNR of the three image coding schemes shown in Table 2 for "Tunnel" at the bit rate of 384 kbits/sec



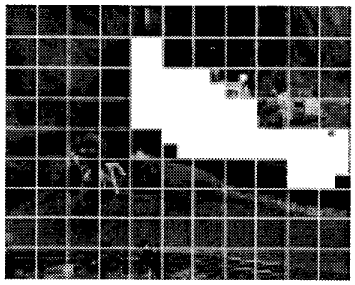
(a)



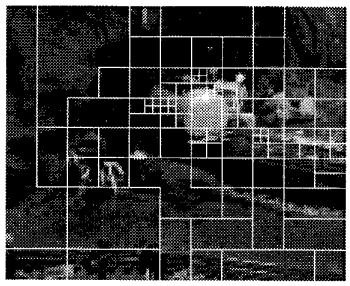
(b)



(c)



(d)



(e)

Figure 4. R-D constrained quadtree segmentation results. (a) Initial split result of frame 1 of "Tunnel"; (b) region classification of (a), white areas are moving regions and the other are static regions; (c) final split result; (d) final region classification results; (e) final segmentation result after merge phase