Dynamic Bit Allocation for Multiple Video Object Coding

Zhenzhong Chen, *Student Member, IEEE*, Junwei Han, and King Ngi Ngan, *Fellow, IEEE*

***Abstract*—In MPEG-4, a visual scene may be treated as a composition of video objects and coded at object level. Such a flexible video coding framework makes it possible to code different video objects with different priority according to human perceptual characteristics. In this paper, we introduce a novel dynamic bit allocation framework to improve the subjective quality in such an object-based video coding system. We incorporate the rate distortion models with the dynamic priorities of the video objects and jointly encode video objects to minimize the weighted distortion within the bit budget constraint. We guarantee the human-interested video objects a better reconstructed quality by using the weighted bit allocation strategy in favour of the video objects with higher priority. To obtain the priority automatically, we apply a visual attention model. Comparing with traditional bit allocation algorithms, the objective quality of the object with higher priority is significantly improved under this framework. These results demonstrate the usefulness of this dynamic bit allocation framework.**

***Index Terms*—Bit allocation, MPEG-4, multiple video object, rate control, rate-distortion.**

I. Introduction

THE second-generation image coding [1] and object-oriented video coding [2] have culminated in the first international object-based video coding standard, MPEG-4 [3]. MPEG-4 treats a scene as a composition of several video objects (VOs) that are separately encoded and decoded. Instances of video object in a given time are called video object planes (VOPs). A natural video object consists of a sequence of VOPs. Fig. 1 presents an MPEG-4 multiple video object coding structure with four video objects. Such a flexible video coding framework makes it possible to code different video objects with different distortion scales. For example, a stationary background can be coded more coarsely than the complex foreground video object without much effect on the subjective quality. More bits can be saved to encode the foreground video object which viewers are more interested in. It is necessary to analyze the priority of the video objects according to its semantic importance, intrinsic properties and psychovisual characteristics such that the bit budget can be distributed properly to video objects to improve the perceptual quality of the compressed video. Due to the bit budget constraint, bit allocation is extremely important since the quality of video objects is sensitive to the bit allocation strategy.

Some typical multiple video object bit allocation algorithms have been studied in [4]–[6] by adopting the traditional frame-based rate distortion analysis technique [7]. They aimed to provide practical bit allocation strategies by considering several bit rate related features of video objects. Lee *et al.* extended the work in [8] to provide a spatio–temporal tradeoff for object based video coding [9]. An operational rate-distortion based approach to allocate bits among multiple video objects and a ρ domain based optimal bit allocation were presented in [10] and [11], respectively. Chen *et al.* provided an optimal bit allocation approach in [12] using object-based rate distortion models [13], [14] and further extended the optimal bit allocation into the joint texture-shape bit allocation framework [15]. However, the video objects were equally treated in these works for objective optimization purpose. Although some distortion smoothing mechanisms were adopted such as defining the quantization parameter variation constraints, the video objects which attract human attention may suffer higher distortion than other video objects.

Recent video coding techniques aim to provide subjectively optimized quality by relying on characteristics of the human visual system [16], [17], appropriate quantizer design can improve the perceptual quality [18], [19], or neurobiological properties of visual perception [20]. It is well known that the human visual system (HVS) is highly space-variant and the spatial resolution is highest at the the foveation point. So the object at the foveation point in the scene coded with better quality significantly contributes to the subjective quality. Foveated video compression algorithms [21] have been proposed to deliver high-quality video at reduced bit rates by seeking to match the nonuniform sampling of the human retina. An optimal rate control approach is then to maximize a foveal visual quality metric, the foveal signal-to-noise ratio (FSNR) to determine the best compression and rate control parameters for a given target bit rate [22]. Since perceptual quality measurements [23]–[28] are complex, an optimal subjective rate control needs to balance the tradeoff between the algorithm complexity and the R-D performance. Priority-based bit allocation for multiple video objects in MPEG-4 was proposed in [29]. The video object with higher priority is assumed to be more important and should be encoded with better quality. However, the priority of the video object has to be set before encoding and could not be adjusted during coding.

This paper aims to provide a dynamic bit allocation framework as shown in Fig. 2. To obtain the priorities of video objects automatically, we apply an object-level visual attention

Manuscript received March 7, 2005; revised February 13, 2006. This work was supported by a grant from the Research Grants Council of Hong Kong SAR under Project CUHK4229/04E supplemented with a Direct Grant 2050316, and is affiliated with the Microsoft-CUHK Joint Laboratory for Human-centric Computing and Interface Technologies. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Chang Wen Chen.

The authors are with the Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, NT, Hong Kong (e-mail: zchen@ee.cuhk.edu.hk; jwhan@ee.cuhk.edu.hk; knngan@ee.cuhk.edu.hk).

Digital Object Identifier 10.1109/TMM.2006.884633

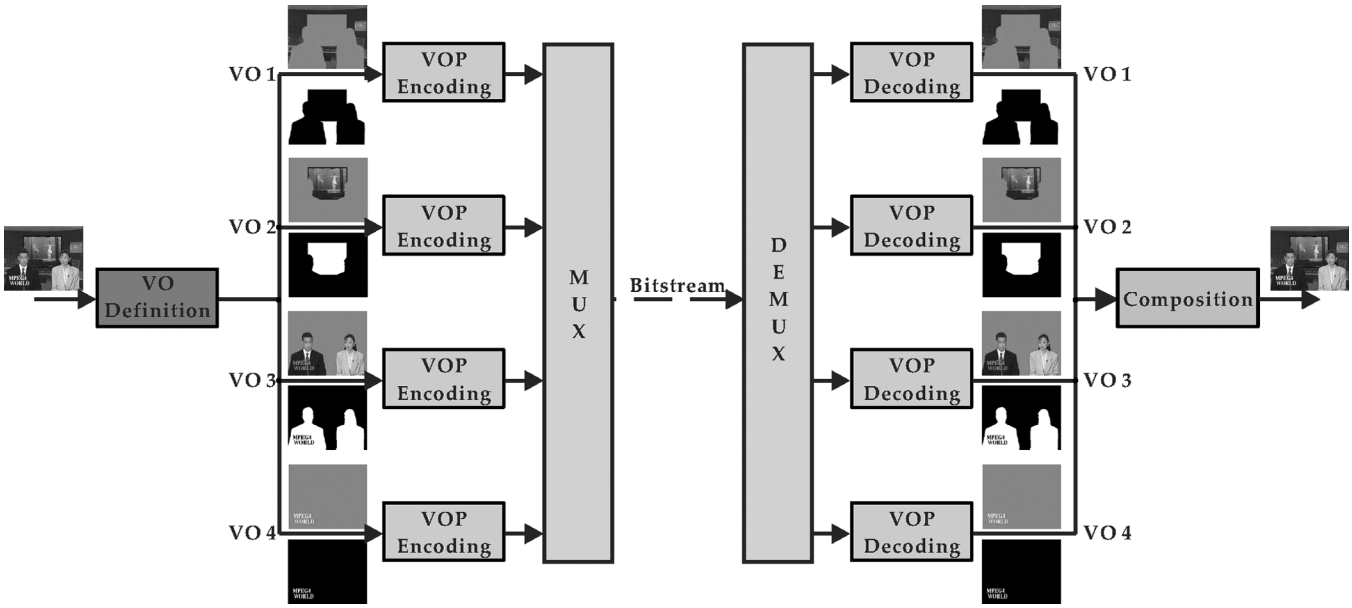


Fig. 1. MPEG-4 multiple video object coding structure.

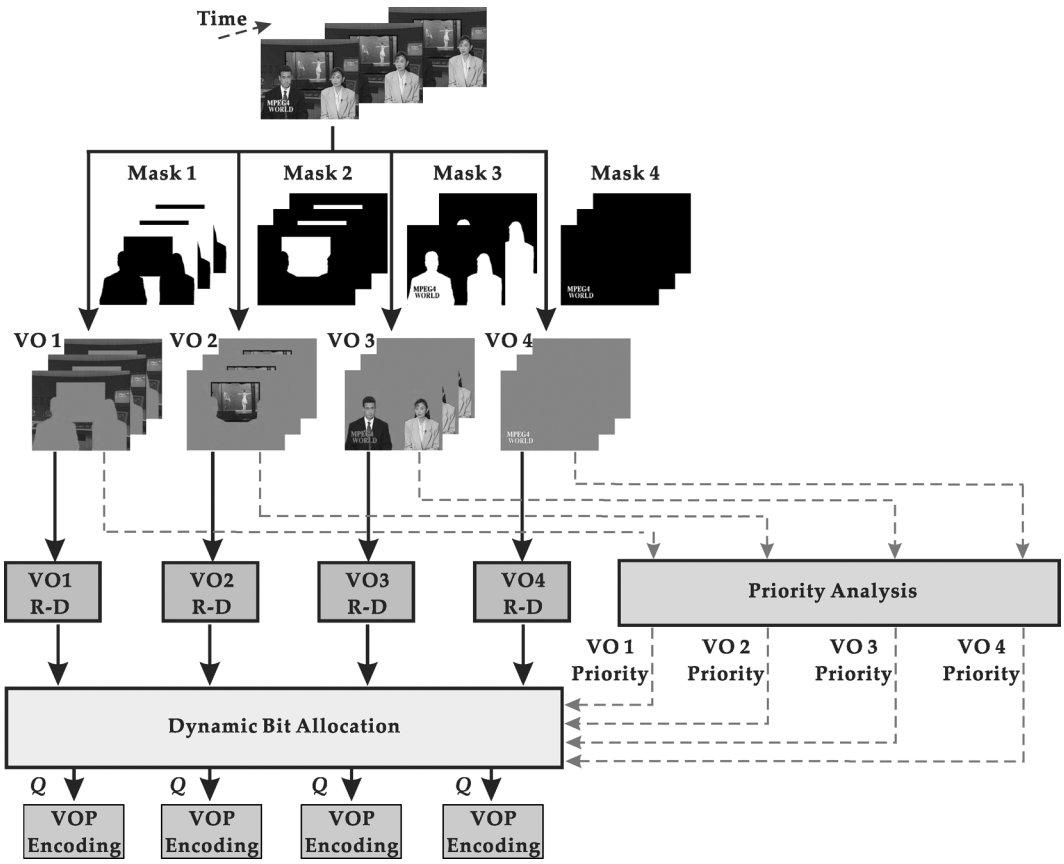


Fig. 2. Framework of dynamic bit allocation for multiple video object coding.

model. We attempt to guarantee the human-interested video object a better reconstructed quality. Due to the bit budget in video coding, constrained optimization must be employed to ensure that the optimal result can be achieved. One significant contribution of this work is that the human visual system characteristics can be incorporated in the video coding optimization process dynamically. Another advantage is that the priority of the video object can be obtained automatically instead of fixing the weighting factors before encoding. It is noted that the performance of the framework relies on the accuracy of the attention model.

The organization of the paper is as follows. The next section gives a brief review of the traditional MPEG-4 multiple video object bit allocation and the optimal multiple video object bit

allocation algorithms. In Section III, the dynamic bit allocation framework is presented. An object-level visual attention model is introduced and applied in the dynamic bit allocation framework as described in Section IV. We provide the experimental results and the concluding remarks in Sections V and VI, respectively.

II. Review of Multiple Video Object Bit Allocation

In traditional multiple video object bit allocation [4], three features are considered: the size, motion, and MAD^2. The target bit rate for object i is given by

$$R_i = R_{\text{target}} \cdot (w_s \mathrm{SIZE}_i + w_m \mathrm{MOT}_i + w_v \mathrm{VAR}_i) \tag{1}$$

where R_{target} is the target bit rate for all the video objects, SIZE_i, MOT_i, and VAR_i are the size, motion and MAD^2 (Mean Absolute Difference)^2 of object i, respectively. The weights $\{w_s, w_m, w_v\} \in [0,1]$ satisfy the constraint $w_s + w_m + w_v = 1$. Typical values of the weighting factors in the MPEG-4 verification model are 0.4, 0.6, 0 for LowMode and 0.25, 0.25, 0.5 for HighMode. LowMode and HighMode are operational modes set by the skip threshold. The weighting factors are defined before encoding. After obtaining the target bit rate of the video object, the corresponding quantization parameter is derived from the rate-quantizer model. The advantage of this bit allocation method is its simplicity. However, the corresponding distortion of the video object is not considered. Since the content of the scene is varying, such a bit allocation is neither flexible nor robust.

An optimal multiple video object bit allocation approach has been proposed in [12] to maximize the objective quality according to the rate-distortion characteristics of the video objects. Generally speaking, the optimal bit allocation aims to distribute the available bit budget amongst the different sources such that the overall distortion can be minimized:

$$\min\left(\sum_{i=1}^{m} D_i\right) \quad \text{subject to} \sum_{i=1}^{m} R_i \le R_{\text{target}} \tag{2}$$

where D_i and R_i denote the distortion and bit rate of the ith VO, respectively. m is the number of the VOs. The rate-distortion characteristics of the individual video object (IVO) are obtained via the rate-distortion model of the video object. The constrained optimal bit allocation solution is solved by dynamic programming. The maximum objective quality of the visual scene can be achieved in this approach. However, all the video objects are equally treated which means that the video object which attracts human attention may suffer lower visual quality compared with other video objects in the scene.

III. Dynamic Bit Allocation Framework

In this section, we present the dynamic bit allocation framework for multiple video object coding as shown in Fig. 2. The object rate-distortion characteristics are explored first by the object-based rate-distortion models. We analyze the priority of the video objects and define the higher weighting factors for higher priority objects. Using these weighting factors, the dynamic bit allocation strategy is realized. Within the bit rate budget constraint, the subjective optimized quantization parameters are then obtained and sent to the VOP encoders to encode the VOs separately.

A. Rate-Distortion Modeling

To explore the rate-distortion characteristics of the video objects, we employ a quadratic rate-quantizer (R-Q) model for the video object i [13]:

$$R_i^k = C_i^k \left(\frac{a_{1,i}^k}{Q_i^k} + \frac{a_{2,i}^k}{Q_i^k \times Q_i^k}\right) \tag{3}$$

where R_i^k is the bits of current VOP k of object i, C is the encoding complexity indicated by the sum of absolute difference (SAD), Q denotes the quantization parameter, a_1 and a_2 denote the model parameters that are updated by linear regression method from the previous coded parameters.

The corresponding distortion model of the video object is defined as [12]

$$D_i^k = b_{1,i}^k \times Q_i^k + b_{2,i}^k \times Q_i^k \times Q_i^k \tag{4}$$

where D_i^k is the distortion for current VOP k of object i, b_1 and b_2 denote the model parameters that are updated by linear regression method from the previous coded parameters.

B. Dynamic Bit Allocation

Assuming we have obtained the priorities of the video objects dynamically, we incorporate the corresponding weighting factors with the rate-distortion models to achieve the minimum weighted distortion:

$$\min\left(\sum_{i=1}^{m} W_i^k D_i^k\right) \quad \text{subject to} \sum_{i=1}^{m} R_i^k \le R_{\text{target}}^k \tag{5}$$

where W_i^k is the weighting factor of VOP k of the video object i. Different to other weighted bit allocation schemes [5], [11], the weighting factor is adaptive in our framework. Since the video object priorities may change with time, it is reasonable to update the object priorities dynamically. However, exploring the dynamic priority of the video object results in higher computational complexity and the coding performance relies on the quality of the obtained priorities. We will use an attention model which can help us to update the priorities automatically, as described in the next section.

The total rate for encoding all the video objects is given by the sum of the rate of each video object as follows:

$$R^k\left(Q_1^k, \cdots, Q_m^k\right) = \sum_{i=1}^{m} R_i^k\left(Q_i^k\right). \tag{6}$$

Similarly, the distortion cost for all the video objects is the sum of the weighted distortions of all the video objects:

$$D^k\left(Q_1^k, \cdots, Q_m^k\right) = \sum_{i=1}^{m} W_i^k D_i^k\left(Q_i^k\right). \tag{7}$$

The Lagrange multiplier method can merge the rate and distortion optimization with a Lagrangian multiplier λ. The

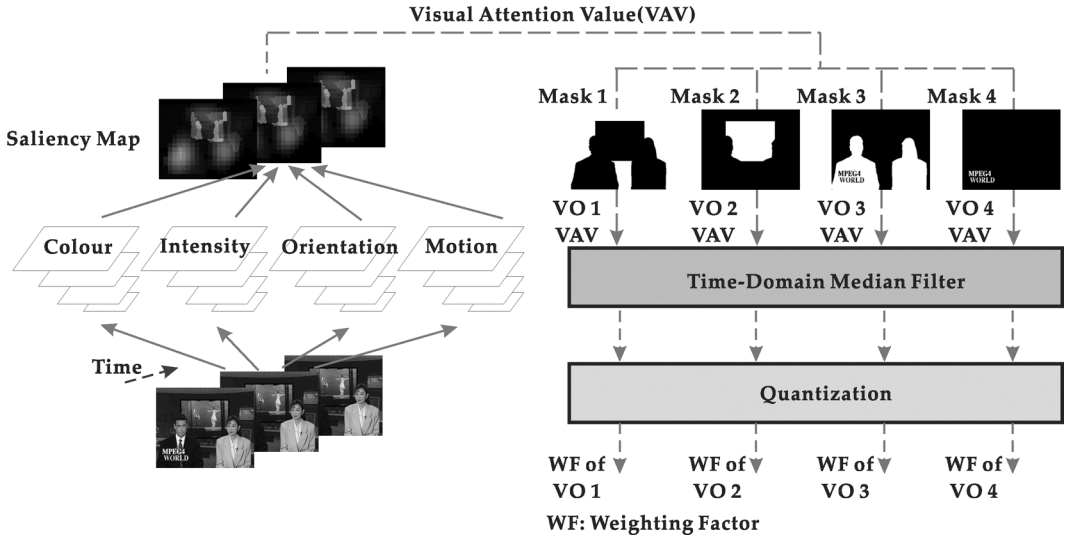


Fig. 3. Video object priority analysis using visual attention model.

constrained problem (5) is converted into the following unconstrained one:

$$\Gamma_{\lambda}^{k}\left(Q_1^k,\cdots,Q_m^k\right)=D^k\left(Q_1^k,\cdots,Q_m^k\right)+\lambda R^k\left(Q_1^k,\cdots,Q_m^k\right). \tag{8}$$

It has been shown in [30], [31] that if a λ^* satisfies

$$\left[Q_1^{k*},\cdots,Q_m^{k*}\right]=\arg\min_{Q_1^k,\cdots,Q_m^k}\Gamma_{\lambda^*}^{k}\left(Q_1^k,\cdots,Q_m^k\right) \tag{9}$$

and

$$R^k\left(Q_1^{k*},\cdots,Q_m^{k*}\right)=R_{\text{target}}^k \tag{10}$$

then $[Q_1^{k*},\cdots,Q_m^{k*}]$ is the optimal solution of (5). The bisection algorithm [32] can be used to find λ^*. The optimal solution of (5) can also be found by using the dynamic programming (DP) algorithm [12], [32], [33].

IV. Object-Level Visual Attention Model

To obtain the priorities and corresponding weighting factors of the video objects, we employ the object-based visual attention model as shown in Fig. 3. We analyze the visual attention property of the visual scene and derive the object-level visual attention model by incorporating the saliency map [34] and the video object masks. The visual attention values of the video objects are calculated and incorporated in the bit allocation mechanism. The visual attention model [34] generates the saliency map to indicate the saliency in the visual scene based on human perceptual characteristics. Three features, color, intensity and orientation, are extracted for stationary visual attention analysis. Then several multiscale maps are formed and center-surround operation is performed to explore the local contrasts for each feature. Finally the saliency map is derived by combing all the feature maps [34]. So far only low-level features are considered in such a saliency map. Here we add a high-level feature, location, since the human generally pays more attention to the object at the center of the visual scene. Therefore we adopt a normalized Gaussian function with the center located at the scene center to redefine the saliency map S_2^k [35]:

$$S_2^k(i,j)=S_1^k(i,j)\cdot G(i,j) \tag{11}$$

where (i,j) is a position of the pixel in the frame k, $S_1^k(i,j)$ is the original saliency map [34], $G(i,j)$ is the normalized Gaussian function centered at $((M-1)/2,(N-1)/2)$ where M and N are the width and height of the picture, respectively.

Actually, the saliency map is a good indictor of the static attention for each pixel. But for simplicity, our basic processing unit is a region of 16×16 pixel block. Consequently, we estimate the attention values for regions. Let Ω_n^k be a region in the frame k, and $p(i,j)$ be a point in the region Ω_n^k. If $S_2^k(i,j)>0$, $p(i,j)$ is regarded as the *attention pixel* with respect to the attention model. $N_t(\Omega_n^k)$ denotes the number of pixels in the region Ω_n^k, and $N_p(\Omega_n^k)$ refers to the number of *attention pixels* in the region Ω_n^k. The static attention value for the region Ω_n^k is defined as

$$S_s^k\left(\Omega_n^k\right)=\frac{N_p\left(\Omega_n^k\right)}{N_t\left(\Omega_n^k\right)}\left[1-\exp\left(-\sum_{(i,j)\in\Omega_n^k}S_2^k(i,j)\right)\right]. \tag{12}$$

From above equation, we can see that a larger number of *attention pixels* results in a larger attention value which indicates that such a region Ω_n^k attracts more human attention.

By integrating the motion feature for moving visual attention, the saliency map of the visual scene in the video can be derived. The global motion estimation and compensation is first required since the background may not be stationary. We adopt the six-parameter affine motion model to estimate the global motion [36]. After the global motion has been compensated, the background of video sequences could be treated as being stationary.

Normally, objects with fast motion are easily perceived by human beings. It has been shown that the intensity-level difference between two consecutive frames is a simple but efficient

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

Fig. 4. Test sequence *News*: (a) original scene, (b) video object 1, (c) video object 2, (d) video object 3, (e) video object 4.

way to localize object motion. The intensity difference at (i,j) is

$$\Delta I^k(i,j) = \left| I^k(i,j) - I^{k-1}(i,j) \right| \tag{13}$$

where $I^k(i,j)$ is the intensity values at (i,j) in frame k. If $\Delta I^k(i,j)$ is greater than a given threshold σ, pixel (i,j) is considered as a *moving pixel*. A significance test [37] is implemented to determine σ and suppress the adverse effect of noise. Let $N_m(\Omega_n^k)$ be the number of *moving pixels* in the region Ω_n^k. The motion attention value of the region Ω_n^k can be derived as

$$S_m^k\left(\Omega_n^k\right) = \frac{N_m\left(\Omega_n^k\right)}{N_t\left(\Omega_n^k\right)} \times \left[1 - \exp\left(-\sum_{(i,j)\in\Omega_n^k, \Delta I^k(i,j)>\sigma} \Delta I^k(i,j)\right)\right]. \tag{14}$$

By integrating $S_s^k(\Omega_n^k)$ and $S_m^k(\Omega_n^k)$, we get the final attention value for the region Ω_n^k

$$S^k\left(\Omega_n^k\right) = \nu_s \overline{S_s^k\left(\Omega_n^k\right)} + \nu_m \overline{S_m^k\left(\Omega_n^k\right)}, 0 \le \nu_s, \nu_m \le 1 \tag{15}$$

where ν_s, ν_m are the weights with typical value of 0.5 for linear combination, and $\overline{S_s^k(\Omega_n^k)}$ and $\overline{S_m^k(\Omega_n^k)}$ are the normalized static and motion attention values, respectively.

After obtaining the final saliency map S^k of the frame, we integrate it with the video object mask and calculate the mean of the visual attention values of the video object to obtain the video object-level visual attention value (VAV):

$$S_{\mathrm{vo},i}^k = S^k \cdot M_i^k \tag{16}$$

$$VAV_i^k = \mathbf{E}\left(S_{\mathrm{vo},i}^k\right) \tag{17}$$

where M_i^k is the binary mask of VOP k of the video object i, VAV_i^k denotes the visual attention value of the video object plane (VOP) k of video object i and $\mathbf{E}$ is the mean operator.

Considering the duration of visual effect, we employ the temporal median filter to guarantee the smoothness of the visual attention:

$$V_i^k = \mathbf{MED}\left[VAV_i^{k-a}, \cdots, VAV_i^k, \cdots, VAV_i^{k+a}\right] \tag{18}$$

where $\mathbf{MED}$ denotes the median filtering procedure, $K = 2a+1$ is the window length. In this paper, we choose $a = 1$ to perform median filtering with window length $K = 3$.

After median filtering, the visual attention value of the video object is quantized to get quantized object attention value as the weighting factor of the video object for the dynamic bit allocation:

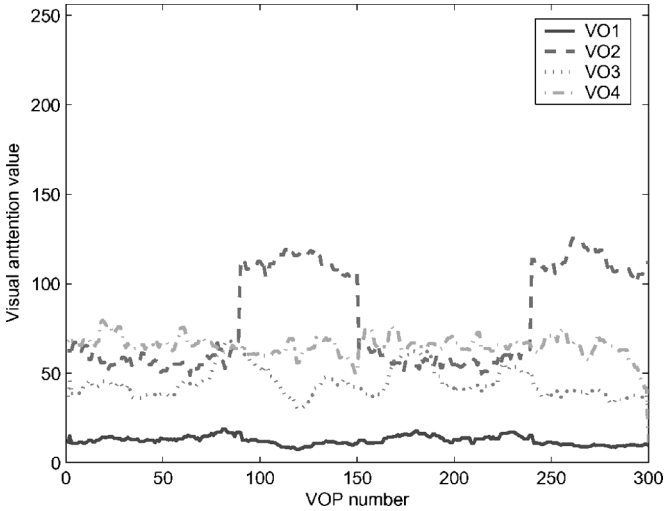


Fig. 5. Visual attention values of video objects in test sequence *News*.

$$W_i^k = \mathrm{round}\left(\frac{V_i^k}{\rho}\right) \tag{19}$$

where W_i^k is the weighting factor of VOP k of the video object i and ρ is the quantization parameter for V_i^k, respectively. The quantization process is used to tolerate small visual attention value differences from different clients by reducing the number of levels of visual attention value.

V. Simulation Results

To evaluate the performance of the proposed approach, we compared it with the traditional verification model bit allocation [38] and the optimal bit allocation algorithms [12]. We used the CIF format test sequence *News* composed of four video objects as shown in Fig. 4 and *Akiyo* which is a videoconferencing sequence consisting of only two objects, i.e., foreground and background objects. The video objects of *News* were coded jointly at a bit rate of 512 kbps and a frame rate of 30 fps and the objects of *Akiyo* were coded jointly at the bit rate of 256 kbps and a frame rate 30 fps. The quantizer ρ for object visual attention value was set to 16.

In the *News* sequence, the video object 2 (VO2) has varying content which catches the human's attention. Using the object-level visual attention model, the high priority of this video object was detected automatically such that more bits were assigned to it to improve the visual quality as indicated in Fig. 5. It is obvious that the visual attention value of the background, VO1, is

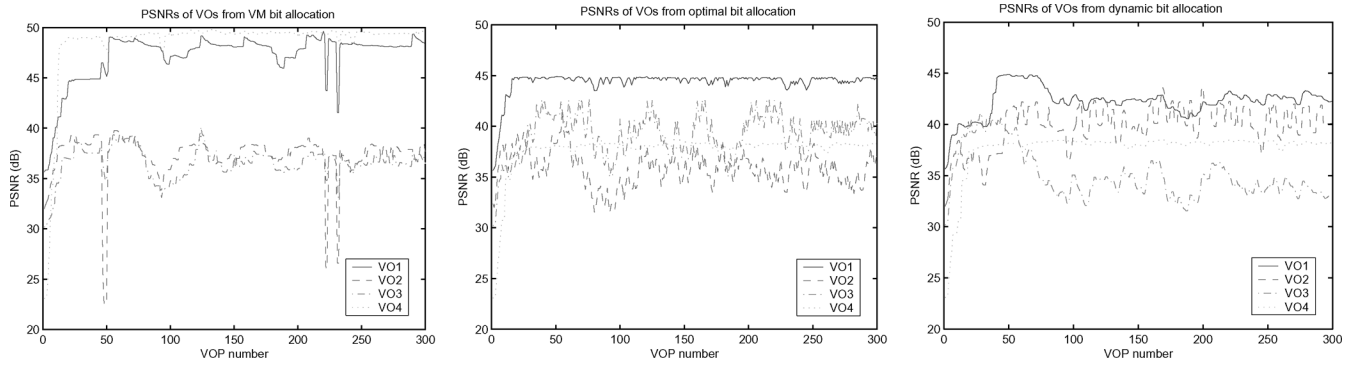


Fig. 6. Comparisons of PSNRs of individual video objects in test sequence *News*: (a) verification model bit allocation, (b) optimal bit allocation, (c) dynamic bit allocation.

TABLE I
SIMULATION RESULTS OF *News*

	PSNR (dB)				
	VO1	VO2	VO3	VO4	Overall
Verification Model Bit Allocation	**45.92**	36.13	35.12	**46.61**	36.48
Optimal Bit Allocation	44.45	34.58	**39.87**	37.03	**37.37**
Dynamic Bit Allocation	42.23	**39.87**	35.05	37.64	36.67

the lowest whilst that of VO2 is the highest in our approach compared to the other two algorithms. As VO3 has little movement compared to VO2, the visual attention value of VO3 is lower that of VO2.

Various distortion measurements such as objective criterion PSNR, perceptual meaningful measurement SSIM [39], and other image fidelity metrics [40] have been proposed to assess the quality of the decoded picture. In this paper, we use the PSNR which is the most widely used criterion in video coding. We assess the objective quality by comparing the PSNR values of the same VO produced by different bit allocation algorithms. Fig. 6 provides the comparisons of PSNRs of individual video objects employing the same algorithm, namely VM bit allocation, optimal bit allocation, and dynamic bit allocation.

The weighted bit allocation employing (5) is to fully utilize the available bits while improving the visual quality of the higher priority objects subject to the bit constraint. It is fair to compare the PSNRs of the same video object employing the three different algorithms, but not the PSNRs of different objects employing the same algorithm. Table I lists the simulation results of the three algorithms. From the results in Table I, we find that our proposed approach encodes VO2, with the highest quality (lowest distortion) where the objective quality improvement is up to 4.7 dB. Although the attention priority of the background VO1 is low, it is noted that the texture of VO1 is uniform such that even when we use a large quantization parameter in VOP encoding, the PSNR of this video object is still high. However, the quality of VO1 from our proposed approach is the lowest of the three algorithms whilst the PSNR of VO2 is the highest. The saved bits on encoding the low priority video object were allocated to other video objects of higher priority to improve their visual quality. This demonstrates the usefulness of our dynamic bit allocation framework.

Fig. 7 presents some samples of the reconstructed frames from three different algorithms. As shown in the pictures, we observe that VO2, which is the human-interested object, has the best subjective quality employing the proposed algorithm as illustrated in Fig. 7(d). The dancer in Fig. 7(d) shows fewer artifacts around the edges and finer texture details. Since we require more calculations to obtain the object-based visual attention value and the perceptual weights for subjective optimization, the proposed approach is more complex than the traditional methods. It should also be noted that the performance relies on the accuracy of the attention model.

For the foreground/background video *Akiyo*, people always pay more attention to the foreground object. The attention model distinguishes the higher priority of the foreground object and sets a larger weighting factor for the foreground object such that more bits were allocated to it. From the results in Table II, we can observe that the quality of the foreground object employing our proposed approach is the highest amongst the three algorithms. The weighting factor incorporated in the dynamic bit allocation framework improves the objective quality of human-interested object, i.e., the foreground video object, such that the subjective quality of the overall reconstructed picture is more pleasing to look at.

VI. CONCLUDING REMARKS

In this paper, we presented a novel dynamic bit allocation framework to improve the overall subjective quality of the image by improving the objective quality of the human-interested object. The priority of the video object is determined by the object-level visual attention model dynamically. According to the priorities of video objects, video objects are jointly encoded to minimize the weighted distortion within the bit rate constraint. The incorporation of the weighted optimal bit allocation and the object visual attention model fully utilizes the bit budget and improves the visual quality of the higher priority objects. As the object-based visual attention value is necessary to obtain the perceptual weights for the dynamic bit allocation, the proposed approach is more complex than the traditional methods. It should also be noted that the performance relies on the quality of the attention model. Compared to the traditional bit allocation algorithms, the objective quality of the interesting object

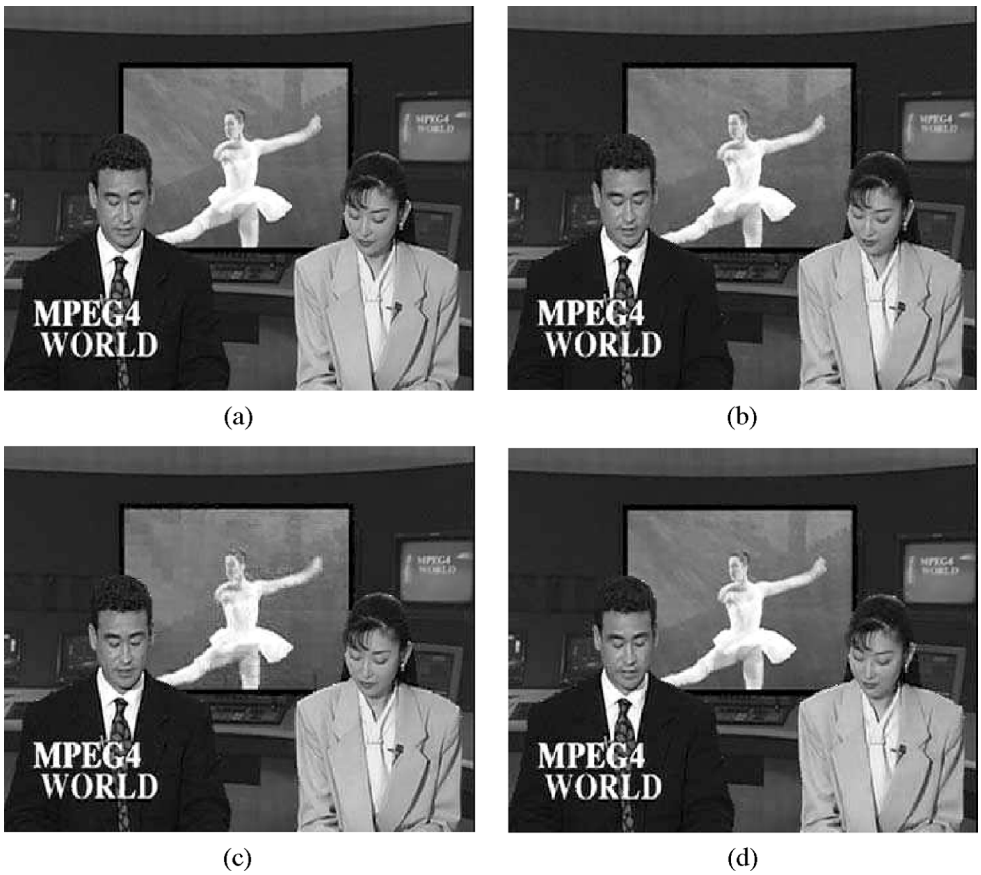


Fig. 7. Reconstructed frame 241 of test sequence *News*: (a) original, (b) verification model bit allocation (PSNR of VO2: 36.89 dB), (c) optimal bit allocation (PSNR of VO2: 31.65 dB), and (d) dynamic bit allocation (PSNR of VO2: 41.97 dB).

TABLE II
SIMULATION RESULTS OF *Akiyo*

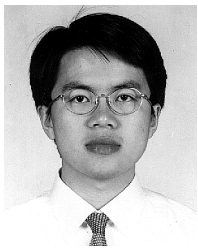
	PSNR (dB)		
	Foreground	Background	Overall
Verification Model Bit Allocation	37.03	43.61	39.32
Optimal Bit Allocation	37.10	**45.01**	**40.16**
Dynamic Bit Allocation	**37.83**	43.15	39.92

is significantly improved. These results demonstrate the usefulness of this dynamic bit allocation framework.

REFERENCES

[1] M. Kunt, "Second-generation image coding techniques," *Proc. IEEE*, vol. 73, pp. 549–574, Apr. 1985.
[2] H. Musmann, M. Hötter, and J. Ostermann, "Object-oriented analysis-synthesis coding of moving images," *Signal Process.: Image Commun.*, vol. 1, pp. 117–138, Oct. 1989.
[3] , ISO/IEC 14496-2:1999, "Information technology—coding of audio/visual objects," Part 2: Visual 1999.
[4] A. Vetro, H. Sun, and Y. Wang, "MPEG-4 rate control for multiple video objects," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, pp. 186–199, Feb. 1999.
[5] J. Ronda, M. Eckert, F. Jaureguizar, and N. Garcia, "Rate control and bit allocation for MPEG-4," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, pp. 1243–1258, Dec. 1999.
[6] H.-J. Lee, T. Chiang, and Y.-Q. Zhang, "Scalable rate control for MPEG-4 video," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, pp. 878–894, Sep. 2000.
[7] T. Chiang and Y.-Q. Zhang, "A new rate control scheme using quadratic rate-distortion modeling," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 1, pp. 246–250, Feb. 1997.
[8] A. Vetro, H. Sun, and Y. Wang, "Rate-distortion optimized video coding with frameskip," in *Proc. IEEE Int. Conf on Image Processing*, Thessaloniki, Greece, Oct. 2001, pp. 383–394.
[9] J.-W. Lee, A. Vetro, Y. Wang, and Y.-S. Ho, "Bit allocation for MPEG-4 video coding with spatio-temporal tradeoffs," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, pp. 488–502, Jun. 2003.
[10] H. Wang, G. M. Schuster, and A. K. Katsaggelos, "Rate-distortion optimal bit allocation scheme for object-based video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, pp. 1113–1123, Sep. 2005.
[11] Z. He and S. K. Mitra, "Optimum bit allocation and accurate rate control for video coding via ρ-domain source modeling," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 10, pp. 840–849, Oct. 2002.
[12] Z. Chen and K. N. Ngan, "Optimal bit allocation for MPEG-4 multiple video objects," in *IEEE Int. Conf. Image Processing*, Singapore, Oct. 2004.
[13] ——, "Rate-constrained arbitrarily shaped video object coding with object-based rate control," *Proc. Inst. Elect. Eng., Vision, Image Signal Process.*, vol. 151, pp. 250–256, Aug. 2004.
[14] ——, "Linear rate-distortion models for MPEG-4 shape coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, pp. 869–873, Jun. 2004.
[15] ——, "Joint texture-shape optimization for MPEG-4 multiple video objects," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 9, pp. 1170–1174, Sep. 2005.
[16] X. Yang and K. Ramchandran, "A low-complexity region-based video coder using backward morphological motion field segmentation," *IEEE Trans. Image Process.*, vol. 8, no. 3, pp. 332–345, Mar. 1999.
[17] Z. Wang, L. Lu, and A. C. Bovik, "Foveation scalable video coding with automatic fixation selection," *IEEE Trans. Image Process.*, vol. 12, no. 2, pp. 243–254, Feb. 2003.
[18] J. Malo, F. Ferri, J. Gutierrez, and I. Epifanio, "Importance of quantizer design compared to optimal multigrid motion estimation in video coding," *Eletron. Lett.*, vol. 36, pp. 807–909, Sep. 2000.

[19] J. Malo, J. Gutierrez, I. Epifanio, F. Ferri, and J. M. Artigas, "Perceptual feed-back in multigrid motion estimation using an improved DCT quantization," *IEEE Trans. Image Process.*, vol. 10, no. 10, pp. 1411–1427, Oct. 2001.
[20] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Trans. Image Process.*, vol. 13, no. 10, pp. 1304–1318, Oct. 2004.
[21] P. L. Silsbee, A. C. Bovik, and D. Chen, "Visual pattern image sequence coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 3, no. 3, pp. 291–301, Aug. 1993.
[22] S. Lee, M. S. Pattichis, and A. C. Bovik, "Foveated video compression with optimal rate control," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 7, pp. 977–992, Jul. 2001.
[23] S. Winkler, "Issues on vision modeling for video quality assessment," *Signal Process.*, vol. 78, pp. 231–252, Oct. 1998.
[24] A. B. Watson, J. Hu, and J. F. McGowan, III, "Digital video quality metric based on human vision," *J. Electron. Imag.*, vol. 10, pp. 20–29, Jan. 2001.
[25] A. B. Watson and J. Malo, "Video quality measures based on the standard spatial observer," in *Proc. IEEE Int. Conf. Image Processing*, Rochester, NY, Sep. 2002.
[26] Z. Wang, A. C. Bovik, and E. P. Simoncelli, "Objective video quality assessment," in *The Handbook of Video Databases Design and Applications*, B. Furht and O. Marqure, Eds. Boca Raton, FL: CRC, Sep. 2003, pp. 1041–1078.
[27] Sarnoff Corporation, JNDmetrix Technology [Online]. Available: http://www. sarnoff.com/products_services/video_vision/jndmetrix/
[28] Video quality experts group (VCEG) [Online]. Available: http://www.vqeg.org
[29] J. I. Ronda, M. Eckert, S. Rienke, F. Jaureguizar, and A. Pacheco, "Advanced rate control for MPEG-4 coders," in *Proc. Visual Commun. Image Processing*, San Jose, CA, Jan. 1998, pp. 383–394.
[30] H. Everett, "Generlized Lagrange multiplier method for solving problems of optimum allocation of resource," *Oper. Res.*, vol. 11, pp. 399–417, 1963.
[31] Y. Shoham and A. Gersho, "Efficient bit allocation for an arbitrary set of quantizers," *IEEE Trans. Acoust, Speech, Signal Process.*, vol. 36, no. 9, pp. 1445–1453, Sep. 1988.
[32] G. M. Schuster and A. K. Katsaggelos, *Rate-Distortion Based Video Compression*. Norwell, MA: Kluwer, 1997.
[33] A. Ortega, K. Ramchandran, and M. Vetterli, "Optimal trellis-based buffered compression and fast approximations," *IEEE Trans. Image Process.*, vol. 3, no. 1, pp. 26–40, Jan. 1994.
[34] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.
[35] J. Han and K. N. Ngan, "Automatic segmentation of objects of interest in video: a unified framework," in *Int. Symp. Intelligent Signal Processing and Communication Systems*, Hong Kong, Nov. 2004.
[36] T. Meier and K. N. Ngan, "Video segmentation for content-based coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 9, no. 12, pp. 1190–1203, Dec. 1999.
[37] S.-Y. Chien, Y.-W. Huang, and L.-G. Chen, "Predictive watershed: a fast watershed algorithm for video segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 5, pp. 453–461, May 2003.
[38] *MPEG-4 Video Verification Model v18.0*, ISO/IEC JTC1/SC29/WG11, Jan. 2001.
[39] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[40] X. Zhang and B. A. Wandell, "Color image fidelity metrics evaluated using image distortion maps," *Signal Process.*, vol. 70, pp. 201–214, Nov. 1998.



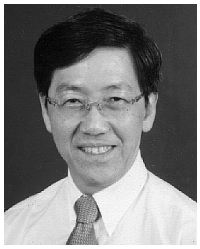
Zhenzhong Chen (S'02) received the B.E. degree from Huazhong University of Science and Technology, Wuhan, China, in 1999. He is currently pursuing the Ph.D. degree at the Department of Electronic Engineering, The Chinese University of Hong Kong (CUHK), Hong Kong.

From 2001 to 2003, he conducted graduate research at Nanyang Technological University, Singapore. His current research interests are visual signal processing and communications.

Mr. Chen is a student member of SPIE and the IEE. He is a recipient of a Microsoft fellowship.

Junwei Han received the Ph.D. degree from Northwestern Polytechnical University, China, in 2003.

He was a Research Associate with the Department of Electronic Engineering, The Chinese University of Hong Kong, in 2004. His research interests include image/video retrieval and image/video segmentation, and gesture recognition.



King Ngi Ngan (M'79-SM'01-F'00) received the Ph.D. degree in Electrical Engineering from Loughborough University of Technology, U.K. He is a Chair Professor in the Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong, and was previously a Full Professor in the Nanyang Technological University, Singapore, and the University of Western Australia, Australia. He has published extensively including three authored books, five edited volumes, and over 200 refereed technical papers in the areas of image/video coding and communications.

Prof. Ngan is an associate editor of the *Journal on Visual Communications and Image Representation*, as well as an area editor of *EURASIP Journal of Signal Processing: Image Communication*, and served as an associate editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY and the *Journal of Applied Signal Processing*. He chaired a number of prestigious international conferences on video signal processing and communications and served on the advisory and technical committees of numerous professional organizations. He is a Fellow of the IEE (U.K.) and the IEAust (Australia).