Content-Aware Video Adaptation in Low Bit-rate Constraint

Ming-Ho Hsiao, Hua-Tsung Chen, Yi-Wen Chen, Kuan-Hung Chou, Suh-Yin Lee
College of Computer Science, National Chiao Tung University
{ mhhsiao, huatsung, ewchen, choukh, sylee}@csie.nctu.edu.tw

ABSTRACT

With the development of wireless and the improvement of mobile device capability, video streaming is more and more widespread applied in such environment. Under the limited resource and inherent constraints, appropriate video adaptation has become one of the most important and challenging issues in wireless multimedia application related areas. We propose a novel approach to adapt video based on content information in order to effectively utilize resource and improve visual perceptual quality in this paper. According to the analyzed characteristics of brightness, location, motion vector, and energy features, combined with capability of client device and correlational statistic model, the attractive or interesting regions of video scene are derived. Therefore, the Region Weighted Rate-Distortion is used for adjusting the bit allocation. Video adaptation scheme dynamically adapt video bitstream through object, frame, and GOP levels. Experimental results show that the proposed scheme is efficient and achieves better visual quality.

1: INTRODUCTIONS

With the development of wireless and the improvement of mobile device capability, the desire for mobile users to access video is becoming stronger. These devices including cellphone (smart phone), PDA, and laptop have enough computing capability to decode and display video and receive video via wireless channel, like 802.11. However, due to some inherent constraints in wireless multimedia application, like limited bandwidth in wireless and high variation in device resource, how to adequately utilize such resource to get better quality is an important issue.

Video adaptation is usually used in response to the huge variation of resource constraints. In traditional video adaptation, the adapter considers available bit rate and network buffer occupancy to adjust the transmitted data while streaming video [1] [2]. Vetro *et al.* presented a general framework that defines the fundamental entities and important concepts related to video adaptation [3]. Furthermore, the authors indicate 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. Some transcoding schemes, such as bit-rate reduction, spatial and temporal resolution reduction, and error resilient transcoding is introduced in [4].

Although the viewpoint of Information Theory, same bitrates deliver same amount of information, it may be not true for human visual perception. Generally speaking, viewers can only be attracted to a relatively small part of the video display with acuity drop-off in peripheral areas at any point in time. Accordingly, by adjusting allocation of bitrate from peripheral regions of the frame to regions-of-interest, viewers can get better visual perceptual quality [5]. In opposition to traditional video adaptation, content-based video adaptation can effectively utilize content information in bit allocation and adaptation and is a promising research direction.

In this paper, a content-aware video adaptation is proposed based on visual attention model. It first analyzes the content of video to derive the important regions which have high degree of attraction level; then allocate bitrate and assign adapting scheme according to the content information in order to acquire better visual quality and avoid unnecessary resource waste in low bitrate constrain. The problem addressed in this paper is to utilize content information for improving the quality of a transmitted video bitstream subject to low bit-rate constraints, which especially applies to mobile device in wireless network environment. Three major issues are considered:

(1) How to quickly derive the important object from video.
(2) How to adapt video streams according to content and mobile device condition.
(3) How to find an appropriate video adaptation approach to get the better quality.

In the following, we will analyze related issues through theory and experiments. Fig. 1 shows the architecture of the proposed content-aware video adaptation scheme. Initially, video streams are processed by video analyzer to derive the content features of each frame/GOP and the important regions which have high degree of attraction. Subsequently, the adaptation decision engine determines the adaptation scheme and parameters according to the content information derived from video analyzer, device capability obtained from profile, correlational statistic model, and region weighted rate-distortion model.

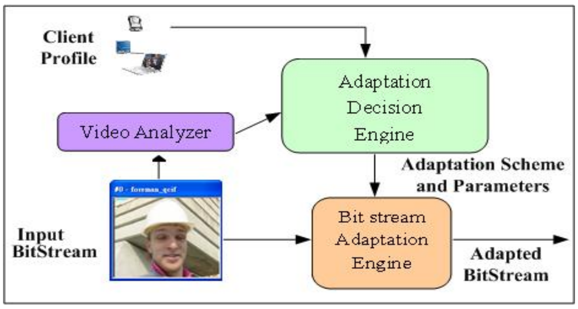

Client Profile
Adaptation Decision Engine
Video Analyzer
Adaptation Scheme and Parameters
#0 - foreman_qcif
Input BitStream
Bit stream Adaptation Engine
Adapted BitStream

Fig. 1 The architecture of the proposed system

Finally, the bitstream adaptation engine adapts video based on the bit allocation scheme.

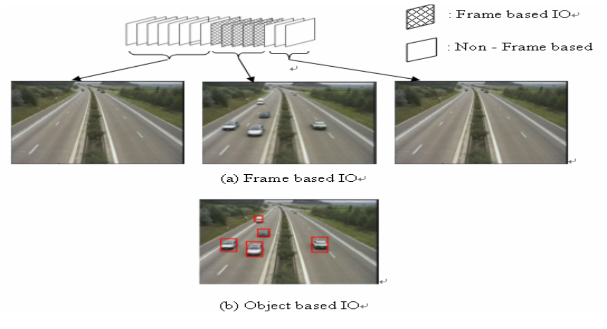
The rest of this paper is organized as follows. Section 2 presents a novel video content analyzer. A hybrid feature-based model for video content adaptation decision is illustrated in Section 3. In Section 4, we describe the proposed bitstream adaptation approaches. The experimental results and discussion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

2: VIDEO ANALYER

In this section, we describe the first component of the proposed system. Video Analyzer is used to analyze features of video content for deriving meaningful information. In Section 2.1, we import the concept of Information Object to represent the content of video. Finally, we introduce the relation of selected features with visual perception effects in Section 2.2.

2.1: INFORMATION OBJECT (IO) DERIVAATION

Different parts of content have different importance values. Attention-based selection [6] allows only attention-getting parts 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 B_i. An information object is an information carrier that delivers the author's intention and catches part of the user's attention as a whole. Fig. 2 is a content representation model example consisted of some Information Objects.



(a) Frame based IO

(b) Object based IO

Fig. 2 Examples of the content representation model

We import the "Information Object" concept, which is a modification of [6] to agree with video content, defined as below.

Definition 1: The basic content representation model for a video shot *S* is defined as a set which has three relative hierarchical levels of Information Objects:

$$S = \{H_i\} \quad 1 \le i \le 3 \quad , \tag{1}$$

$$H_i = \{B_j\} \quad 1 \le j \le N \quad , \tag{2}$$

and

$$B_j = (IMP_j, CON_j) \tag{3}$$

where H_i, is the perception of object, frame, or GOP level of *S*, respectively

B_j, is the j^{th} Information Object in H_i of *S*

IMP_j, is the importance value of B_j

and *CONj*, describes the members the Information Object contained.

2.2: VISUAL FEATURE SELECTION

A video bitstream contains a lot kinds of information that can be extracted from pixel or compressed domain. Since fast processing is required to suit the presented application scenario, we consider only the features that can be derived from data in compressed domain.

Visual effect is considered in four feature domains — brightness, spatial location, motion, and energy. For each feature, we briefly discuss the extraction methods (i.e. the relationship with data extraction), visual perceptive effect, and possible limitation caused by certain video content.

2.2.1: BRIGHYNESS

Generally speaking, the human perception is always attracted by the brighter part, which is referred to as brightness attraction property. For example, the brightly colored, or strongly brightness contrasted parts of a video frame, even in the background, always have high attraction. So, the brightness characteristic is an important feature to identify the Information Objects. Here, we use the DC value of the luminance of I frame to derive the block brightness. Our brightness attention model containing mean of brightness, variance of brightness, and location based brightness histogram presents as the following:

$$B = \frac{DCvalue}{B\_level} \times B\_\mathrm{var} \tag{4}$$

where *DCvalue* is the DC value of luminance, *B_level* is obtained from the average luminance of the previous calculated frame, and *B*_var denotes the DC value variance of current and surrounding eight blocks.

In order to improve the brightness attention model in response to attraction, we design a location based brightness histogram which utilizes the correlation between brightness and distribution to identify the important brightness bin and roughly discriminate foreground from background. The module calculates mainly distribution of each brightness bin 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 will apply the average region value of the brightness bin to adjust the *B* obtained from Eq. (4) when the proportion of the brightness bin is greater than certain degree. The function of adjustment is as follows:

$$B' = \begin{cases} 0 & \text{if } lbbh\ (B) \le 1 \\ B-1 & \text{if } 2 \le lbbh\ (B) < 1 \\ B & \text{if } 3 \le lbbh\ (B) < 2 \\ B+1 & \text{if } 4 \le lbbh\ (B) < 3 \\ 5 & \text{if } lbbh\ (B) > 4 \end{cases} \qquad (5)$$

where *B'* is the adjusted brightness attention value using location based brightness histogram model, and the *lbbh*() function denotes the average region value of the brightness bin.

In Fig. 3, we can evidently discover that the results of the location based brightness histogram have large refinement against pure brightness attention model.

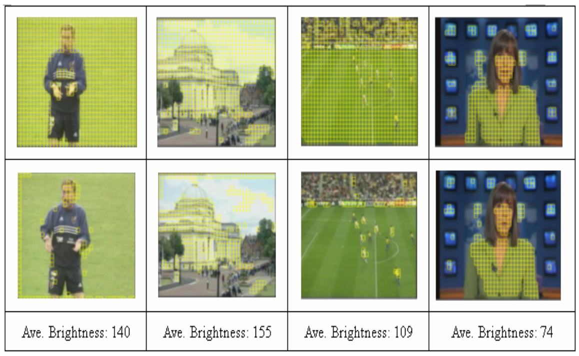


Fig. 3 IO derived from brightness without (first row) and with combining the location based brightness histogram (second row)

2.2.2: 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 always operate the camera to focus on the main object, i.e. put the primary object on the center of the camera view, in the technique of photography. 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. Fig. 4 illustrates the weighting map and the adapted result based on the location factor. But, for different type 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 statistic information of IO distribution.

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

Fig. 4 Location weighting map and adapted video according to the location feature

2.2.3: MOTION

After extensive observation of a variety of video shots, 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 10~15%. The second class is fixed camera and some objects 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 20~30%. The third class, the camera move while no change in the scene, is about 30~40%. For instance, some shots of scenery scene belong to this type. The fourth class, the camera is moving while some objects move in the scene, such as object tracking shots. The proportion of this class is also about 30~40%.

Because the meaning and the importance degree of the motion vector 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 different representations in 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 the zero motion have a certain degree proportion. 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 camera motion cases. However, the major direction of motion vectors 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. According to the above discussions, 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 . More than 80% of test video sequences can be classified into the correct motion class by our proposed Motion Class model.

Table 1 Video are classified according to motion vector

			Motion magnitude		Zero motion
Class	Camera	Object	Mean	Variance	(%)
1	Fixed	Static	near 0	quite small	near 100%
2	Fixed	Moving	smaller	smaller	middle
3	Moving	Static	larger	middle/larger	small
4	Moving	Moving	larger	larger	small

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

In Motion Class 1 and 2:

$$Mattention = \frac{magnitude}{\alpha - \beta} \times MA, \quad \text{when } \alpha \ge magnitude \ge \beta \tag{6}$$

In Motion Class 3 and 4:

$$Mattention = \frac{magnitude}{\alpha - \beta} \times (1 - |0.5 - MA|), \quad \text{when } \alpha \ge magnitude \ge \beta \tag{7}$$

where *Mattention* is the motion attention value, *magnitude* denotes motion magnitude, *MA* represents the bin proportion of the motion angle histogram for each block, and α, β are two thresholds for noise elimination and normalization.

2.2.4: ENERGY

Another influence on perceptual attention is the texture complexity, i.e. the distribution of edges. People usually pay more attention to the objects which have greater or less magnitude of edge than average and referred to as energy attraction property. For example, the 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 to extract edges. The two horizontal and vertical edge features can be formed by two-dimensional DCT of a block.

$$\begin{aligned} &Horizontal\ Feature: H = \{H_i : i = 1,2,...,7\} \\ &Vertical\ Feature: V = \{V_j : j = 1,2,...,7\} \end{aligned} \tag{8}$$

in which H_i and V_j correspond to the DCT coefficients $F_{u,0}$ and $F_{0,v}$ for *u, v* = 1, 2, …, 7.

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:

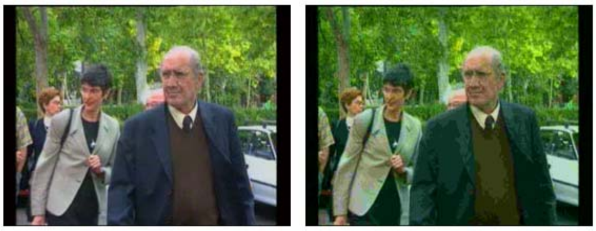
$$E = \sqrt{H^2 + V^2} \tag{9a}$$

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

The gradient energy of I frame is then obtained which is represented as the edge energy feature.

The gradient energy of I frame is then obtained which is represented as the edge energy feature.

However, the influence of perceptual distortion in parts with large edge energy or small edge energy is little. As shown in Fig. 5, we can discover high energy regions like tree have less visual distortion than other regions like walking person in (b) under the same quantization constraint. Although we have explained that objects which have greater or less magnitude of edge than average attract more human attention. On the contrary, the visual perceptual distortion introduced by quantization is small in extremely high or low energy cases.



(a) Original frame (b) Uniform quantization adapted frame

Fig. 5 Comparison of the visual distortion in different edge energy regions

Accordingly, our energy model combined the above two aspects is illustrated as below. We introduce four thresholds which are derived from the mean and variance of energy of the previous calculated frame. According to the energy *E* obtained from Eq. (9a), assign each block the energy attention value, as shown in Fig. 6. When *E* is near the energy mean of the frame, we assign the block medium energy attention value. When *E* belongs to higher or lower regions, we assign the block high energy attention value. In extreme energy case, we assign such blocks the lowest energy attention value because their visual distortion is unobvious.

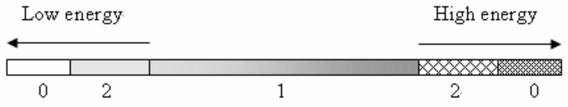


Fig. 6 The energy attention model

3: 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 utilize the feature characteristic 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. In the first class, due to the motions are almost zero motions and meaningless, 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, in the third class the features we considered, i.e. brightness, location, and energy while motion ignored, 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 it depends on the speed of the camera motion.

4: BITSTREAM ADAPTATION

Bitstream Adaptation Engine is used to control the bit rate and adapts the bitstream based on Video Analyzer and Adaptation Decision Engine. In Section 4.1, we present bit allocation scheme of our content aware adaptation. Subsequently, we introduce the concept of Region Weighted Rate-Distortion Model used to execute rate control in Section 4.2.

4.1: BIT ALLOCATION SCHEME

In our allocation scheme, we consider two major principles: improve visual perceptual quality and avoid unnecessary resource waste.

For the first principle, we shift the bit rate from the non-attention regions to the attention regions, which are discriminated by Video Analyzer. In order to consolidate the effect of adaptation, our bit allocation scheme is also divided into three relative hierarchical levels, i.e. object, frame, and GOP levels as the adaptation decision principle.

In GOP level, we consider the GOP based Importance Value which aggregates the *IMP* of frames to obtain, average motion mean, and average motion variance to determine adapted scheme. For example, when the motion of video is slight, more frames can be dropped without producing motion jitter and keeping acceptable temporal quality. There are three schemes as the following.

(1) Full frame rate. No frames are dropped to maintain full temporal quality.

(2) 1/3 frame rate. Suppose the GOP structure of compressed video stream is "IBBPBBPBBPBBPBB". All the B frames are dropped, and all the saving bitrate is assigned to I/P frames.

(3) Skip all frame except I frame. It is used in very motionless video.

In frame level, the whole frame adjustment utilizes the aggregation of the *IMP* in a frame to judge the importance of the frame as shown in Fig. 7 and then we can expect higher coding efficiency and higher quality.

In object level, we utilize the above-mentioned region-weighted RD model as described in Section 4.2 to adjust the quantization parameters in different attention Information Object regions. The bit allocation in object level must base on the GOP level determination.

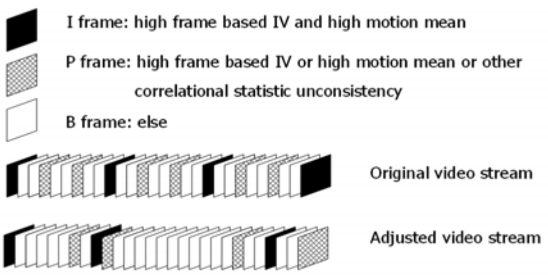

I frame: high frame based IV and high motion mean
P frame: high frame based IV or high motion mean or other correlational statistic unconsistency
B frame: else
Original video stream
Adjusted video stream

Fig. 7 Frame based dynamic bit allocation scheme

4.2: REGION WEIGHTED RATE DISTORION MODEL

Rate control is a fundamental technique in the coding process, which is based on the rate distortion theory. Based on the fact that regions with different attention level have different sensitivity to coding error, [7] proposed a video region-weighted rate-distortion (R-D) function:

$$D_i(R_i) = w_i * \sigma_i^2 * e^{-\gamma R_i}\ , \qquad (10)$$

where D_i denotes the mean square value of the error of Regions-of-Interest$_i$ (ROI$_i$) between decoded video frame and original video frame, w_i denotes the weight coefficient of ROI$_i$, which is determined by the attention level A_i of ROI$_i$, γ is a constant number, σ_i^2 is the variance of the encoding signal, and R_i is the bit rate (bits/pixel) used to encode the ROI$_i$.

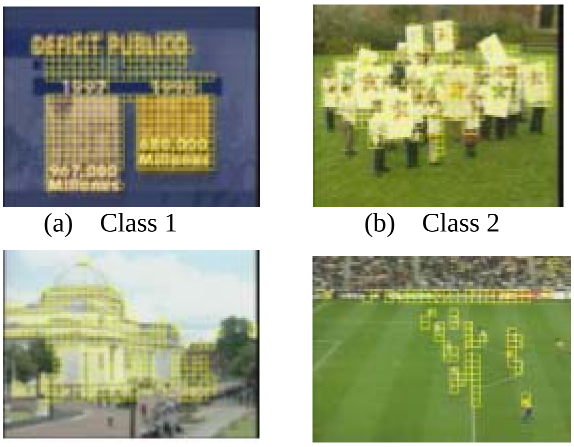
Based on the analysis of the region weighted RD model, we can obtain the appropriate bit rates and quantization parameters of each attention level region for content aware bitstream adaptation.

5: EXPERIMENTAL RESULTS

The proposed approach of content-aware video adaptation is applied to various kinds of videos. The contents of testing video sequences mainly include news, interview, walking person, soccer, baseball, tennis, and scenery. We present the experimental results, including Information Object masks region of content analysis, bit allocation scheme, and visual perceptual quality.

5.1: IO REGION

First, we experiment on the performance of Information Object derived from Video Analyzer. Our Video Analyzer is more general for content type of videos. Some experimental results of Video Analyzer are demonstrated in Fig. 8.



(a) Class 1 (b) Class 2

(c) Class 3 (d) Class 4

Fig. 8 Information Object results of Video Analyzer

5.2: BIT ALLOCATION SCHEME

In order to judge the rationality of the GOP based adaptation and bit allocation scheme, the Fig. 9 shows the relation between video content and the bit allocation scheme. When the motions of the interval are larger, like main object moving as (a) of Fig. 9 and camera panning as (c) of Fig. 9, the adapter adopts GOPscheme 1 to keep full frame rate and maintain smooth motion. On the contrary, when the motions of the interval are smaller, like (b) and (d) of Fig. 9, the adapter adopts GOPscheme 2 to drop 2/3 frames without introduce evident motion jitter.

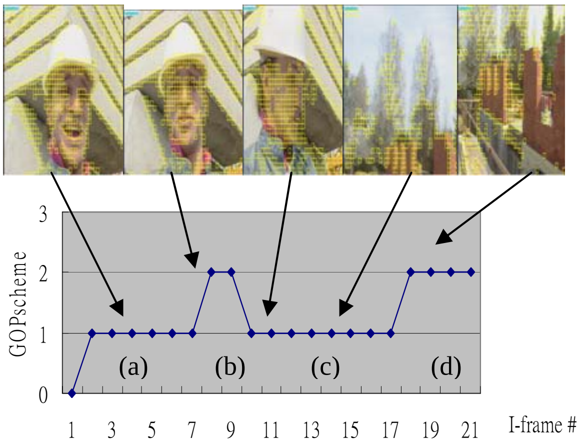


Fig. 9 Example of Bit allocation scheme

5.3: VISUAL PERCEPTUAL QUALITY

Finally, we compare the visual quality with adapting video using our approach, which referred to as Content-aware coding and with adapting video using conventional uniform approach, which referred to as normal coding under the same bitrate constraint. Several video sequences of four Motion Classes are used to test. The original video, Information Object, visual perceptual quality of normal coding, and visual perceptual quality of Content-aware coding are shown in Fig. 10, respectively.

We can see that, the visual quality of our proposed Content-aware coding is better than conventional normal coding, especially in attraction regions. It proved that our content-aware video adaptation is effective.

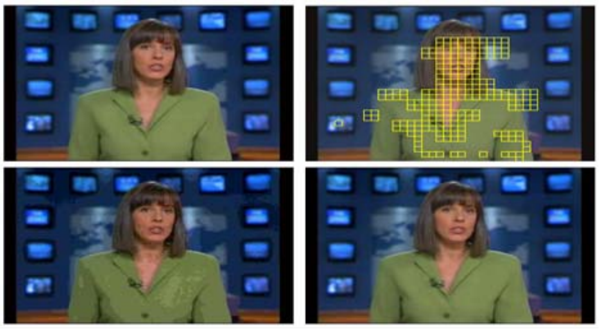


Fig. 10 Comparison of visual quality

6: CONCLUSION

In this paper, we proposed a video analyzer to determine visual attention regions and a video adapter to dynamically adjust bitstream in accordance with the information of content and variations of resource. Video analyzer first analyzes some features of video content such as brightness, location, motion, and energy to determine Information Objects. Then, adaptation decision engine decides the adapting scheme and determines the target bit rate of each region for bitstream adaptation engine to suitably adapt video. Our experimental results have shown that the proposed method is effective and achieves better subjective quality than conventional method under the same bandwidth constant.

REFERENCES

[1] J. R. Smith, R. Mohan and C. Li, “Scalable Multimedia Delivery for Pervasive Computing,” Proceeding of ACM Multimedia 99, pp. 131-140, Orlando, FL, Oct. 1999.

[2] S. F. Chang, “Video Adaptation: Concepts, Technologies, and Open Issues,” IEEProceedings of the IEEE, ISSN: 018-9219, Vol. 93, Issue 1, pp. 148-158, Jan. 2005.

[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, Mar. 2003.

[4] 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, Issue 2, pp. 150-183, Jun. 2003.

[5] Y. F. Ma, L. Lu, H. J. Zhang and M. Li, “A User Attention Model for Video Summarization,” Proceeding of ACM Multimedia 02, pp. 533-542, Juan-les-Pins, France, Dec. 2002.

[6] X. Xie, W. Y. Ma, H. J. Zhang, “Maximizing Information Throughput for Multimedia Browsing on Small Displays,” Proceeding of International Conferences on Multimedia and Expo 2004 (ICME 04), Vol. 3, pp. 2143-2146, Jun. 27-30, 2004.

[7] W. Lai, X. D. Gu, R. H. Wang, W. Y. Ma, H. J. Zhang, “A Content-Based Bit Allocation Model for Video Streaming,” Proceeding of International Conferences on Multimedia and Expo 2004 (ICME 04), Vol. 2, pp. 1315-1318, Jun. 27-30, 2004.