Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2007, Article ID 17179, 17 pages
doi:10.1155/2007/17179

Research Article

Content-Aware Video Adaptation under Low-Bitrate Constraint

Ming-Ho Hsiao, Yi-Wen Chen, Hua-Tsung Chen, Kuan-Hung Chou, and Suh-Yin Lee

College of Computer Science, National Chiao Tung University, 1001 Ta-Hsueh Road, Hsinchu 300, Taiwan

Received 1 September 2006; Revised 25 February 2007; Accepted 14 May 2007

Recommended by Yap-Peng Tan

With the development of wireless network and the improvement of mobile device capability, video streaming is more and more widespread in such an environment. Under the condition of limited resource and inherent constraints, appropriate video adaptations have become one of the most important and challenging issues in wireless multimedia applications. In this paper, we propose a novel content-aware video adaptation in order to effectively utilize resource and improve visual perceptual quality. First, the attention model is derived from analyzing the characteristics of brightness, location, motion vector, and energy features in compressed domain to reduce computation complexity. Then, through the integration of attention model, capability of client device and correlational statistic model, attractive regions of video scenes are derived. The information object- (IOB-) weighted rate distortion model is used for adjusting the bit allocation. Finally, the video adaptation scheme dynamically adjusts video bitstream in frame level and object level. Experimental results validate that the proposed scheme achieves better visual quality effectively and efficiently.

1. INTRODUCTION

With the development of wireless network and the improvement of mobile device capability, for mobile users, the desire to access videos is becoming stronger. More and more client users in a heterogeneous environment are desirous of universal access, that is, one can access any information over any network through a great diversity of client devices. Today, mobile devices including cellphone (smart phone), PDA, and laptop have enough computing capability to receive and display videos via wireless channels. However, due to some inherent constraints in wireless multimedia applications, such as the limitation of wireless bandwidth and high variation in device resource, how to appropriately utilize resource for universal access and to achieve high visual quality becomes an important issue.

Video adaptation is usually employed in response to the huge variation of resource constraints. In traditional video adaptation, the adapter considers the available bitrate and network buffer occupancy to adjust the data transmission while streaming video [1, 2]. Vetro et al. provided an overview of the video transcoding and introduced some transcoding schemes, such as bitrate reduction, spatial and temporal resolution reduction, and error resilient transcoding [3]. Chang and Vetro presented a general framework that defines the fundamental entities and important concepts related to video adaptation [4]. Furthermore, the authors indicated that most innovative and advanced open issues about video adaptation require joint consideration of adaptation with several other closely related issues, such as analysis of video content, understanding and modeling of users and environments. This work took video contents into consideration for video adaptation.

Much attention has focused on visual content adaptation [5]. Most traditional video communication systems consider videos as low-level bitstreams, ignoring the underlying visual content information. However, content analysis plays a critical role in developing effective solutions meeting unique resource constraints and user preferences under low-bitrate constraints. From the viewpoint of information theory, although the same bitrate delivers the same amount of information, it may not be true for human visual perception. Generally speaking, viewers can only be attracted and focused on a relatively small portion of a video frame. Hence, by adjusting different bit allocation to peripheral regions and regions-of-interest (ROI) of a frame, viewers can get better visual

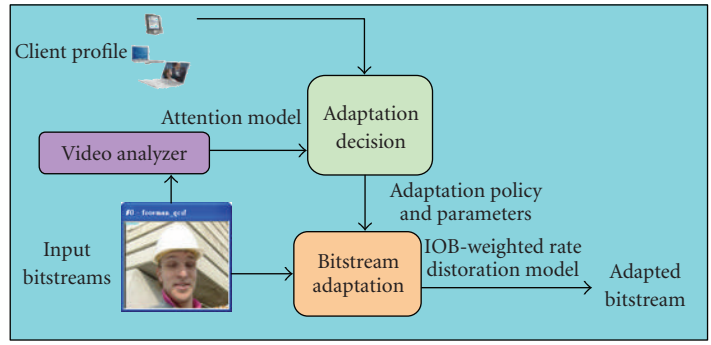


Figure 1: The architecture of the video adaptation system.

perceptual quality. In contrast to traditional video adaptation, content-based video adaptation can effectively utilize content information in bit allocation and in video adaptation.

In a content-aware framework for video communication, it is reasonable to assume videos belonging to the same class exhibit similar behaviors of resource requirements due to their similar features [6]. The comprehensive and high-level audio-visual features can be extracted from the compressed domains directly [7–9]. Low-level features like color, brightness, edge, texture, and motion are usually extracted for representing video content information [10]. Reference [11] presented a visual attention model based on motion, color, texture, face, and camera motion to simulate how viewers' attention are attracted based on analyzing low-level features of video content without fully semantic understanding of video content. Furthermore, different applications influence user preferences, while different contents cause various attention responses. The tradeoff between spatial quality (image clarity) and temporal quality (motion smoothness) under a limited bandwidth is considered to maximize user satisfaction in video streaming [5, 12]. Lai et al. proposed a content-based video streaming method based on visual attention model to efficiently utilize network bandwidth and achieve better subjective video quality [13]. Features like motion, color, texture, face, and camera motion are utilized to model the visual effects.

Attention is neurobiological conception [14]. It means the concentration of mentality on an attraction region in the content. Attention analysis breaks the problem of content object understanding into a computationally less demanding and a localized analytical problem. Thus, fast content analysis facilitates the decision making of video adaptation in adaptive content transmission.

Although there have been many approaches for adapting visual contents, most of them focus only on developing visual attention model in order to meet the bit-rate constraint and then to achieve high visual quality without considering the device capability. Hence the results may not be consistent with human perception due to excessive resolution reduction. The problem addressed in this paper is to utilize content information for improving the quality of a transmitted video bitstream subject to low-bitrate constraints, which especially applies to mobile devices in wireless network environment. Three major issues are concerned:

(1) how to quickly derive the important objects from a video?
(2) how to adapt video streams according to visual attention model and various mobile device capabilities?
(3) how to find an appropriate video adaptation approach to achieve better visual quality?

In this paper, a content-aware video adaptation mechanism is proposed based on visual attention model. Due to real time and low-bitrate constraints, we choose to derive content features from compressed domain to avoid expensive computation and time consumption involved in decoding and/or re-encoding. The content of video is first analyzed to derive the important regions which have high degree of attraction level. Then, bitrate allocation and adaptation assignment scheme is performed according to the content information in order to achieve better visual quality and avoid unnecessary resource waste under low-bitrate constraint. Finally, we will analyze the issues related to device capabilities through theory and experiments and thereupon present a system to deal with it.

The rest of this paper is organized as follows. Section 2 presents an overview of the proposed scheme. A novel video content analyzer is presented in Section 3 and a hybrid feature-based model for video content adaptation decision is illustrated in Section 4. In Section 5, we describe the proposed bitstream adaptation approaches. The experimental results and discussion will be presented in Section 6. Finally, we conclude the paper and describe the future works in Section 7.

2. OVERVIEW OF THE VIDEO ADAPTATION SCHEME

In this section, we introduce the overview of the proposed content-aware video adaptation scheme, as shown in Figure 1. Initially, video streams are processed by video analyzer to derive the content features of each frame/GOP and then to obtain the important regions with high attraction. Subsequently, the adaptation decision engine determines the adaptation policy according to the attention model derived

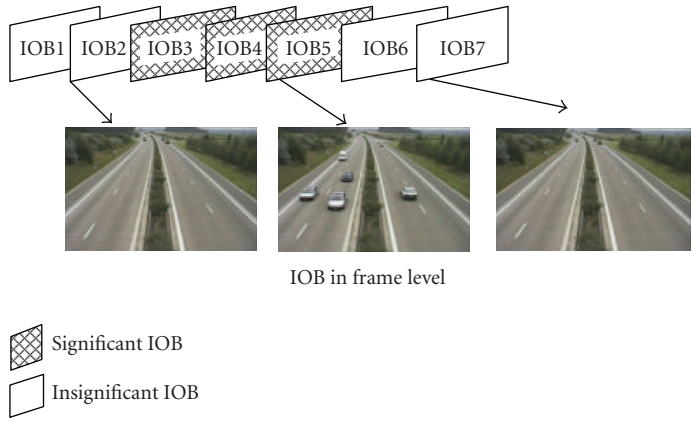


Figure 2: An example of content attention model.

from video analyzer. Besides, the device capability obtained from client profile, the correlational statistic model, and the region-weighted rate distortion model [13] will be applied to adapt video bitstrem at the same time. Finally, the bitstream adaptation engine adapts video based on adaptation parameters and IOB-weighted rate distortion model.

3. VIDEO ANALYZER

In this section, we describe the video analyzer which is used to analyze the features of video content for deriving meaningful information. Section 3.1 describes the input data we use for video analyzer. In Section 3.2, we import the concept of Information object to model user attention. Finally, we introduce the relation between the extracted features and visual perception effects in Section 3.3.

3.1. Data extraction

The features are extracted from the coded stream in compressed domain, which is computationally less demanding, in order to meet the real-time requirement of the application scenario. The DC and AC coefficients of the DCT transformed blocks represent the illumination and texture in the corresponding blocks. The motion vectors are also extracted for describing the motion information of the frames.

Since the DC and AC coefficients in P or B frames are resulted from DCT transformation of residuals, they provide less semantic description of the video data than those in I frames. Therefore, in this paper, we choose to extract the DC and AC coefficients in I frames only. Moreover, the content of B frames is similar in general to the neighboring I or P frames due to the characteristics of temporal coherence. Thus, we drop the extraction of motion information in B frames to speed up the computation of data extraction.

To sum up the procedure of data extraction, we choose the DC and AC values of I frames plus motion magnitudes and motion directions of P frames as input data of the video analyzer. These input data can be easily extracted from compressed video sequences. The relations and visual effect of extracted features including brightness, color, edge, energy, and motion will be further described in Section 3.3.

3.2. Information object (IOB) derivation

Different parts of video contents have different attraction values for user perception. Attention-based selection [14] allows only attention-catching parts to be presented to the user without affecting much user experience. For example, human faces in a photo are usually more important than the other parts. A piece of media content P usually consists of several information objects IOB_i. An information object is an information carrier that delivers the author's intention and catches the user's attention as a whole. We import the "information object" concept, which is a modification of [14] to agree with video content, defined as below.

Definition 1. The basic content attention model for a video shot S is defined as a set which has two related hierarchical levels of information objects:

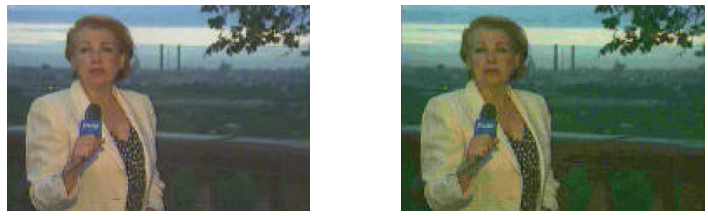
$$\begin{aligned} S &= \{\mathrm{HIO}_i\}, \quad 1 \le i \le 2, \\ \mathrm{HIO}_i &= \{\mathrm{IOB}_j, \mathrm{IMP}_j\}, \quad 1 \le j \le N_i, \end{aligned} \tag{1}$$

where HIO_i is the perception of frame or object level of S, respectively, IOB_j is the jth information object in HIO_i of S, IMP_j, is the importance attraction value (**IMP**) of IOB_j, and N_i, is total number of information objects in HIO_i of S.

Figure 2 gives an example of content attention model consisted of some information objects in different levels. The information objects generated by content analyzer are basic units for video adaptation.

3.3. Feature selection for visual attention

By analyzing a video content, we can extract many visual features (including brightness, spatial location, motion, and energy) that can be used to generate a visual attention model. In the following, we discuss the extraction methods, visual perceptive effect, and possible limitation for each feature. Some



(a) Original frame

(b) An adapted frame using uniform quantization parameter

Figure 3: Perceptual distortion comparison between different brightness.

features might be meaningless for some kinds of videos, such as motion feature for rather smooth scenes or no motion videos.

Brightness

Generally speaking, the human perception is attracted by the brighter part. For example, the brightly colored or strongly contrasted parts within a video frame always have high attraction, even those in the background. Integrating the preceding analysis with an observation in Figure 3, even the same bitrate is assigned, the visual distortion of the dark regions is usually more unobvious. Chou et al. mentioned that visual distortion of regions in the midgrey level close to the midgrey luminance is more obvious than in the brighter and darker regions [15, 16]. Therefore, the brightness characteristic is an important feature to identify the information Objects for visual attention.

Consequently, for each block the importance value of the proposed brightness attention model containing mean of brightness and variance of brightness is presented in the ollowing:

$$\mathrm{IMP_{BR}} = \frac{\mathrm{DCvalue} \times \mathrm{BR\_weight}}{\mathrm{BR\_level}} \times \mathrm{BR\_var}, \tag{2}$$

where DCvalue is the DC value of luminance for each block, BR_level is obtained from the average luminance of the previous frame, BR_var denotes the DCvalue variance of current and neighboring eight blocks, and BR_weight is assigned according to the error visibility threshold presented in [15]. When the luminance is close to midgrey (127), the weight is higher to reduce visual distortion [15]. Moreover, in order to reduce the computing time, weight can be assigned as follows:

$$\mathrm{BR\_weight} = \begin{cases} 2^0, & \text{if DCValue} < 64, \\ 2^2, & \text{if } 64 \le \text{DCValue} \le 196, \\ 2^1, & \text{if } 196 < \text{DCValue}. \end{cases} \tag{3}$$

In order to further normalize the brightness attention values of different video content, we use $\mathrm{IMP_{BR}}$ value of each block to represent the brightness attention histogram. We divided the brightness attention histogram into **L** levels and then assigned them the value from 1 to **L** (here, **L** = 5), respectively.

However, the brightness attraction property may lose its reliability when the overall frame/scene has higher brightness. As illustrated in the first row of Figure 4, the IOBs presented with yellow mask suffuse the overall fame so that we cannot distinguish which regions are more attracted, if we just use the DC values of the luminance of I frames to derive the brightness of blocks. Moreover, in some special cases, the regions with large brightness value do not cause human attention, such as the scene containing the white wall background, the cloudy sky, the vivid grasslands.

In order to improve the brightness attention model in response to attraction, we design a location-based brightness distribution histogram (lbbh) which utilizes the correlation between brightness distribution and position to identify the important brightness bin and roughly discriminate foreground from background. In Figure 5(a), the blocks near central regions of a frame are assigned high region value and they are considered as foreground IOBs. We use DC value of each block to represent the brightness histogram. The brightness histogram of each frame is computed while the region value of the block is also recoded at the same time. Then, for each bin, the average or the majority of (block) region values is computed to indicate the representative region value (location) of that bin. This is called the location-based brightness histogram as shown in Figure 5(b). The approach calculates mainly average region value of each bin of brightness distribution to decide whether the degree of brightness is attractive. For instance, the same brightness distributed over center regions or peripheral regions will cause different degree of attention, even if they both are quite bright.

We apply the location-based brightness histogram to adjust the attention model of brightness. After obtaining $\mathrm{IMP_{BR}}$ value from (2) and (3), we adjust the $\mathrm{IMP_{BR}}$ depending on whether the proportion of the brightness bin is greater than a certain degree or not. The function of adjustment is as follows:

$$\mathrm{IMP_{BR'}} = \begin{cases} 0, & \text{if lbbh}(bi) \le 1, \\ \mathrm{IMP_{BR}} - 1, & \text{if } 1 < \text{lbbh}(bi) \le 2, \\ \mathrm{IMP_{BR}}, & \text{if } 2 < \text{lbbh}(bi) \le 3, \\ \mathrm{IMP_{BR}} + 1, & \text{if } 3 < \text{lbbh}(bi) \le 4, \\ 5, & \text{if } 4 < \text{lbbh}(bi). \end{cases} \tag{4}$$

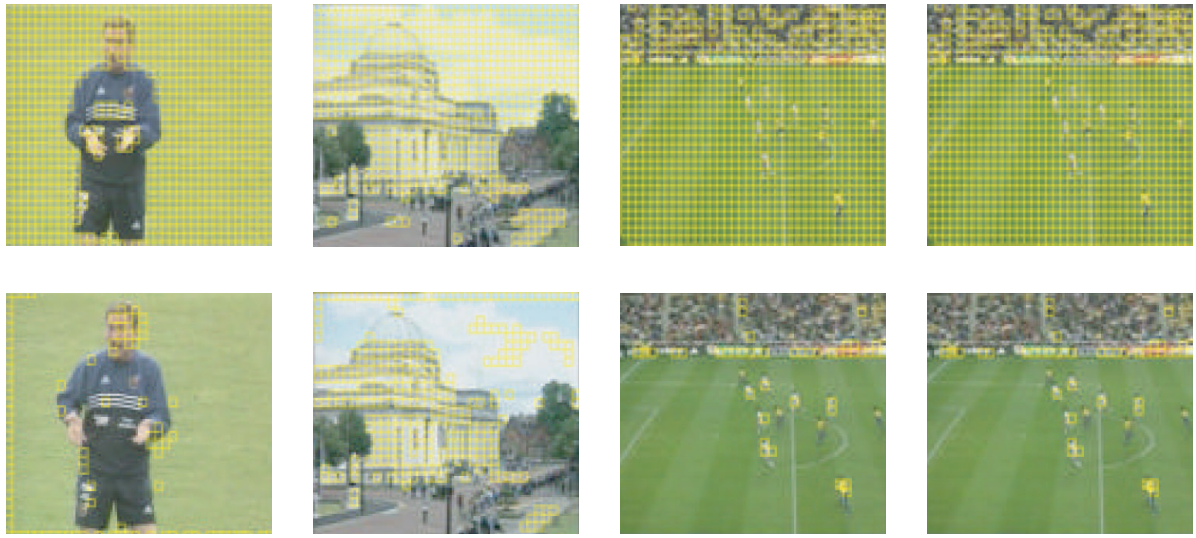
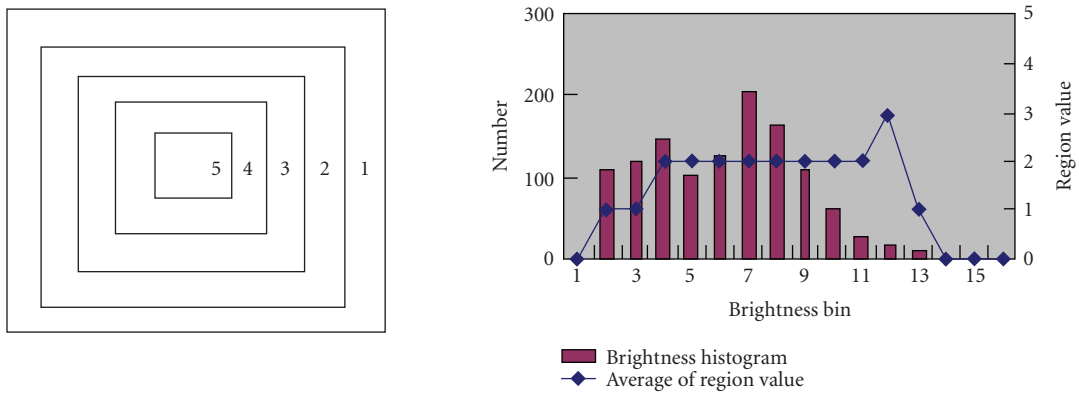


FIGURE 4: IOBs derived from brightness without (first row) and with combining the location-based brightness histogram (second row).



(a) The centricity region and weight used to estimate the distribution of brightness bin

(b) An example of location-based brightness histogram

FIGURE 5: Location-based brightness histogram.

$IMP_{BR'}$ is the adjusted brightness attention value using location based brightness histogram model. The lbbh*bi* denotes the region value of block *bi* derived from the location in its brightness distribution bin in the range $[1 \sim 5]$. For each bin, if the average region value of the blocks falling in to this bin is close to the centricity region value, the weight assigned to those blocks is higher to increase the importance. In Figure 5(b), the IMP value of blocks whose luminance fall into bin 12 will be assigned higher weight than others because bin 12 has the larger region value 3. As a result, those blocks assigned large IMP values will be considered as important IOBs.

We can evidently discover that the IOBs derived from (4) really attract human visual perception as shown in the second row of Figure 4. Hence, the adjusted results of IOBs employing the location-based characteristic have better refinement against pure brightness attention model.

Location

Human usually pay more attention to the region near the center of a frame, referred to as location attraction property. On the other hand, the cameramen usually operate the camera to focus on the main object, that is, put the primary object on the center of the camera view, in the technique of photography. So, the closer to the center the object is, the more important the object might be. Even the same objects may have different important values depending on their location

1	1	1	1	1
1	2	2	2	1
1	2	4	2	1
1	2	2	2	1
1	1	1	1	1

Figure 6: Location weighting map and adapted video according to the location feature.

Table 1: The video types are classified according to motion vector.

			Motion magnitude		Zero motion (%)	Maximum motion direction proportion
Class	Camera	Object	Mean	Variance		
1	Fixed	Static	Near 0 (M1 = 0.1)	Quite small (V1 = 1.5)	Near 95%	—
2	Fixed	Moving	Small (M2 = 2)	Smaller (V2 = 5)	Medium (> 40%)	—
3	Moving	Static	Larger	Midium/large	Small	Quite large (> 0.33)
4	Moving	Moving	Larger	Larger	Small	Smaller

of appearance. To get better subjective perceptual quality, the frames can be generated adaptively by emphasizing the regions near the important location and deemphasizing the rest regions. The location-related information can be generated automatically according to the centricity.

We introduce a weighting map in accordance with centricity to reflect the location characteristic. Figure 6 illustrates the weighting map and an adapted frame example based on the location. However, for different types of videos, the centricity of attraction may be different. A dynamic adjustment of location weighting map will be introduced in Section 4.3 according to the statistical information of IOB distribution.

Motion

After extensive observation of a variety of video shots in our experiments, the relation between the camera operation and the object behavior in a scene can be classified into four classes. The first class, the camera is fixed and all the objects in the scene are static, such as partial shots of documentary or commercial scenes. The percentage of this type of shots is about 10 $\sim$ 15%. The second class is fixed camera and some objects are moving in the scene, like anchor person shots in the news, interview shots in the movie, and surveillance video. This type of shots is about 20 $\sim$ 30%. The third class, the camera moves while no change in the scene, is about 30 $\sim$ 40%. For instance, some shots of scenery scene belong to this type. The fourth class, the camera is moving while some objects are moving in the scene, such as object tracking shots. The proportion of this class is also about 30 $\sim$ 40%.

Because the meaning and the importance degree of the motion feature are dissimilar in the four classes, it is beneficial to first determine what class a shot belongs to while we derive information objects. We can utilize the motion vector field to distinguish the target video shot into applicable class. In the first class, all motion vectors are almost zero motions because the adjacent frames are almost the same. In the second class, there are partial zero motions due to the fixed camera and partial similar motion patterns attributed to moving objects, so that the average and the variance of motion magnitude are small and there is a certain proportion of zero motion.

In the third class, all motions have similar motion patterns when the camera moves along the XY-plane or Z-axis, while the magnitudes of motions may have larger variance in other cases of camera motion. The major direction of motion vectors also has a rather large proportion in this class. In the fourth class, the overall motions may have large variation while some regions belonging to the same object have similar motion patterns.

Generally speaking, the mean and variance of motion magnitudes in the cases of moving camera are larger than those in fixed camera motion. Besides, the motion variances in the fourth class are larger than the variances in the third class due to the moving objects mixed with camera moving resulting in difference motion patterns. However, in the fourth class the motion variance may be not larger than that in the third class if moving objects are small sized. The motion magnitude only might not be a good criterion to distinguish between the third and fourth classes. We can observe that the major direction of motion vectors has a rather large proportion in third class because almost all the motions have similar motion direction following the moving camera. Hence, we can utilize the maximum motion direction proportion to distinguish the two video classes in the cases of moving camera. If the proportion is larger than the predefined threshold (say 30%), the video type belongs to the third class.

According to the above discussion, we use the mean of motion magnitude, the variance of motion magnitude, the proportion of zero motion, and the histogram of motion direction to determine the video type, as shown in Table 1. M1, M2, V1, and V2 are thresholds for classification and

are described in Section 5.1. More than 80% of test video sequences can be correctly classified into their motion class by our proposed motion class model. Because the P frames of the first GOP sometimes use intracoding mode, that is, no motion vector, the accuracy of motion class in the first GOP is lower than others. Therefore, we adjust the adapting scheme after the first GOP in our video adaptation mechanism.

People usually pay more attention to large motion objects or objects which have distinct motion activity from others, referred to as motion attraction property. Besides, motion feature has different importance degree and different meaning according to its motion class. So, our motion attention model will depend on the above mentioned motion classes and is illustrated as below.

In motion classes 1 and 2,

$$\mathrm{IMP}_{\mathrm{MAtt}} = \frac{\mathrm{MV\_magnitude}}{\tau - \lambda} \quad \text{when } \tau \geq \mathrm{MV\_magnitude} \geq \lambda. \tag{5}$$

In motion classes 3 and 4,

$$\mathrm{IMP}_{\mathrm{MAtt}} = \frac{\mathrm{MV\_magnitude}}{\tau - \lambda} \times \left(\frac{|\mathrm{MV\_ang} - \mathrm{DMV\_ang}|}{\mathrm{DMV\_ang}} \right) \quad \text{when } \tau \geq \mathrm{MV\_magnitude} \geq \lambda, \tag{6}$$

where $\mathrm{IMP}_{\mathrm{MAtt}}$ is the motion attention value for each block of P frame, MV_magnitude denotes motion magnitude, MV_ang represents motion angle, DMV_ang represents the dominate motion angle, and τ, λ are the two dynamic thresholds for noise elimination and normalization accounting for different video content. τ and λ adopted are the maximum and the minimum motion magnitude in our model, respectively.

For each block of a video frame, we calculate the histogram of the motion angle. The MA represents the bin proportion of the motion angle distribution histogram for each block. In this paper, we use 30 degrees as a bin, and then the histogram (distribution) can be obtained. The MAs of each block can be computed as the ratio of bin value to the sum of all bin values. Then the motion angle of maximum MA can be treated as the DMV_ang to compute the correct $\mathrm{IMP}_{\mathrm{MAtt}}$ value of moving objects in the motion classes 3 and 4, because camera motion should be taken into consideration to compensate the motion magnitude for the global motion. In (6), the $\mathrm{IMP}_{\mathrm{MAtt}}$ value of each block can be calculated to acquire motion magnitude to further identify the attention value. If the motion angles of blocks are close to the DMA_angle, those blocks are assigned low attention value and they are considered as background IOBs.

Energy

Another factor that influences perceptual attention is the texture complexity, that is, the distribution of edges. People usually pay more attention to the objects which have larger or smaller magnitude of edge than average [17], referred to as energy attraction property. For example, an object with complicated texture in smooth scene is more attractive, and vice versa. We use the predefined two edge features of the AC coefficients in DCT transformed domain [9, 18] to extract edges. The two horizontal and vertical edge features can be formed by two-dimensional DCT of a block [19],

$$\begin{aligned} \text{Horizontal Feature} &: H = \{H_i : i = 1, 2, \ldots, 7, \} \\ \text{Vertical Feature} &: V = \{V_j : j = 1, 2, \ldots, 7\} \end{aligned} \tag{7}$$

in which H_i and V_j correspond to the DCT coefficients $F_{u,0}$ and $F_{0,v}$ for $u, v = 1, 2, \ldots, 7$. Equation (8) describes the AC coefficients of DCT:

$$F_{u,v} = \frac{2}{\sqrt{MN}} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} x_{i,j} \cos \frac{(2i+1)u\pi}{2M} \cos \frac{(2j+1)v\pi}{2N}, \tag{8}$$

where $u = 1, 2, \ldots, M-1$, and $v = 1, 2, \ldots, N-1$. Here $M = N = 8$ for an 8×8 block.

In the DCT domain, the edge pattern of a block can be characterized with only one edge component, which is represented by projecting components in the vertical and horizontal directions, respectively. The gradient energy of each block is computed as

$$E = \sqrt{H^2 + V^2}, \tag{9a}$$

$$H = \sum_{i=1}^{7} |H_i|, \qquad V = \sum_{j=1}^{7} |V_j|. \tag{9b}$$

The gradient energy of I frame represents the edge energy feature.

However, the influence of perceptual distortion with large edge energy or small edge energy is not so significant. As shown in Figure 7, we can discover that high-energy regions

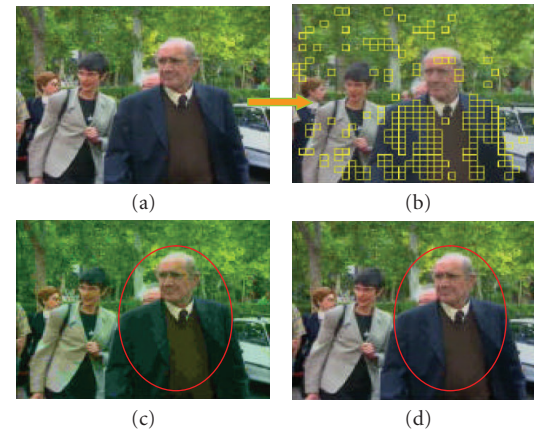


Figure 7: Comparison of the visual distortion in different edge energy regions. (a) The original frame. (b) The IOBs are derived from energy. (c) The uniform quantization frame. (d) The energy adapted frame.

like tree have less visual distortion than other regions like walking person in Figure 7(b) under the uniform quantization constraint. In other words, the visual perceptual distortion introduced by quantization is small in extremely high- or low-energy cases.

Our energy model which integrates the above two aspects is illustrated as below. According to the energy E obtained from (9a), each block is assigned the energy attention value, as shown in Figure 8. Because the energy distribution of each video frame is different, the energy of a block may be higher in some frames, but lower in other frames. We use the ratio of the block energy to average energy of a frame to dynamically determine the importance value. When E is close to the energy mean of a frame, we assign a medium energy attention value to the block. When E belongs to higher-energy (or lower) regions, we assign a high-energy attention value to the block. In extreme energy cases, we assign the lowest-energy attention value to such blocks because their visual distortion is unobvious. The IMP of energy attention model, $\mathrm{IMP}_{\mathrm{AE}_i}$, of the block i in the frame j is coputed as

$$\mathrm{IMP}_{\mathrm{AE}_i} = \begin{cases} 1, & \text{if } \dfrac{E_i}{E\_\mathrm{mean}} > \dfrac{E\_\mathrm{Max}}{E\_\mathrm{mean}} \times Eb + (1 - Eb) \\ & \text{or } \dfrac{E_i}{E\_\mathrm{mean}} < \dfrac{E\_\mathrm{Min}}{E\_\mathrm{mean}} \times Eb + (1 - Eb), \\ 2, & \text{if } \dfrac{E_i}{E\_\mathrm{mean}} < \dfrac{E\_\mathrm{Max}}{E\_\mathrm{mean}} \times Ea + (1 - Ea) \\ & \text{or } \dfrac{E_i}{E\_\mathrm{mean}} > \dfrac{E\_\mathrm{Min}}{E\_\mathrm{mean}} \times Ea + (1 - Ea), \\ 4, & \text{otherwise,} \end{cases} \tag{10}$$

where Ei is the energy of block i, and $E_$Max, $E_$Min, and $E_$mean are the maximum block energy, the minimum block energy, and the average energy of frame j, respectively. Ea and Eb are two parameters used to dynamically control the weight assignment. If the ratio of block energy Ei to $E_$mean is higher than $Ea \times (E\_\mathrm{Max}/E\_\mathrm{mean})$ and lower than $Eb \times (E\_\mathrm{Max}/E\_\mathrm{mean})$, the weight will be 4. Ea and Eb are derived from the result of training video shots, and are set to be 0.6 and 0.8, respectively. According to the IOBs derived from the energy attention model as shown in Figure 7(b), we can observe that the energy-adapted frame in Figure 7(d) achieves better visual quality than the uniform quantization frame.

4. ADAPTATION DECISION

Adaptation decision engine is used to determine video adaptation scheme and adaptation parameters for subsequent Bitstream adaptation engine to obtain better visual quality. We describe the adaptation approaches and decision principle according to the video content in Section 4.1, while we present device capability-related adaptation in Section 4.2. In Section 4.3, we propose the concept of correlational statistic model to improve the content-aware video adaptation system.

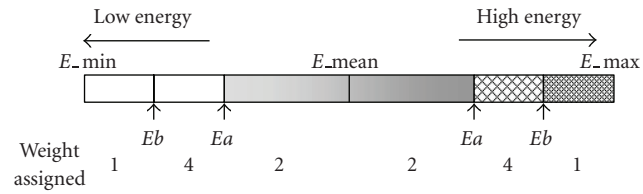


Figure 8: The energy attention model.

Table 2: The importance of feature for obtaining IOBs in different video classes.

Class	Camera	Object	Brightness	Location	Motion	Energy
1	Fixed	Static	◎	◎	—	◎
2	Fixed	Moving	—	—	◎	—
3	Moving	Static	◎	◎	—	◎
4	Moving	Moving	◎	◎	◎	◎

4.1. Content

Our content-related adaptation decision is based on the extracted features and the attention models discussed in the Section 3. We utilize brightness, location, motion, and energy features to derive the information objects of video content. A lot of factors affect human perception. We adopt integration model to aggregate attention values from each feature, instead of intersection model. One object gaining quite high score in one feature may attract viewers while another object gaining medium high score in several features may also attract viewers. For example, a quite high-speed car appearing in a scene will attract viewers' attention, while a brightly, slowly walking person appearing in the center of a screen also attracts the sight of views.

In addition, due to vast variety in video content, the decision principle for adaptation scheme must be adjustable according to the content information. We utilize the feature characteristics to roughly discriminate content into several classes. In our opinions, the motion class is a good classification to determine the weight of each feature in the information object derivation process. Table 2 shows the details of the selected features to compute important value of IOBs in each motion class. In the first class, due to the motions being almost zero motions, we do not need to consider the motion factor. In the second class, the motion is the dominant feature because the moving objects are especially attractive in this class. Although the selected features for obtaining IOBs in third class are the same as the first class, the adaptation schemes are entirely different. In the first class, the frame rate can be reduced considerably without introducing the motion jitter. Nevertheless, whether the frame rate can be reduced in the third class depends on the speed of the camera motion. The features in the attraction of viewer's attention are not practically distinguishable in the fourth class. Hence, all the features are adopted to derive the information objects of video content.

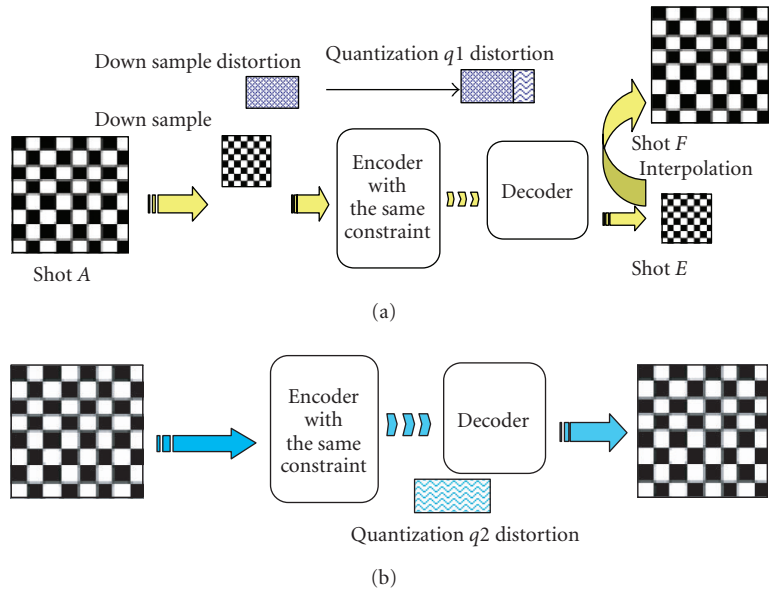


Figure 9: The above process (a) is the resolution-considered adaptation. The below process (b) is the original encoding process.

4.2. Device capability

In order to reduce the unnecessary waste and increase the utilization of resource, it is essential to consider the device capability in adapting video. Especially, as a great amount of new devices with diverse capabilities are making a popular boom, their limited resolution, available bandwidth, weaker display support, and relatively powerless computation are still obstacles to streaming video even in traditional environments. Without appropriately adapting video, the resource cannot be efficiently utilized and the received visual quality may be quite poor.

In our video adaptation scheme related to client device capability, we consider the spatial resolution, color depth, brightness, and computation power of the receiving device. In the following, we will describe the adjusting methods in different aspects.

Spatial resolution

In hand-held devices, there is one common characteristic or shortcoming, small resolution. If we transmit a higher-resolution video, like 320×240, to a lower-resolution device, like 240×180, it is easy to understand that much unnecessary resource is wasted with quite little quality gain or just the same quality. Besides, picture resolutions of video streams need not be equal to the screen resolutions of multimedia devices [20]. When the device resolution is larger than the video resolution, the device can easily zoom the pictures by interpolation. Under the same bitrate constraint, higher-resolution video streams certainly need to use larger quantization parameter, and smaller resolution video streams naturally can use smaller quantization parameter. Actually, it is a tradeoff between picture resolution and quantization precision. Reference [20] concluded that appropriately lowering picture resolution combined with decent interpolation algorithms can achieve better subjective quality in a target bitrate. However, their proposed tradeoff principle used to determine the appropriate picture resolution is heuristic and computation-intensive, which requires preencoding attempt.

As to the issue of how to adjust the video resolution properly accommodating the device resolution under various bitrate constraints, some experiments related to the determination of appropriate resolution are presented and described below. In the simulation, the video sequences were MPEG-2 encoded, the resolution is 320×240, and the device resolution is 240×180. We observe the video quality of different resolutions and various bitrates under the same constraint. Due to the dissimilar behavior in different bitrate environments, the bandwidth constraint in the experiments varies from high to very low, that is, 1152 kbps to 52 kbps. The resolution varies from original (320×240) to 80×60.

The process of Figure 9(a) is the resolution-considered adaptation. The process of Figure 9(b) is the original encoding process. Under the same bitrate constraint, the quantization step of process Figure 9(b) is much larger than that of process Figure 9(a). In Figure 9, we can find that the distortion introduced by down sampling, encoding quantization, and interpolation is smaller than that introduced just by encoding quantization under the same bitrate constraint.

As to the influence of device capability, we discuss the tradeoff between the appropriate picture resolutions and quantization precision. The PSNR is most commonly used as a measure of quality of reconstruction in compression. However, the device capability is not considered during the computation of traditional PSNR. Since in PSNR the same resolution is considered, hence we modify the definition of PSNR to reasonably reflect the objective quality accommodating the device capability by linear interpolation before imitating the PSNR, which is referred to as MPSNR. MPSNR is

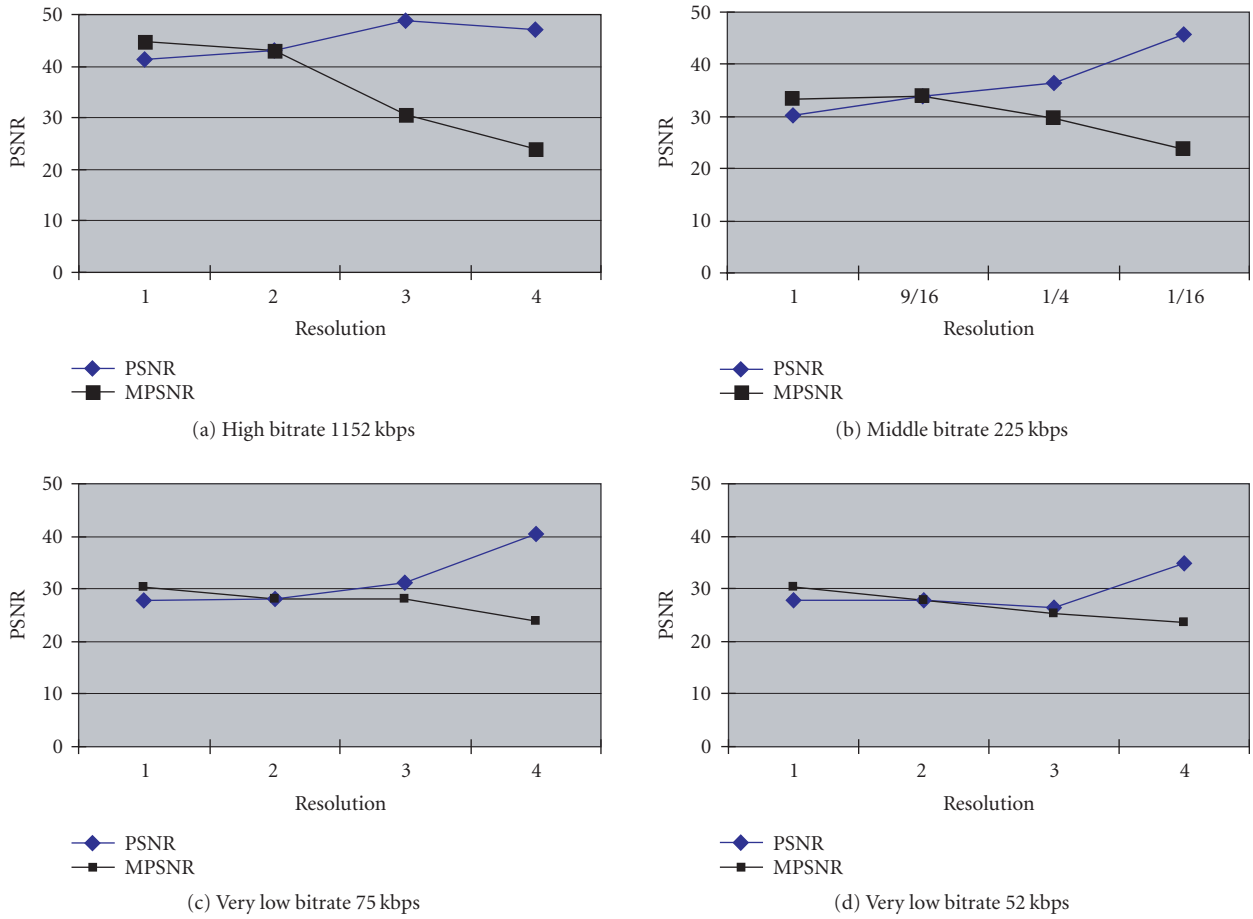


Figure 10: Comparison of PSNR and MPSNR in various bitrates. The x-axis is the percentage of original video resolution.

proposed to measure the quality of reconstruction video shot accommodating the device resolution.

For example, assume the resolution of an original shot (Shot A) in Figure 9(a) is 320×240 and device resolution is 240×180. If the resolution of the encoded shot E is 80×60 as shown in Figure 9(a), then the shot E needs to be upsampled from 80×60 to 320×240 when we measure the PSNR of the shot E constructed. In addition, if we want to calculate the MPSNR of the shot E, we need to interpolate the downsampled shot E of resolution 80×60 to the interpolated shot F in Figure 9(a) of the device resolution 240×180. Then the resolution of the original shot needs to be adjusted from 320×240 to 240×180. The PSNR between the constructed shot A and interpolated shot F in the resolution of display is called MPSNR.

For objective quality, The PSNR and MPSNR values are measured to compare the distortion in various bitrate constraints, as illustrated in Figure 10. The resolution of original shot is 320×240 and device resolution is 240×180. In order to validate the effectiveness of MPSNR, the encoded resolutions of the original shot are 320×240, 240×180, 160×120, and 80×60 in various bitrates, respectively. From the experimental results measured in MPSNR instead of PSNR, we can verify that reducing the video resolution to device resolution or to 1/4 device resolution while increasing quantization precision will achieve better visual quality in low bitrate, such as 75 to 100 kbps.

The idea which utilizes the downsampling approach in device-aware video adaptation as illustrated in Figure 9 is beneficial to obtain better visual quality. It can be observed that the visual quality of Figure 11(b) is better than that of Figure 11(a), which validates the effectiveness of the approach.

Color depth and brightness

The reason for considering the color depth of the device capability is similar to the spatial resolution. Some hand-held devices may not support full color depth, that is, eight bits for each component of color space. To avoid unnecessary resource waste, we may utilize the color depth information of the device in video adaptation. For example, it is necessary to avoid transmitting video streams with 24-bit color depth to the device with only 16-bit color depth. The effect of reducing the color depth is similar to quantization. Therefore, the rate controller will choose higher quantization parameter when the device supports less color depth.

Furthermore, visual perception is not so sensitive to the variation in very bright (or very dark) regions and some restrictions are inherent in hand-held device display screens, such as low brightness contrast. It is reasonable to remove the extreme value without influencing viewers' experience. Although the improvement in the utilization of resource based on extreme brightness removing property is pretty limited, the entropy is reduced without perceived distortion during encoding process.

Computation

Due to the weak computation capability of mobile devices, there may be not enough time to decode and display video at the frame rate defined at the encoder. Appropriately reducing the frame rate transmitted not only can avoid the asynchronous problem but also can exploit the bitrate saving in spatial quality. Another advantage is to extend the duration of power on while reducing the frame rate for transmission. In contrast to general computer, mobile devices have a significant difference, that is, the power source is limited. If we reduce the temporal resolution, the power consumption will slow down owing to the reduction in the computation for receiving and decoding.

4.3. Correlational statistic model

Due to the high correlation between adjacent frames, we apply the correlational statistic model to explore the relation between frames. The location which has higher-importance attention value in the preceding frames will have higher probability of being information object in the current frame. Similarly, the frame of which the preceding frames have higher-importance value will have higher probability of being information object in frame level. In order to reduce the computation, we analyze the interdependence of information objects in spatial and temporal domain.

From this observation, we can predict information objects utilizing the spatial/temporal information when the motions of a shot are small. In the proposed correlational statistic model, the spatial/temporal information including the statistics of each features, IOB distribution, IOB density, and motion class of preceding frames is adopted to determine the weight of each content feature in the evaluation of IMP value in the current frame in order to adjust adaptation decision. In the following, we will describe the purpose and the effect of the statistics of information objects.

Observing the dispersedness of information objects in a frame, we can discover large variation in distinct videos. For instance, the density and centricity of information objects in Figures 12(a) and 12(b) are eminently different. We adopt the centricity region of Figure 5(a) to calculate the percentage of information objects in each region as shown in Figure 13. Therefore, we will dynamically adjust the weighting map of the location feature according to the statistics of the IOB density. The five candidates of location weight map are demonstrated in Figure 14(a) and some examples of weight location maps are in Figure 14(b). When the IOBs are centralized, the candidate location map at the bottom right of Figure 14(a) is used. On the contrary, the candidate location map at the top left of Figure 14(a) is used.

(a) (b)

Figure 11: Comparison of visual quality in very low bitrate constraint, that is, 75 kbps. (a) The adapted result in 240 × 180 resolution. (b) The adapted result in 180 × 120 resolution.

(a) video c41 (b) video c162

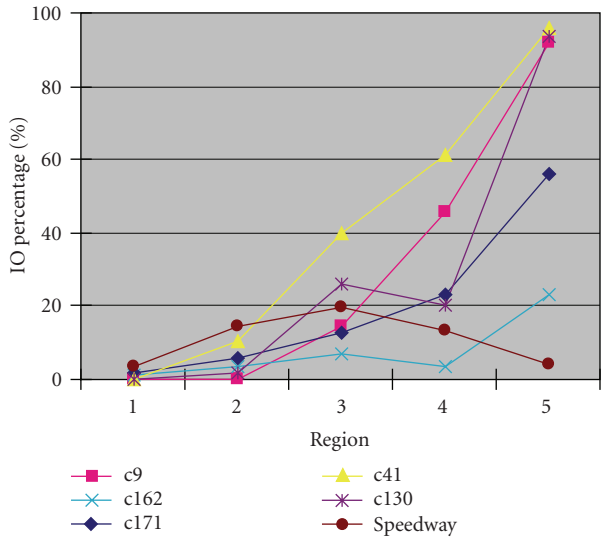
Figure 12: The IOBs are marked as yellow masks in distinct videos.

5. BITSTREAM ADAPTATION

In this section, we present the dynamic bit allocation framework for bitstream adaptation. Bitstream adaptation engine controls the bitrate and adapts the bitstream based on adaptation policy and parameters obtained from adaptation decision engine. Under the bitrate constraint, the optimized quantization parameters for achieving better visual quality are then obtained to encode the IOBs separately. In Section 5.1, we present bit allocation scheme of the proposed content-aware adaptation. Subsequently, the concept of IOB-weighted rate distortion model used to execute rate control is introduced in Section 5.2.

5.1. Bit allocation scheme

Based on the attention analysis results, we establish a content-aware video adaptation model. When the bandwidth is insufficient for the transmission of original full quality video stream, the adaptation system must have an efficient mechanism to modulate videos following certain principles, such as high resource utilization, better temporal quality, better spatial quality, and/or low computation complexity. In the proposed allocation scheme, we incorporate two major principles: improve visual perceptual quality and avoid unnecessary resource waste.



Video name	Video content
c9	Anchor person
c162	Full court football
c171	Close-up football
c41	Dense objects
c130	Large object
speedway	Speed way

Figure 13: The relation between densities of IOB and centricity regions as shown in Figure 5(a).

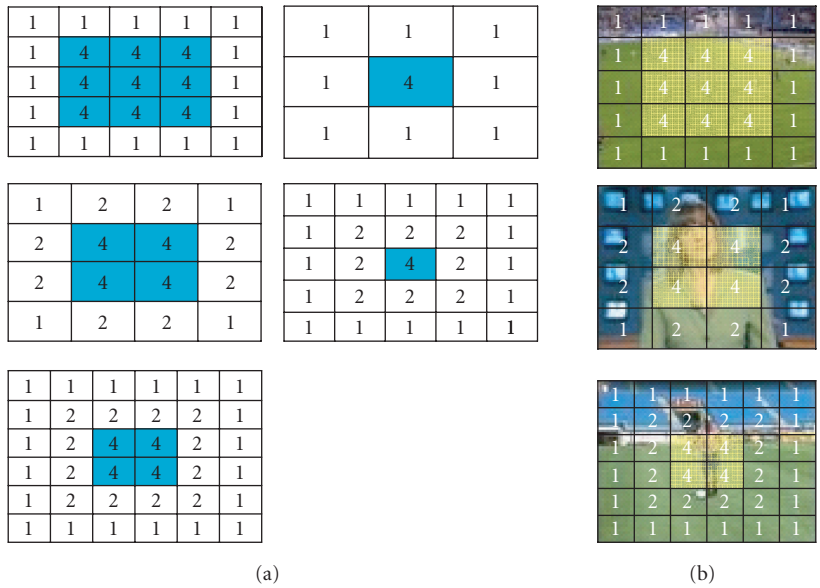


Figure 14: (a) The candidates of location weighting map, (b) the suitable one for each video.

For the first principle, we shift the bitrate from the nonattention IOBs to the attention IOBs, which are discriminated by video analyzer. In order to consolidate the effect of adaptation, the bit allocation scheme is also divided into two related hierarchical levels, frame level and object level.

High level and dynamic feature, like motion activities, are more meaningful and significant than other static characteristics, for example, color distribution, to represent semantic content within videos shots. Therefore, the motion feature used to describe the video content is a key factor to attract users' attention. To save computation in the adaptation decision, we consider GOP-based prediction. For a GOP, if the average mean and the average variance of motion magnitude of frames over the GOP are large, then this GOP within a video shot is highly probable high motion. Each frame within the GOP is important and attractive. Hence, we need to sustain full frame rate of this GOP and no frames are dropped to maintain full temporal quality. If the motion of video is slight, insignificant frames can be dropped without producing motion jitter and still maintaining acceptable temporal quality.

In frame level bit allocation, we take into account the average importance attention value (AIMP), average motion mean (AMM), and average motion variance (AMV), which are obtained by aggregating those values of all frames within a GOP. Meanwhile, we assign larger weight to I frames. There are three adaptation schemes proposed in the frame level bit allocation. In the first scheme, if the AMM and AMV within

```
For each GOP:
  Obtain target bit R for GOP
  Set bit Ib for I frame = 0 and bit Pb for p frame = 0
    Set predefined threshold M1, M2, V2
    1st Step: initialization:
      Calculate AIMP of each I frame (I_AIMP) and
      each P frame (P_AIMP), respectively
      Calculate total AIMP of I frames (TI_AIMP) and
       total AIMP of P frames (TP_AIMP), respectively
        Calculate AMM and AMV within a GOP
    2nd Step: Decision
          If AMM > M2 and AMV > V2
                Keep full frame rate
            Else If AMM > M1 and AMV <V2
              Drop all the B frames
                  Ib = R∗I_AIMP∗WI/
                  (TI_AIMP∗W1+TP_AIMP∗W2)
                  Pb = R∗P_AIMP∗W2/
                  (TI_AIMP∗W1+TP_AIMP∗W2)
            Else
                Skip all the frames except I frames
                 Ib = R∗I_AIMP/TI_AIMP
```

Algorithm 1: Bit Allocation of frames within a GOP.

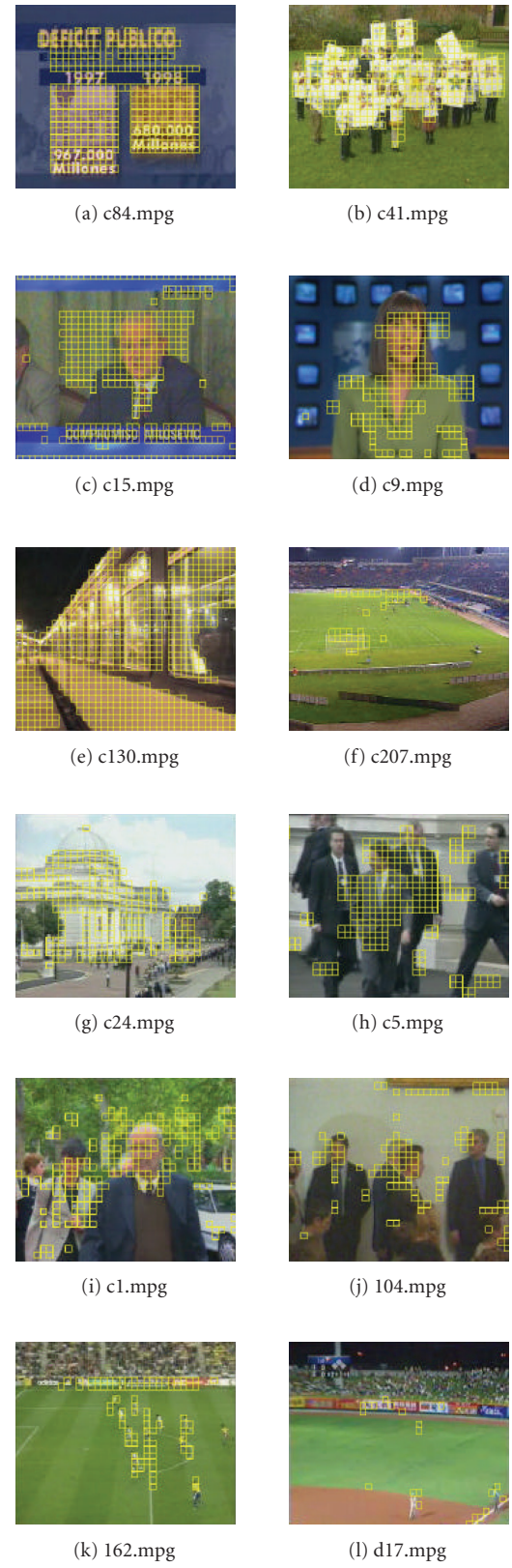
a GOP are larger than the predefined thresholds M2 and V2, as mentioned in Table 1, respectively, then the video shot is a high motion and we keep full frame rate. If the AMM is larger than M1 but the AMV is smaller than V2 within a GOP, then the video shot is moderate motion and we adopt the second scheme. All the B frames are dropped and all the bitrate saved is assigned to remaining I/P frames according to the AIMP values. In the third case, if the motion of video is slight, we can skip all the frames except I frames in motionless video. The three adaptation schemes used in frame level are summarized in Algorithm 1.

In object level, the bit allocation of each IOB in HIO2 is dependent on the frame level determination. After the bitrate for each frame is adjusted, we control the quantization parameters for different attention information objects in object level applying IOB-weighted rate distortion model.

For the second principle, we take into account the capability of client device in order to avoid transmitting redundant or useless data. For example, the mobile device is weaker in spatial resolution, color depth, computation power, and so forth. If the video is adjusted for transmission following the profile of client device as described in Section 4.2, the utilization of the resource is more effective.

5.2. IOB-weighted rate distortion model

Rate control based on the rate distortion theory is a fundamental technique in the coding process. By the fact that regions with different attention level have different sensitivity to coding error, we apply the video region-weighted Rate dis-



(a) c84.mpg (b) c41.mpg

(c) c15.mpg (d) c9.mpg

(e) c130.mpg (f) c207.mpg

(g) c24.mpg (h) c5.mpg

(i) c1.mpg (j) 104.mpg

(k) 162.mpg (l) d17.mpg

Figure 15: Information object results of video analyzer. (a) belongs to the motion class 1, fixed camera and static object. (b)–(e) are motion Class 2, fixed camera and moving objects. (f)-(g) are motion Class 3, moving camera and static scene. (h)-(l) are motion class 4, moving camera and moving objects, that is, object tracking.

tortion (RD) function [13] as IOB-weighted RD model in rate control:

$$D_i(R_i) = w_i * \sigma_i^2 * e^{-\gamma R_i}, \tag{11}$$

where D_i denotes the mean square value of the error of information object$_i$ (IOB$_i$) between decoded video frame and original video frame. w_i is the weight coefficient of IOB$_i$, which is determined by the importance attention value IMP$_i$ of IOB$_i$. γ is a constant number. σ_i^2 denotes the variance of the encoding signal, and R_i is the bitrate (bits/pixel) allocated to encode the IOB$_i$.

The Lagrange multiplier method is applied to solve the global optimization issue of rateallocation, and the result can be simplified to

$$R_i = R + \frac{1}{\gamma N_f S} \sum_{j \neq i} S_j \cdot \log \frac{w_i}{w_j} \quad (i = 1, 2, \ldots, N_f), \tag{12}$$

where Sj is the area size of IOB$_j$, S is the frame resolution, and Nf is the total number of IOBs for frame f. The bitrate R varies with the bit allocated in frame level adaptation scheme.

By the theory of acoustics, the human's perception of sound is a logarithmic form of the energy of sound, and hence the same discipline is for light. The importance attention value of IOB$_i$, IMP$_i$, represents the human's visual perception. The weight coefficient w_i denotes the weight of attention model, so IMP$_i$ is a logarithmic form of w_i, while w_i is an exponential form of IMP$_i$. Assume w_i has an exponential form as

$$w_i = C^{\cdot \mathrm{IMP}_i/k}, \tag{13}$$

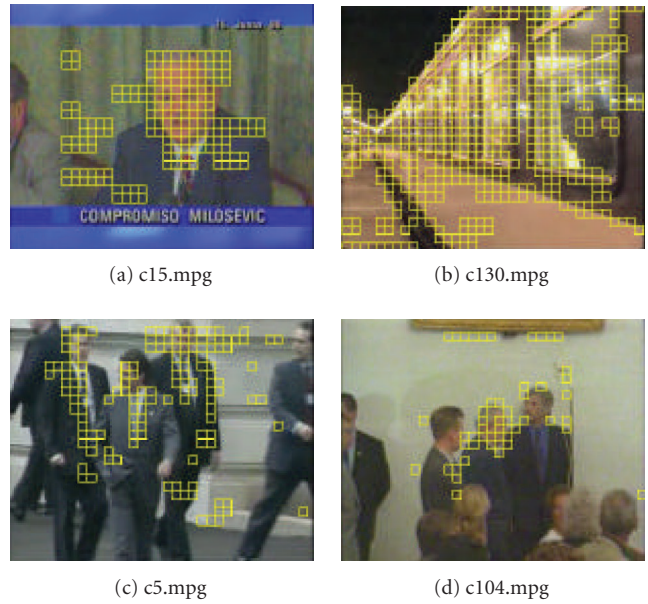
where k and C are constants. Here k and C are obtained from training videos and are set to be 32 and 0.0025, respectively. The IOBs are encoded by different quantization parameter QP$_i$, to meet the target bitrate R_i, according to the R-Q model [21]:

$$R_i = \alpha - \beta \log \mathrm{QP}_i \Longrightarrow \mathrm{QP}_i = e^{(\alpha - R_i)/\beta}, \tag{14}$$

where α accounts for overhead bits and β could be considered as a measure of complexity for each video segment. By applying the IOB-weighted RD model, we can obtain the appropriate bitrates and quantization parameters of each attention IOB for content-aware bitstream adaptation.

6. EXPERIMENTAL RESULTS AND DISCUSSION

To show the effectiveness of the proposed framework, we simulated the content-aware video adaptation using MPEG-7 test dataset [22], which includes various programs such as documentaries, news, interview, walking person, soccer, baseball, tennis, and scenery. In the test dataset, the degree of strength of the motions in these shots ranged from low, medium to high, and the size, of moving objects were classified as either small, medium, or large. We present the experimental results, including information object masks region of content analysis, bit allocation scheme, and visual perceptual quality. The aims of the experiments are to (1) evaluate the retrieval performance of IOBs using attention model, (2) analyze the adaptation policy when motion variance of frames was used in the adaptation process, and (3) evaluate the visual quality of the proposed content-aware video adaptation approach.



(a) c15.mpg (b) c130.mpg

(c) c5.mpg (d) c104.mpg

Figure 16: Information objects of improved video analyzer adopting correlational statistic model.

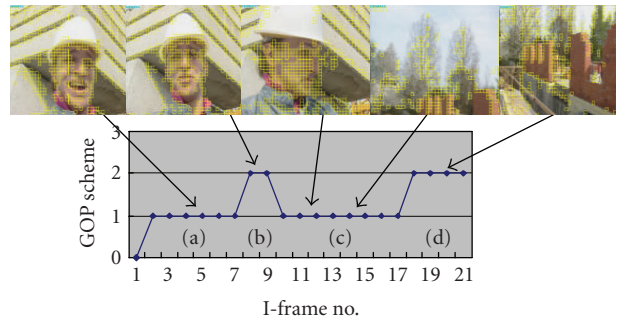


Figure 17: An example of bit allocation scheme.

IO mask region (content analysis)

First, we experiment on the performance of information object derived from Video Analyzer. The original input video rate before adapting was MPEG-2 encoded at 1.5 Mbps, a frame rate of 25 fps, and GOP parameters $N = 15$ and $M = 3$. All video sequences are 320×240 resolutions. Many previous researches about video analyzer are applicable only to one or two classes of videos, such as static background video analysis, like surveillance video analysis, and restricted domain video analysis, like tennis video analysis. Our video Analyzer is more general for different content types of videos. The four types of motion class as described above are used to

(a) Motion class 1 video: c84.mpg

(b) Motion class 2 video: c9.mpg

(c) Motion class 3 video: c207.mpg

(d) Motion class 4 video: c104.mpg

Figure 18: Comparison of visual quality. (1) The upper-left is original video. (2) The upper-right is information object result of video analyzer. (3) The bottom-left is the result video of normal uniform adaptation. (4) The bottom-right is the result video of our proposed adaptation.

verify the accuracy of the proposed video analyzer. Some significant IOBs derived from video analyzer are demonstrated with yellow mask in Figure 15.

In order to further validate the improvement of video analyzer, we apply the correlational statistic model in Section 4.3. Based on this model, the information of the foregoing frames will be utilized in the later analysis. Some insignificant IOBs, which are deemed as significant in Figures 15(h)–15(j) without using correctional model, are removed using correctional model as demonstrated in Figures 16(c) and 16(d) while significant IOBs are still remained. Compared with Figure 15, we can obtain better performance of information object analysis to validate the effectiveness of correlational statistic model as illustrated in Figure 16.

Bit allocation scheme

In order to judge the rationality of the GOP-based adaptation and bit allocation scheme, Figure 17 shows the relation between video content and the bit allocation scheme. When the motion variance of frames is larger, like main object moving as Figure 17(a) and camera panning as Figure 17(c), the adapter adopts GOPscheme 1 to keep full frame rate and maintain smooth motion. On the contrary, when the motion variance of frames is smaller, like Figure 17(b) and Figure 17(d), the adapter adopts GOPscheme 2 to drop 2/3 frames without introducing evident motion jitter.

Visual perceptual quality

Finally, under the same bitrate constraint, we compare the visual quality of video adapting using the proposed approach referred to as content-aware coding with that using conventional uniform approach, referred to as normal coding. Several video sequences of four motion classes are used for testing. The original video, information Object, visual perceptual quality of normal coding, and visual perceptual quality

Table 3: The subjective visual quality in different video classes.

	Good (%)	Fair (%)	Poor (%)
Class 1	69	18	13
Class 2	81	11	8
Class 3	74	15	11
Class 4	65	14	21
Avg.	72.7	14.3	13

of content-aware coding are shown in Figure 18, respectively. The output video rate after adapting was 120 kbps. α and β used in IOB-weighted RD model are set to be 1.91 and 0.52, respectively, which are calculated from some experimental video shots. Simultaneously, the parameters of each GOP are adjusted according to visual attention model.

We can see that, the visual quality of our proposed content-aware coding is better than conventional normal coding, especially in attraction regions, such as the two data charts in Figure 18(a), anchor person in Figure 18(b), football gate in Figure 18(c), and major walking person in Figure 18(d). It validates that the proposed content-aware video adaptation is effective.

A subjective experiment was designed to further evaluate the visual perceptual quality of the proposed content-aware adaptation framework. Thirtynine shots divided into four motion classes, with the lengths varying from 1 minute to 3 minutes, were selected as the test dataset. Then, fourteen observers were invited to give their subjective scores in the user study. In the experiments, the observers were required to give a comment of good, fair, or poor.

The statistical results of the experiment are listed in Table 3. Obviously, more than 70% of observers considered the proposed content-aware adaptation better than the traditional adaptation method and only 13% of them considered it poor. However, it is worth noting that in the motion class 4, our solution had a low score. In fact, this is reasonable because a codec maintains original frame rate to present a high-motion shot and all the frames are important objects.

7. CONCLUSION AND FUTURE WORK

In order to effectively utilize resource and improve visual perceptual quality, content-aware video adaptation is essential, especially in limited resource environments with very low-bitrate constraint. In this paper, we proposed a video analyzer to determine information objects (regions) of visual attention and a video adapter to dynamically adjust bitstream in accordance with the information of content and variations of resource. Information objects which attract more attention of viewers should be allocated more bits. In the proposed approach, video analyzer first analyzes features of video content such as brightness, location, motion, and energy to determine information objects. Those features are all extracted from compressed domain and hence it is computationally less demanding. Then, adaptation decision engine decides the adapting scheme and determines the target bitrate of each information object for bitstream adaptation engine to adapt video appropriately. The scheme is not restricted to specific codecs and can be easily implemented in many popular video-coding standards, such as MPEG-1, MPEG-2, MPEG-4, and H.264. Our experimental results have shown that the proposed mechanism is effective and achieves better subjective quality than conventional method under the same bandwidth constraint.

Though most of the thresholds are dynamically determined adapted to the content, we can find that some thresholds are not easy to be determined in the experiments due to the wide variation in video content. In order to improve the performance of the proposed approach, we can deploy further the classification of videos during the video analysis process. Therefore, we will continue to investigate the characteristics or domain knowledge of different content classes in our future work.

ACKNOWLEDGMENTS

S. Y. Lee's research was sponsored in part by the Lee and MTI Center for Networking Research, NCTU, and by NSC under Grant no. 95-2221-E-009-076-MY3 and NSC 95-2221-E-009-069-MY3.

REFERENCES

[1] A. Fox and E. A. Brewer, "Reducing WWW latency and bandwidth requirements by real-time distillation," in *Proceeding of the 5th International Conference on World Wide Web*, Paris, France, May 1996.

[2] J. R. Smith, R. Mohan, and C.-S. Li, "Scalable multimedia delivery for pervasive computing," in *Proceedings of the 7th ACM International Multimedia Conference (MULTIMEDIA '99)*, vol. 1, pp. 131–140, Orlando, Fla, USA, October-November 1999.

[3] A. Vetro, C. Christopoulos, and H. Sun, "Video transcoding architectures and techniques: an overview," *IEEE Signal Processing Magazine*, vol. 20, no. 2, pp. 18–29, 2003.

[4] S.-F. Chang and A. Vetro, "Video adaptation: concepts, technologies, and open issues," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 148–158, 2005.

[5] W. Lai, X.-D. Gu, R.-H. Wang, L.-R. Dai, and H.-J. Zhang, "Perceptual video streaming by adaptive spatial-temporal scalability," in *Proceedings of the 5th Pacific Rim Conference on Multimedia (PCM '04)*, pp. 431–438, Tokyo, Japan, November-December 2004.

[6] S.-F. Chang and P. Bocheck, "Principles and applications of content-aware video communication," in *Proceedings of the IEEE Internaitonal Symposium on Circuits and Systems (ISCAS '00)*, vol. 4, pp. 33–36, Geneva, Switzerland, May 2000.

[7] H. Wang, A. Divakaran, A. Vetro, S.-F. Chang, and H. Sun, "Survey of compressed-domain features used in audio-visual indexing and analysis," *Journal of Visual Communication and Image Representation*, vol. 14, no. 2, pp. 150–183, 2003.

[8] D.-Y. Chen, S.-Y. Lee, and H.-Y. M. Liao, "Robust video sequence retrieval using a novel object-based T2D-histogram

descriptor," *Journal of Visual Communication and Image Representation*, vol. 16, no. 2, pp. 212–232, 2005.

[9] D.-Y. Chen, M.-H. Hsiao, and S.-Y. Lee, "Automatic closed caption detection and filtering in MPEG videos for video structuring," *Journal of Information Science and Engineering*, vol. 22, no. 5, pp. 1145–1162, 2006.

[10] H.-C. Wu, Y.-W. Chen, M.-H. Hsiao, and S.-Y. Lee, "Robust video retrieval using a temporal edge pattern descriptor," in *Proceedings of the 19th IPPR Conference on Computer Vision, Graphics, and Image Processing (CVGIP '06)*, pp. 253–260, Taoyuan, Taiwan, August 2006.

[11] Y.-F. Ma, L. Lu, H.-J. Zhang, and M. Li, "A user attention model for video summarization," in *Proceedings of the 10th ACM International Conference of Multimedia (MULTIMEDIA '02)*, pp. 533–542, Juan les Pins, France, December 2002.

[12] S. F. Chang, "Content-based video summarization and adaptation for ubiquitous media access," in *Proceeding of the 12th International Conference on Image Analysis and Processing (ICIAP '03)*, pp. 494–496, Mantova, Italy, September 2003.

[13] W. Lai, X.-D. Gu, R.-H. Wang, W.-Y. Ma, and H.-J. Zhang, "A content-based bit allocation model for video streaming," in *Proceedings of IEEE International Conference on Multimedia and Expo (ICME '04)*, vol. 2, pp. 1315–1318, Taipei, Taiwan, June 2004.

[14] X. Xie, W.-Y. Ma, and H.-J. Zhang, "Maximizing information throughput for multimedia browsing on small displays," in *Proceedings of IEEE International Conference on Multimedia and Expo (ICME '04)*, vol. 3, pp. 2143–2146, Taipei, Taiwan, June 2004.

[15] C.-H. Chou and C.-W. Chen, "A perceptually optimized 3-D subband codec for video communication over wireless channels," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 6, no. 2, pp. 143–156, 1996.

[16] C.-H. Chou and Y.-C. Li, "Perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 5, no. 6, pp. 467–476, 1995.

[17] S. R. Gulliver and G. Ghinea, "Region of interest displays: addressing a perceptual problem?" in *Proceedings of 6th IEEE International Symposium on Multimedia Software Engineering (ISMSE '04)*, pp. 2–9, Miami, Fla, USA, December 2004.

[18] M. H. Lee, S. Nepal, and U. Srinivasan, "Edge-based semantic classification of sports video sequences," in *Proceedings of International Conference on Multimedia and Expo (ICME '03)*, vol. 1, pp. 157–160, Baltimore, Md, USA, July 2003.

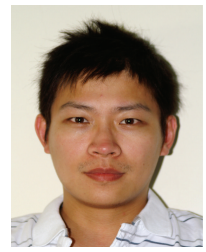
[19] M. H. Lee, S. Nepal, and U. Srinivasan, "Role of edge detection in video semantics," in *Proceedings of Pan-Sydney Area Workshop on Visual Information Processing (VIP '02)*, vol. 22 of *Conferences in Research and Practice in Information Technology*, pp. 59–68, Sydney, Australia, 2002.

[20] Y. Yuan, D. Feng, and Y. Zhong, "A mixed scheme to improve subjective quality in low bitrate video," in *Proceedings of IEEE Wireless Communications and Networking Conference (WCNC '04)*, vol. 3, pp. 1754–1759, Atlanta, Ga, USA, March 2004.

[21] W. Ding and B. Liu, "Rate control of MPEG video coding and recording by rate-quantization modeling," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 6, no. 1, pp. 12–20, 1996.

[22] ISO/IEC JTC1/SC29/WG11/N2466, "Licensing Agreement for the MPEG-7 Content Set," Atlantic City, NJ, USA, 1998.

Ming-Ho Hsiao received the B.S. degrees in computer sciences and information engineering from Fu Jen Catholic University, Taiwan, in 2000. He received the M.S. degree in computer sciences and information Engineering from National Chiao Tung University, Taiwan, where he is currently pursuing the Ph.D. degree. His research interests are video signal processing, content-based indexing and retrieval, and distributed multimedia system, in particular, media server architecture and peer-to-peer system.



Yi-Wen Chen is currently a Ph.D. candidate of computer science and information Engineering in National Chiao Tung University, Taiwan. He received the B.S. and M.S. degrees in computer science and information engineering from National Chiao Tung University, Taiwan, in 2000 and 2002, respectively. He has been engaged in the research areas of computer vision and video/image compression.

Hua-Tsung Chen received his B.S. and M.S. degrees in computer science and information Engineering from National Chiao Tung University, Taiwan, in 2001 and 2003, respectively. Currently, he is pursing the Ph.D. degree in computer science and information Engineering in National Chiao Tung University, Taiwan. His research interests include computer vision, video signal processing, content-based video indexing and retrieval, multimedia information system, and music signal processing.

Kuan-Hung Chou was born in Taipei, Taiwan, in 1981. He received the B.S. degree in computer information science from National Chiao Tung University, Hsinchu, Taiwan, in 2003, and the M.S. degree in computer science and information engineering from National Chiao Tung University, Hsinchu, Taiwan, in 2005. In 2005, he joined the Information and Communications Research Laboratories, Industrial Technology Research Institute, Hsinchu, Taiwan, where he is currently a Design Engineer. His current research interests include video streaming and content-aware video adaptation.

Suh-Yin Lee received the B.S. degree in electrical engineering from National Chiao Tung University, Taiwan, in 1972, and the M.S. degree in computer science from University of Washington, USA, in 1975, and the Ph.D. degree in computer science form Institute of Electronics, National Chiao Tung University. Her research interests include content-based indexing and retrieval, distributed multimedia information system, mobile computing, and data mining.