Published on Picture Coding Symposium 1999, March 1999

Efficient Video Coding with R-D Constrained Quadtree Segmentation

Chia-Wen Lin
Computer and Communication Research Labs
Industrial Technology Research Institute
Hsinchu, Taiwan 310, ROC
ljw@n100.ccl.itri.org.tw

Yao-Jen Chang , Eryin Fei, and Yung-Chang Chen
Department of Electrical Engineering
National Tsing Hua University
Hsinchu, Taiwan 300, ROC
{kc,eryin,ycchen}@benz.ee.nthu.edu.tw

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 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 estimation scheme as well as a face model-assisted conditional optimization strategy is proposed to drastically reduce the large computation cost required in the R-D optimization process. A novel skin-color based face detection scheme is also proposed for fast locating face regions. The simulation results show that the proposed method can achieve more than 0.5 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 [3-4]. 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 [2]. The optimization of segmentation in rate-distortion sense is, however, very computation intensive thereby making it impractical in real-time applications. In this paper, we develop a framework of R-D constrained quadtree segmentation for video coding. A fast hierarchical motion estimation scheme as well as a model-assisted conditional optimization strategy is proposed to drastically reduce the huge computation load required for R-D optimization process.

2. R-D CONSTRAINED VIDEO CODING

Let $v_i \in \boldsymbol{V}$ be the displacement vector corresponding to the block i of the image, where $\boldsymbol{V}$ 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 $R_{\boldsymbol{target}}$ for transmitting the displacement vector field, the segmentation information and the error image. This corresponds to the following constrained optimization problem:

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

subject to

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

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, ... , N } of the unconstrained minimization of the cost function $C(\lambda)$:

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

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 [3] 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 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.

3. THE PROPOSED R-D CONSTRAINED QUADTREE SEGMENTATION

The proposed architecture to realize the aforementioned framework of R-D constrained video coding is depicted in Figure 1. A hierarchical splitting scheme based on motion information is used to segment an image into variable-size blocks with uniform motion. The R-D optimization process performs quadtree segmentation and motion estimation in an iterative fashion to find the best match which is too computation intensive to be used for 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 [1] is utilized to effectively reduce the computation load in motion 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 face model-assisted conditional optimization strategy can also save much computation time while still maintaining good video quality, especially on the regions of special interest.

3.2 A Novel Skin-Color Based Face Detectior

As aforementioned, face regions are of special interest to be R-D optimized in our work. Therefore an efficient detection scheme is required to locate face regions. It was shown in [5] that human skin-color based face detection approach can effectively identify the face location in real-time. In this work, we propose a novel low-complexity skin-color based face detection scheme which is robust against camera noises and skin-color-like interference of non-face objects by using joint motion/color probability model assisted decision. As shown in Figure 2, each pixel on the input video frame is classified as either skin color or non-skin color via the pre-determined skin-color model. A binary-tree split-and-merge segmentation process is performed to group the skin-color pixels into candidate face regions. Subsequently, A joint motion/color probability model is used to determine the most probable face region with the highest joint probability. After locating the face region of the first frame, the fast face tracking mode is used to speed up the detection process. Each functional unit is elaborated below.

A. The Motion Probability Map for Each Pixel

For video sequences, the frame differences provide the motion information. We have made an assumption that most motions in the video sequences are caused by objects in the head. The assumption is reasonable for video sequences mainly containing a person with head and shoulders, such as, the news programs, the videophone and the videoconference video, etc. Therefore, the location with higher the frame differences, the more probable the location is in the face region. Actually, by our observation, most motions are caused by the eyes' blinking, the mouth's opening and closing, and the head movements. We use the following equation to calculate the frame differences:

$$Dif(i,j) = \sum_{t=0}^{n} | f_t(i,j) - f_{t-1}(i,j) | \qquad (5)$$

where the $f_t\ (i, j)$ denotes the pixel value in the location (i, j) of the video sequence at time t. And the $f_{t-1}\ (i, j)$ denotes the pixel value at time t-1. The probability for the pixel at location (i, j) belongs to the face region is calculated in Eq. (6)

$$P_{face}(i,j) = Dif(i,j) \Big/ \sum_k \sum_l Dif(k,l) \qquad (6)$$

B. Color Probability Map for Each Pixel

The Bayesian decision rule described in [5] is adopted to set up the color probability model for each pixel in the Y-Cb-Cr color coordinate. In this work, the threshold TH for skin color classification is automatically set under the assumption that the face region cannot occupy more than sixty percent area of the whole frame. In this way, background noise can be reduced for detection of the face block. A binary search algorithm is used for fast auto-threshold setting.

C. Binary-Tree Split-and-Merge Segmentation

After thresholding by using TH, there may still exist some background noise or false detection of other objects in the image with skin-like colors. Therefore, further segmentation process is required to eliminate the noise region and separate different objects. Here, we propose a binary-domain split-and-merge algorithm with binary-tree partitioning. Neighboring pixels are grouped to form a face block candidate. Since the binary-tree segmentation is performed in the binary domain, the computation cost is very low. The split-and-merge algorithm is summarized as follows:

Split Phase:

1. The search region is set to be the whole picture.
2. Find the upper and lower boundaries of the skin-color block by thresholding the horizontal integral projection of the search region.
3. Find the left and right boundaries of the skin-color block by thresholding the vertical integral projection of the picture between the upper and lower boundaries.
4. If the fullness of the skin-color block is higher than 60% or the depth of the split is higher than the predefined maximum depth (max depth = 3 in the proposed scheme), goto step 5.
   Otherwise, divide the block equally in the horizontal direction to two search regions and repeat steps 2~4 for each of the two search regions with an increased depth.
5. Stop the split process for the current block and record the block boundary information.

Merge Phase:

For each recorded block resulting from the split phase, merge those connected blocks that have small differences in block width to form a skin-color group. And the width of the group is the weighted averages of the blocks belong to the group.

D. Joint Motion/Color Probability Model Assisted Decision

The normalized motion probability for each group is defined as follows:

$$Pm(Group_n) = \frac{\frac{1}{Area(Group_n)} \sum_{(i,j)\in Group_n} P_{face}(i,j)}{\sum_{k=1}^{\#\text{ of Groups}} \frac{1}{Area(Group_k)} \sum_{(i,j)\in Group_k} P_{face}(i,j)} \qquad (7)$$

And the normalized fullness of skin-color for each group is defined as :

$$F(Group_n) = \frac{\text{ratio of skin - color pixels in Group}_n}{\sum_{k=1}^{\#\text{ of Groups}} \text{ratio of skin - color pixels in Group}_k} \qquad (8)$$

With these two kinds of information for each group, we set different weighting factors W_m and W_f for Pm(.) and F(.) respectively. The probability for a group to be the face block can be calculated by Eq. (9):

$$P_{fb}(Group_n) = W_m \cdot P_m(Group_n) + W_f \cdot F(Group_n) \qquad (9)$$

The group with maximum value of P_{fb} is determined as the face block. The weighting factors W_m and W_f are empirically set as 1/3 and 2/3 respectively.

E. Fast Face Tracking Mode

After we extract the average color value in the detected group, unlike the general skin-color distribution that covers moderate ranges in the Cb-Cr plane, the skin-color for a specific person is restricted to a much smaller range. Hence, the extracted average color of the face region in the previous frame can be used as a very reliable reference model for tracking the face in the current frame. Meanwhile, the detected location of previous frame is taken as the initial guess to search the current face location within a small search window to speed up the tracking process.

3.2 Model-Assisted Video Coding with R-D Constrained Quadtree Segmentation

To save computation, the R-D optimization process is performed only on those regions of special interest. In video phone and videoconference applications, face regions are the most important parts for human perception. As shown in Figure 1, a change detector is used to detect the moving regions and the static ones within the face region. Firstly, the difference image between the faces in 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). As mentioned above, only those blocks which are classified as "moving blocks" within face region are quadtree segmented.

4. EXPERIMENTAL RESULTS

Table 1 compares the simulation results of average PSNR, bits/pixel and relative computation time with the test image sequence "Claire" and "Miss America" encoded at 384 kbits/sec using the proposed jointly R-D constrained quadtree segmentation and motion estimation method and the traditional TMN5 H.263 coder respectively. The experimental results show that the proposed scheme can achieve more than 0.5 dB PSNR improvement over the TMN5 coder. The computation cost required for the proposed scheme is, however, a bit higher than the TMN5 coder. Table 1 also indicates the average 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 considerably reduced such that more data bits can be assigned to encode the residuals thus leading to performance improvement. This coding strategy will be especially advantageous in very low bit rate applications. Figure 3 illustrates the quadtree segmentation results at each stage. Only about 15% regions are classified as moving regions as shown in Figure 3, thus the R-D computation only needs to be performed on a small portion of an image with the proposed method. The proposed method can also be combined with some model-assisted rate control schemes to further emphasize the quality on the face regions [6]

5. REFERENCES

[1] 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, Aug. 1998.

[2] 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.

[3] 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.

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

[5] Hualu Wang, and Shuh-Fu Chang, *"A Highly Efficient System for Automatic Face Region Detection in MPEG Video", IEEE Trans. on Circuits and Systems for Video Technology,* Vol.7, No. 4, 1997.

[6] J.-B. Lee and A. Eleftheriadis, "Spatio-Temporal Model-Assisted Compatible Coding for Low and Very Low Bitrate Videotelephony", ICIP-96, Lausanne, Switzerland, Sept. 1996, pp. II.429-II.432.

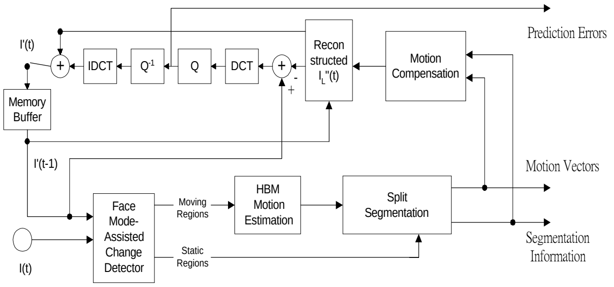

Figure 1. The proposed R-D contrained video encoder

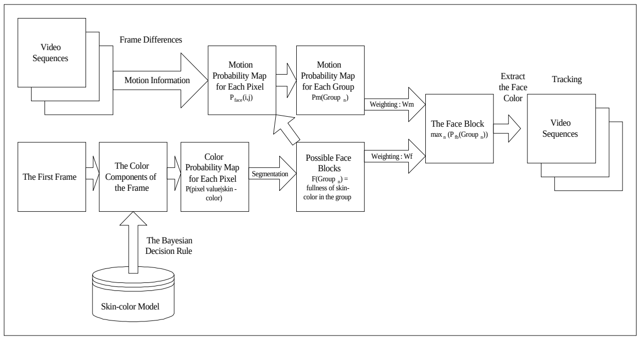

Figure 2. The flow chart of the propose skin-color based face detection scheme

	Clair (384 Kb/s)		
	PSNR (dB)	bits/pixel	Time
H.263 (TMN5)	36.51	0.0885（M:0.0026, E:0.0355, I:0.0504）	100%
Proposed Scheme	37.05	0.0881（M:0.0014, S:0.0004, E:0.0361, I:0.0504）	114.3%

(a)

	Miss America (384 Kb/s)		
	PSNR (dB)	bits/pixel	Time
H.263 (TMN5)	38.03	0.0479（M:0.0051, E:0.0217, I:0.0211）	100%
Proposed Scheme	38.61	0.0482（M:0.0021, S:0.0014, E:0.0236, I:0.0211）	145.6%

(b)

Table 1. 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.

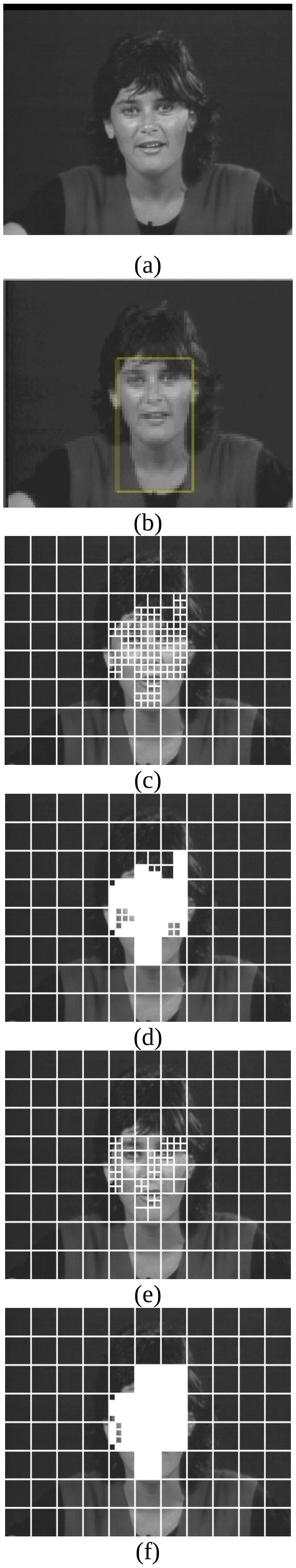
(a)
(b)
(c)
(d)
(e)
(f)

Figure 3. (a) original image of frame 0, (b) face detection result of (a), (c) initial segmentation result, (d) the indication of moving objects in (c) (white areas), (e ）, (f) the final segmentation result.