Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2007, Article ID 25415, 12 pages
doi:10.1155/2007/25415

Research Article

An Attention-Information-Based Spatial Adaptation Framework for Browsing Videos via Mobile Devices

Houqiang Li, Yi Wang, and Chang Wen Chen

Department of Electronic Engineering and Information Science (EEIS), University of Science and Technology of China, Hefei 230027, China

Received 1 September 2006; Revised 8 February 2007; Accepted 3 May 2007

Recommended by Chia-Wen Lin

With the growing popularity of personal digital assistant devices and smart phones, more and more consumers are becoming quite enthusiastic to appreciate videos via mobile devices. However, limited display size of the mobile devices has been imposing significant barriers for users to enjoy browsing high-resolution videos. In this paper, we present an attention-information-based spatial adaptation framework to address this problem. The whole framework includes two major parts: video content generation and video adaptation system. During video compression, the attention information in video sequences will be detected using an attention model and embedded into bitstreams with proposed supplement-enhanced information (SEI) structure. Furthermore, we also develop an innovative scheme to adaptively adjust quantization parameters in order to simultaneously improve the quality of overall encoding and the quality of transcoding the attention areas. When the high-resolution bitstream is transmitted to mobile users, a fast transcoding algorithm we developed earlier will be applied to generate a new bitstream for attention areas in frames. The new low-resolution bitstream containing mostly attention information, instead of the high-resolution one, will be sent to users for display on the mobile devices. Experimental results show that the proposed spatial adaptation scheme is able to improve both subjective and objective video qualities.

1. INTRODUCTION

Recent advances in wireless networks, especially with the emergence of 3G network, have enabled a new array of applications in image and video over wireless networks beyond traditional applications in voice and text. Real-time multimedia applications, such as video streaming, have become feasible in the wireless environment. In particular, with the growing popularity of mobile devices, users can enjoy videos anyplace and anytime over wireless networks. However, for video streaming application in the wired environment, the videos stored in the video server are generally encoded at high resolution (HR) and high bitrate to guarantee users' browsing experiences. There are two serious difficulties in enjoying these videos with mobile devices over wireless networks. One is the lower bandwidth of wireless networks. The other critical constraint is the limited display size of mobile devices, which often hinders the users to fully enjoy the video scene. It is very much desired for mobile users to access videos via mobile devices with limited display size but with an enhanced viewing experience.

Pervasive media environment including different types of terminals and networks brings critical difficulties in achieving universal multimedia access (UMA) [1] which refers to the access and consumption of multimedia content over heterogeneous networks by using diverse terminals in a seamless and transparent way. Video adaptation [2] is an emerging research field that offers a rich set of techniques to address all kinds of adaptation problems for supporting UMA. In general, it transforms the input video to an output in video or augmented multimedia form in order to meet diverse resource constraints and user preferences. One common adaptation solution to the constraint from the display size of mobile devices is through spatial transcoding [3, 4]. Through simply downsizing HR videos into low-resolution (LR) ones by a factor of integer or fraction, users may be able to browse video scenes with limited display size. The bitrates will also be reduced accordingly. Though the two constraints are addressed by this solution, excessive simple resolution reduction will cause significant loss in the perception of desired information. This is because the simple downsizing will result in unacceptable reduction of the attention area within the video frames.

Several researchers have also proposed adaptation solutions based on region of interest (ROI) [5–7]. These methods improve the visual quality by increasing the number of bits allocated for ROI in frames. However, they did not address the problem caused by the limited display size of mobile devices. In [8], the author proposed an ROI-based image transcoding scheme for browsing images in heterogeneous client displays. The proposed approach cannot be easily applied to video transcoding that requires smooth transition between frames.

In order to meet the constraints of display size and bandwidth while optimizing the perceived video quality, we propose an attention-information-based spatial adaptation framework which is composed of two processes: the preprocessing of video content and the video adaptation stage. During video encoding to generate the compressed bitstream, a special attention model is adopted to detect the attention objects within frames. Then the attention information will be embedded into the bitstreams with proposed supplement-enhanced information (SEI) structure [9]. Furthermore, based on the detected attention information, we develop an approach to improve both coding performance of original video and transcoding performance and visual quality of video adaptation by adjusting bit allocation strategy for attention and nonattention areas within a video frame. When the video server sends the HR video bitstream containing attention information to the client, our adaptation system will crop attention areas in frames and intelligently assemble them into a new video sequence containing as much desired information as possible. Then it compresses the new sequence into a bitstream by a technique of fast-mode decision [10] which utilizes the motion and residue information included in the original bitstream. The size of attention areas can be adaptively adjusted according to different display sizes of mobile devices.

The rest of this paper is organized as follows. Section 2 gives an overview of the spatial adaptation framework. Section 3 introduces the details of video content generation which includes the detection of attention objects, proposed SEI structure, and adaptive QP adjustment approach. Section 4 presents the procedure to perform adaptation operation with the embedded attention information. Several experimental results have been demonstrated in Section 5. Section 6 concludes this paper with a summary.

2. THE ATTENTION-INFORMATION-BASED SPATIAL ADAPTATION

The proposed attention-information-based spatial adaptation is based on an ROI transcoding scheme that we developed earlier. In this section, we will first present an overview of the spatial adaptation scheme based on ROI transcoding. Even though the spatial adaptation based on ROI transcoding is able to provide high-quality ROI for display on the mobile devices, we will point out that there is a need to design a new scheme in order to overcome two inherent shortcomings associated with the ROI transcoding-based video adaptation. Finally, we will present an overview of the proposed attention-information-based spatial adaptation.

2.1. Spatial adaptation system based on ROI transcoding

In our previous work, we have developed a spatial adaptation scheme based on region-of-interest transcoding [10]. We assume that a video server prestores high-quality videos and serves various mobile terminals, including PCs, smart phone, and PDA. When a mobile user client requests for a service, the server sends a video to the client. We assume that this system is placed on a server or a proxy and will adapt the HR video to generate an LR video suitable for the display size of the user's mobile device and the bandwidth of the mobile link. The adaptation will improve the user's perceptual experiences by appropriately transcoding the HR video to generate the LR video for mobile devices. For different users, the reduction of the video resolution can be different and will be decided by the real display sizes of mobile devices. The system consists of three main modules: *decoder, attention area extractor*, and *transcoder*. The module of decoder is to decode the HR bitstream. Decoded information will be transmitted to the module of attention area extractor and transcoder. The module of attention area detector includes several submodules: motion, face, text, and saliency detectors to extract attention objects and the combiner to output smooth attention areas for the following transcoder. Based on the output areas, the last module, transcoder, will produce the LR bitstream. The transcoding module is composed of three submodules: mode decision, motion vectors adjustment, and drifting error removal. The details of ROI-based transcoding are given in [10].

2.2. The need for a more intelligent solution

Although the previously developed ROI-based transcoding is able to perform video adaptation for mobile devices with small display and limited bandwidth, this system has two critical shortcomings that need to be overcome to maximize mobile video browsing experiences. The first shortcoming of the ROI-based transcoding is the need to perform the detection of four types of attention objects separately in order to obtain a robust ROI within a given video. The computational operations to perform these detections and to combine the detection results will become a significant burden for either server or proxy. The second shortcoming of the ROI-based transcoding is the need to perform ROI detection for different users every time these users request the video browsing service. Such repeated operations will sometimes overwhelm the proxy server.

However, these shortcomings can be overcome if the compressed video at the original server can be augmented with ROI information for proxy server to access. If we are able to embed the ROI information into the bitstream of

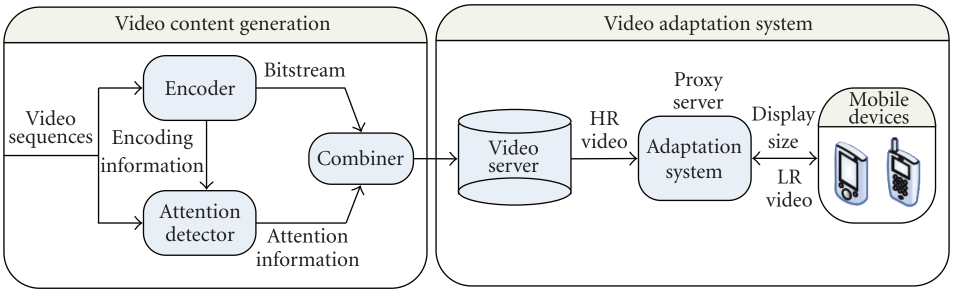


Figure 1: The structure of attention-information-based spatial adaptation framework.

compressed video and extract by the proxy server, the burden of the computational complexity can be shifted from the transcoding to encoding. It is this analysis that motivated us to design an attention-information-based spatial adaptation framework for browsing video via mobile devices.

This intelligent design of spatial adaptation is based on the assumption that the videos stored in the video server are usually offline generated and the computational complexity is not an issue with offline operations. Furthermore, we also assume that the attention objects in each video frame may remain the same even for different mobile users. This is because the attention model we adopted is quite generic for a wide variety of users. If we are able to move the attention detection operation from the transcoding process to the encoding process, then, we will be able to shift the complexity from proxy to the video server.

2.3. Overview of the attention-information-based spatial adaptation

Based on the above analysis, we propose an intelligent spatial adaptation framework in this research as shown in Figure 1. This framework has two parts: video content generation and video adaptation operation. During the generation of compressed video bitstreams, the attention information will also be detected simultaneously. Then the bitstreams and the attention information will be stored together in the video server which serves not only mobile devices users but also high-resolution PC users. When the server offers services for mobile devices users, the adaptation system placed on a proxy or server will perform adaptation manipulation on the HR video by making use of the attention information in the video to meet the display constraint of the mobile devices. It should be noted that the adaptation operation will not be performed for high-end users even though the attention information is available. That is because the original HR videos have better perceptive experiences than adapted videos generated by the adaptation system for high-end PC users. The separation of attention information detection and adaptation operation has two benefits. First, since the detection process needs to be performed only once and the detected attention information can be used for all users, the detection process can be moved to the video content generation server and the workload of this new adaptation system will be reduced greatly while the system still remains flexible. Now the adaptation system only needs to complete the function in transforming HR videos into LR videos. This will facilitate the implementation of real-time adaptation process for mobile devices. The second benefit of the proposed scheme is that we can actually improve the video adaptation performance by fully utilizing the predetected attention information. This will be described in detail in the next section.

3. VIDEO CONTENT GENERATION

In order to produce video bitstreams with embedded attention information, we need to integrate two modules of *attention detector* and *SEI* into the traditional video coding structure as shown in Figure 2. During video compression, each original frame and its motion information acquired from the module of motion estimation (ME) will be input into the module of *attention detector*. A group of attention areas within the frame will be detected as attention objects. Then the attention information will be encapsulated in the SEI module and embedded into video bitstreams. Another added module is *QP adjustment* which controls the encoding QP for attention and nonattention areas, respectively. Based on the detected attention information, we propose an approach of adaptive QP adjustment for attention and nonattention areas. The details of the three modules will be introduced respectively in the following subsections.

3.1. Visual attention modeling

In this subsection, we present a visual attention model [11] to reveal the regions that attract the user's attention in each video frame. The attention detector adopts attention objects (AOs) defined in (1) as the information carriers:

$$\text{definition1}: \{\text{AO}_i\} = \{(\text{SR}_i, \text{AV}_i, \text{MPS}_i)\}, \quad 1 \leq i \leq N. \tag{1}$$

Each AO owns three attributions: SR, AV, and MPS. SR is referred to as a spatial region corresponding to an AO. The attention value (AV) indicates the weight of each AO in contribution to the information contained in the image. Since the delivery of information is significantly dependent on the dimension of presentation, minimal perceptible size (MPS) is introduced as an approximate threshold to avoid excessively subsampling during the reduction of display size.

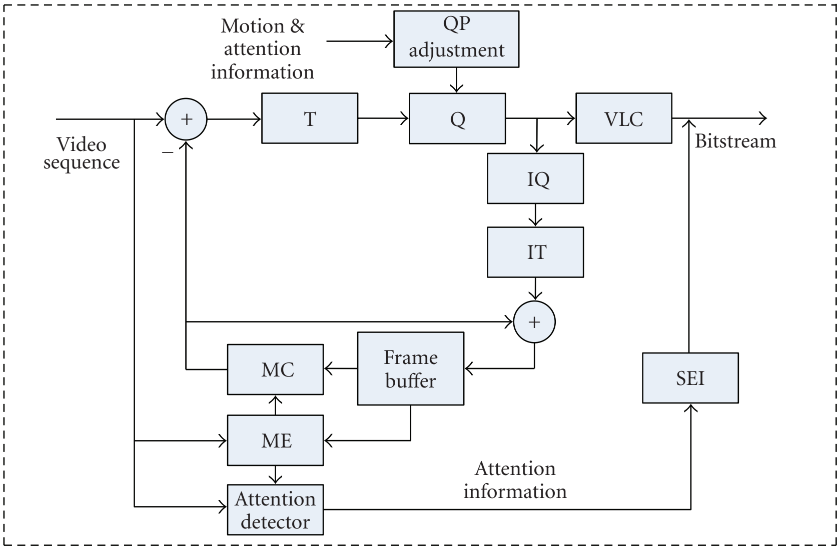


Figure 2: The block diagram of video content generation.

Accordingly, three attributions of AOs will be measured by an automatic modeling method. Up to now, four types of attention objects are taken into account in our model: motion objects, face objects, text objects, and saliency objects.

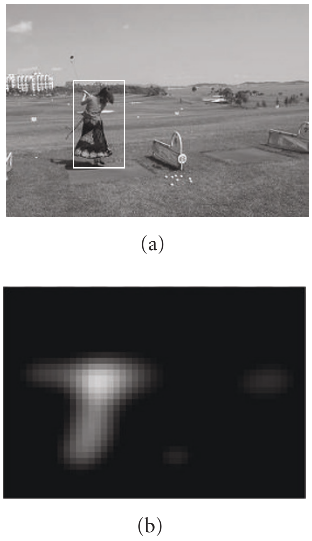
It is different from the modeling of static pictures in that moving parts in a video are usually noticeable. In our implementation, video sequences are stored in H.264 format and the motion vector field (MVF) of a frame can approximately measure the motion information:

$$I(i,j) = \sqrt{dx_{i,j}^2 + dy_{i,j}^2}, \tag{2}$$

where $(dx_{i,j}, dy_{i,j})$ denote two components of the motion vector. We consider $I(i,j)$ as an intensity map and employ following image processing methods to determine the SR attribution of a motion object. Firstly, we adopt median filter to remove the noise and then adjust the map by the histogram equalization. Several seeds points are chosen to get some larger segmented regions by the region growing method. We regard these regions as SRs of motion AOs. The AV of a motion object is estimated by its size, spatial/temporal coherence, and motion intensity. It is based on the assumption that an object with larger magnitude, greater motion intensity, or more consistent motion will be more important:

$$\text{AV}_{\text{motion}} = \text{Area}_{\text{motion}} \times W_{\text{motion}}^{\text{intensity}} \times W_{\text{motion}}^{\text{coherence}}. \tag{3}$$

Early stages of attention processing are deployed by ensemble of low-level features such as contrast, orientation, and intensity. Due to the heavy computations of the traditional saliency model, we adopt a contrast-based model [12] to produce the saliency map and determine the attention-getting areas. An example image and its saliency map are shown in Figure 3.


(a)

(b)

Figure 3: An example of saliency detection.

The AVs of saliency objects are calculated as

$$\text{AV}_{\text{saliency}} = \sum_{(i,j\in R)} B_{i,j} \cdot W_{\text{saliency}}^{i,j}, \tag{4}$$

where $B_{i,j}$ denotes the value of pixel (i,j) in the saliency map. Since people often pay more attention to the region near the center of an image, a normalized Gaussian template centered at the image is used to assign the position weight $W_{\text{saliency}}^{i,j}$.

Some heuristic rules are employed to calculate the MPS of above two types of AOs. For example, bigger regions can be scaled down more aggressively than smaller ones.

In addition, faces and texts often carry the semantic information, which users expect and can be detected with good accuracy currently. Face objects and text objects are defined

attention_information (PayloadType, PayloadSize) {	Descriptor
attention_object_number	ue(v)
if (attention_object_number > 0) {	
for (i = 0; i < attention_object_number; i++ {	
attention_value	ue(v)
left_top_x	ue(v)
left_top_y	ue(v)
right_bottom_x_minus_left_top_x	ue(v)
right_bottom_y_minus_left_top_y	ue(v)
}	
}	
}	

Figure 4: Proposed SEI structure for attention information.

and generated in the similar way as the work in [11]. In our solution, a simple face detection algorithm is adopted mainly to detect the frontal faces in video frames. In order to decrease the computational cost, we carried out the face detection every three frames. A fast text detection algorithm based on [13] is employed to find text regions in each video frame. AVs of face and text objects are estimated by their sizes and the MPS values are predefined. In order to combine different types of AOs into a unified attention model, the AV of each AO is normalized to (0,1) and the final AV is computed as

$$\mathrm{AV}_i = w_k \cdot \frac{\mathrm{AV}_i^k}{\sum_i \mathrm{AV}_i^k}, \quad (5)$$

where AV_i^k represents the AV of AO_i detected in model k and w_k is the weight of model k, for example, face model, text model, or motion model, which manifests the contribution during the attention-guiding function. In our system, motion objects are considered most attention-getting and semantic objects play a more important role than saliency objects.

3.2. SEI-based attention information

Through attention detection, the information of attention objects in frames has been acquired. We intend to find a solution to embed the attention information into bitstreams for the future adaptation operation. There are two basic requirements for such a solution. One requirement is that the video bitstream with embedded attention information should still conform to the video coding standard. For clients who do not need any adaptation operation, the embedding will be transparent, and thus will not cause any burden while decoding the bitstream. The other requirement is that the embedded attention information can be easily extracted and conveniently used in the adaptation process. This means that the embedding should introduce minimum additional computational complexity and negligible overhead. The tool of SEI [9] in H.264 is a good solution. SEI is a technique developed in H.264 standard. It can take some auxiliary information and will assist in the processes related to decoding, display, or other purposes of video signals. Several SEI messages have been defined in H.264 for special purposes, such as spare picture, buffering period, and picture timing. More details of SEI messages can be found in [9]. SEI can perfectly meet the two requirements mentioned above. So, we adopt it as the carrier of the information of attention objects in our scheme. The key work is to design a special SEI message designed to signal the attention information, and utilize it in the adaptation process. The proposed SEI structure is shown in Figure 4.

The SEI message is designed for the attention information of only one frame and it can be stored closely with the related frame, which will make the use of the attention information flexible. The message includes several items: *attention_object_number, attention_value, left_top_x, left_top_y, right_bottom_x_minus_left_top_x*, and *right_bottom_y_minus_left_top_y*, for the desired attention information according to the attention model defined above. They are all efficiently coded by ue(v) which is unsigned integer Exp-Golomb-coded syntax element and can be easily decoded to recover the information of attention objects for the adaptation operation. The meanings of these items are explained as follows.

attention_object_number: the number of attention objects in a frame; *attention_value*: the attention value of an object; *left_top_x, left_top_y, right_bottom_x_minus_left_top_x* and *right_bottom_y_minus_left_top_y*: the coordinates of each attention object. It should be noted that the values of the latter two terms are the differences between the left-top point and the right-bottom one.

3.3. Adaptive QP adjustment for balanced encoding and transcoding

Since original sequences are unavailable in the transcoding process, reconstructed frames after decoding are generally regarded as the input video. The better the quality of the reconstructed frames is, the higher coding performance the transcoder will achieve. In this research, the new video sequence generated by the transcoding and suitable for the target mobile device will consist of mostly attention areas in each frame, instead of the entire frame. Therefore, if we can perform an attention-biased bit allocation strategy in the original encoding within a given frame, we will be able to improve the quality of transcoded videos. That is to say, if it is known to us that the majority of the clients are mobile device users, we can move some bits allocated for nonattention areas to attention areas when we encode the original video frame.

We expect that, at the same bitrate, the quality of attention areas will be improved if we apply the attention area-aware bit allocation strategy. Since the true display size of mobile devices is known only when the video server receives the request from the mobile users, the information is unavailable at the time of encoding original video. In this case, the attention area we extract in each frame will need to be the one that covers all attention objects in the whole frame rather than the one which is restricted by the true display size which will keep the flexibility and improve the quality of transcoded videos for various mobile devices with different display size. The cost is that the improvement will be lowered in the case that the real display size is smaller than the size of the maximal attention area.

We have also carried out some preliminary bit allocation experiments in an attempt to uncover the relationship between the performance of bit allocation and motion characteristics of the video sequences. We found that excessively increasing the amount of bits allocated for attention areas will cause obvious coding performance loss especially for the video sequences with low motion. One of the reasons is that attention-biased bit allocation causes a frame to generate significant difference in quality between attention and nonattention areas. This will bring about a negative influence on the motion compensation performance of the next frame. The other reason is that the overhead of signaling different quantization strategy for attention and nonattention area may become substantial.

In our preliminary experiments, we also observed that, for high-motion sequences, the coding performance loss is negligible since traditional coding of such sequences also allocates more bits to high-motion areas within each frame. Since the attention areas are often have high motion, the influence caused by the attention-biased bit allocation strategy on the next frame is less than that of low-motion sequences. At the same time, the overhead of signaling the different quantization strategy is negligible comparing to the amount of bits for high-motion sequences.

Based on the preliminary analysis and experimental results, we can draw a valuable conclusion that increasing the bits for attention area may cause noticeable loss for low-motion frames but negligible loss for high-motion frames. The loss of the encoding performance is undesired even though the quality of the video transcoding can be improved. Since the clients consist of not only mobile devices users but also PC users and the adaptation operation is unnecessary for PC users, the apparent encoding performance loss will impair the perceptual experience of the high-end PC users. As a result, the key problem in this research on attention-biased bit allocation strategy is how to achieve a fine balance between video encoding performance and video transcoding performance. That is to say, we must improve the transcoding performance as much as possible while maintaining the high encoding performance. This problem can be formulated as follows:

$$r_{\text{best}} = \underset{r_i \in R}{\arg\min}\, \Delta D_{\text{encoding}}(r_i) + \alpha \cdot f_{\text{trascoding}}(\Delta D_{\text{attention}}(r_i)), \tag{6}$$

where r_i denotes the amount of bits moved from nonattention areas to the attention areas, R is the set of available adjusting rates, $\Delta D_{\text{encoding}}(r_i)$ is the increased distortion in encoding caused by r_i, $\Delta D_{\text{attention}}(r_i)$ denotes the improvement of the attention area quality, $f_{\text{trascoding}}(\Delta D_{\text{attention}}(r_i))$ indicates the transcoding gain from the improved quality of the attention area, and α is a parameter used to balance the performance between encoding and transcoding. Optimal bit allocation strategy r_{best} can be computed according to (6).

It is computationally complicated and unnecessary to accurately model these relationships presented in (6), such as $\Delta D_{\text{encoding}}(r_i)$, $\Delta D_{\text{attention}}(r_i)$, and $f_{\text{trascoding}}(\Delta D_{\text{attention}}(r_i))$. Based on the results from our preliminary study, we propose a frame-level adaptive QP adjusting approach according to the motion characteristics of a given frame. This is also consistent with our analysis on the impact of motion activity on the bit allocation strategy.

In the process of encoding video sequences, the encoder searches for optimal motion vectors for each macroblock in the given frames during motion estimation. Motion vectors will indicate the displacement of one macroblock relative to the reference frame. The more complicated motion the frame may have, the greater the values of motion vectors are. Therefore, we can measure the motion characteristics of a frame by its motion intensity defined as follows:

$$I = \frac{1}{MN} \sum_{j=0}^{M-1} \sum_{i=0}^{N-1} \sqrt{(mvx_{j,i})^2 + (mvy_{j,i})^2}, \tag{7}$$

where M and N are the height and width of a frame, $mvx_{i,j}$ and $mvy_{i,j}$ are two components of the motion vector for the pixel (i, j). Here, we assume pixel (i, j) has the same motion vector as the macroblock to which it belongs. A frame can be classified into three types: *high, medium, and low*, by its motion intensity I. However, since motion information is unavailable before encoding, we may adopt the motion intensity of previous frame to predict the type of motion in the current frame,

$$\text{type}_i = \begin{cases} \text{high} & I_{i-1} > T_h, \\ \text{medium} & T_l < I_{i-1} < T_h, \\ \text{low} & I_{i-1} < T_l, \end{cases} \tag{8}$$

where T_h and T_l are two thresholds. It is well known in the video coding community that, for a frame, the rate-distortion optimization (RDO) adopted by H.264 will affect

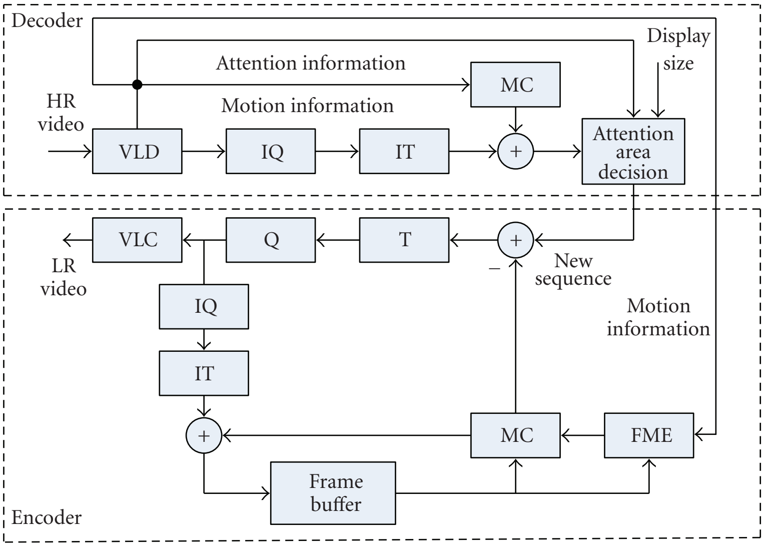


Figure 5: The structure of video adaptation system.

the motion intensity when the video is coded at different bitrates. Therefore, T_h and T_l may vary with different bitrate. Because all videos are encoded at high bitrate in the proposed application scenario, we can train the two thresholds using small QP (such as 20). Extensive experiments based on various training sequences have led to the conclusion that these two parameters may be set as $T_h = 10$ and $T_l = 3$.

Given any type of video frame, the encoding QP for attention areas will be adjusted by

$$\mathrm{QP}_{\text{attention}} = \begin{cases} \mathrm{QP} - 3 & \text{type} = \text{high}, \\ \mathrm{QP} - 2 & \text{type} = \text{medium}, \\ \mathrm{QP} - 1 & \text{type} = \text{low}, \end{cases} \tag{9}$$

where QP is the quantization parameter for nonattention areas. The QP adjustment is large for high-motion frame and small for low-motion frame. Such a heuristic approach based on motion activity classification captures the essential objective of the optimization as shown in (6).

4. VIDEO ADAPTATION SYSTEM

When the HR video bitstreams containing attention information are sent to mobile users over wireless channel, our adaptation system placed on a proxy or server will perform adaptation manipulation on the videos. That is, it will transform the HR videos into LR videos to meet the constraint from the limited display size. To generate a new bitstream, a direct method is to re-encode each new frame of the sequence, which is referred to as the cascade of decoder and encoder scheme. But the computational complexity with such scheme is too high to be applicable to such a system. Instead, we adopt an effective transcoding approach developed in our earlier work [10] to solve this problem.

In order to make the proposed scheme more comprehensible to the readers, we give a simple introduction of the adaptation system though it is not the main work of this paper. As shown in Figure 5, the adaptation system is composed of a decoder and encoder. Firstly, the decoder decodes motion information, reconstructed frames, and attention information from the HR bitstream. The module of *attention area decision* will decide the most attractive area which contains as much attention information as possible while meeting the constraint from the display size of mobile devices. In our scheme, the size of attention area in a frame is constrained to a set of specific sizes rather than arbitrary sizes, for example 352 × 288 pixels for CIF, 176 × 144 pixels for QCIF, and so forth, in order to simplify the development of fast transcoding method and to guarantee the transcoding performance. We define the attention area as a rectangle, whose size can be adaptively adjusted to the nearest possible value in the specific sizes according to different display sizes of mobile devices. It should be noticed that the size of attention area is fixed in a certain video sequence in the present solution. A branch and bound searching algorithm [11] is utilized to crop the rectangular attention area in a frame. If the size of the cropped attention area is not equal to the predetermined specific size, we will clip or expand the edges from the center of attention area to meet the constraint. In order to avoid the jittery results caused by producing these regions directly, the technique of virtual camera control [14] is adopted in our system to adjust positions of the cropped regions. After cropping from each reconstructed frame, the attention areas will be assembled into a new LR sequence and input into the encoder. By making use of original motion information, the module of ME (motion estimation) in the encoder can be replaced by the module of FME (fast motion estimation). A fast-mode decision algorithm [10] is adopted in FME. Because attention area decision and fast transcoding algorithm have been developed in our earlier work, we will not give the details of them in this paper. More details can be found in [10]. After transcoding, the LR video will be sent to mobile device users for browsing.

5. EXPERIMENTAL RESULTS

In this section, we present some experiments to evaluate the performance of the attention-information-based spatial adaptation framework as it compares with our previous framework. The system has been implemented based on the reference software of H.264, jm61e [15]. Eight standard sequences, Foreman, Coastguard, Football, Table, Tempete, Paris, Stefan, and Mobile (CIF 90 frames, 30 Hz), are selected as the test sequences.

Table 1: Overhead of attention information.

Sequences	Original bitrate (kb/s)	Attention information bitrate (kb/s)	Percent
Foreman	512	2.5	0.49%
Table	1024	6.6	0.64%
Coastguard	1024	3.1	0.30%
Tempete	1024	3.6	0.35%
Paris	1024	0.7	0.09%
Stefan	1024	7.3	0.71%
Mobile	2048	3.4	0.17%
Football	2048	11.4	0.56%

5.1. Overhead of attention information

Attention information is additional information in bitstreams and may be regarded as a sort of overhead. We calculate the bits spent on attention information encapsulated in SEI structure and show them in Table 1. Original bitrate refers to bits for motion and residue. As shown in Table 1, comparing with the amount of bits for motion and residue, the overhead of attention information is negligible. It will cause little coding performance loss in video content generation.

5.2. Performance comparison with and without QP adjustment

It is anticipated that the QP adjustment will not degrade the encoding performance. Given the thresholds $T_h = 10$ and $T_l = 3$, the eight sequences are encoded with several QPs. The encoding results of adaptive QP adjustment have been shown in Figure 6, and compared with that of without QP adjustment, that is, fixed QP encoding. For sequences of Table, Football, Tempete, Paris, Stefan, and Mobile, there are virtually no performance loss with our approach while for the sequences of Foreman and Coastguard, the performance loss has been consistently less than 0.2 dB. The results prove that adaptive adjusting QP for attention area has little effect on the encoding performance.

5.3. Comparison of transcoding performance

We anticipate that the QP adjustment will improve the transcoding performance since the attention areas are encoded with higher quality. We compare the proposed attention-information-based adaptation framework with QP adjustment against our previous framework without such adjustment [10]. In this experiment, eight sequences are firstly encoded at different bitrates: Foreman 512 kb/s, Coastguard 1024 kb/s, Table 1024 kb/s, Tempete 1024 kb/s, Paris 1024 kb/s, Stefan 1024 kb/s, Mobile 2048 kb/s, and Football 2048 kb/s. For the proposed framework in this research, attention information is included in bitstreams and adaptive QP adjusting has been performed. We apply the same transcoding algorithm for both frameworks. Without loss of generality, we may set the target display size as QCIF. Then the original CIF bitstreams are transcoded into QCIF bitstreams containing attention areas at a variety of bitrates. In order to calculate PSNR, we extracted attention areas of original videos and regarded them as the original version of new sequences. Then PSNR can be calculated between the original version and the transcoded videos. As shown in Figure 7, comparing with previous framework, the proposed framework in this research is able to obtain R-D (rate distortion) performance improvement at all bitrates. Especially for the video sequence Foreman at high bitrate, the gain can be up to 0.5 dB. For Paris sequence, the improvement is not obvious. That is because paris is low motion so that our QP adjustment algorithm has little improvement on quality of its attention areas.

5.4. Subjective testing

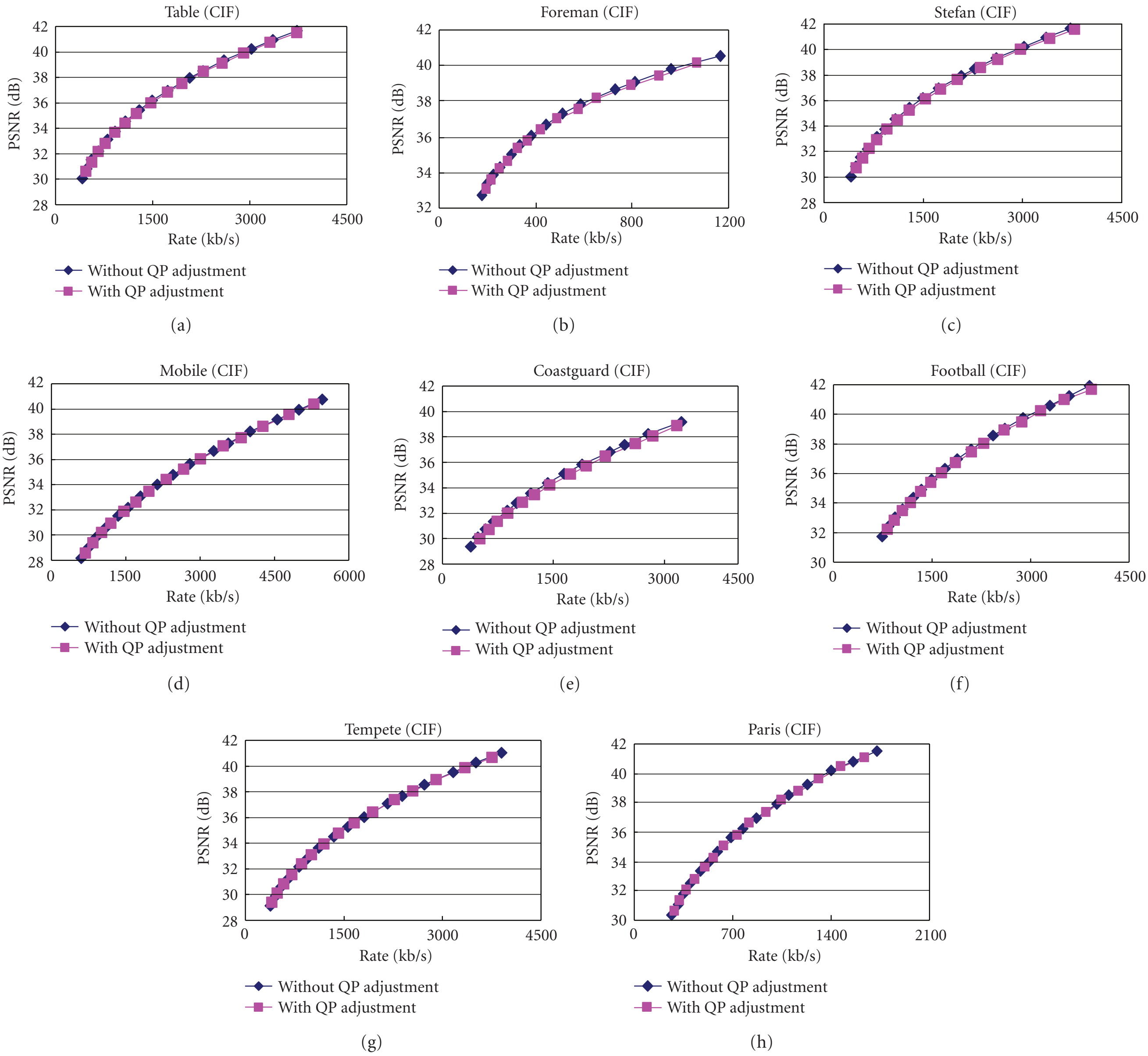
In order to evaluate the perceptual impression of the output videos of our adaptation system, a user study has been carried out. Six volunteers who have no knowledge of our system were invited to score for two subjective assessment questions. These two questions are as follows.

Question 1. Compared to the original sequence, is the output region of our system the one you are interested in? (4-definitely, 3-mostly, 2-possibly, 1-rarely).

Question 2. Do you think the visual quality of the result is acceptable for small displays? (3-good, 2-fair, 1-poor).

The average scores have been shown in Table 2.

From the user study experiments, we can learn that with our adaptation scheme, the perceptive experiments of browsing videos via mobiles have been improved. Viewers obtain most attention information from the LR videos of our adaptation system. The scores of Mobile and Paris for Question 1 is lower than those of other sequences. That is because the two sequences have multiple moving objects in one frame. Due to the constraint of the display size, the cropped areas by our system cannot cover all attention objects, which results in the lower scores.



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

FIGURE 6: Encoding performance comparison between with and without QP adjustment.

5.5. *Visual quality comparison*

In this research, we expected that the visual quality of the video adaptation will also be improved compared with previous ROI-based scheme. Figure 8 gives an example of visual quality comparison between three methods, downsizing, without QP adjustment, and with QP adjustment, which consists of some frames from Coastguard sequence. For fair comparison, the outputted bitstreams of the three methods are at the same bitrate. The first line is the results of simple downsizing. By directly downsizing video sequences from CIF into QCIF, videos are adapted to the display size. However, the details in frames, for example, the boat and the man, are too small to be recognized. The second line is the results of our previous framework in [10] with adaptation to attention regions. The results of this research are shown in the third line. Comparing with downsizing method, our algorithm can supply more attention information and better perceptive experiences. Comparing with our previous framework, the adaptive QP adjustment based on attention information is able to further improve the visual quality of attention areas as shown in Figure 8.

6. CONCLUSION

In this paper, a novel video adaptation solution has been developed to overcome the constraint from limited display sizes of mobile devices in video browsing and the limited

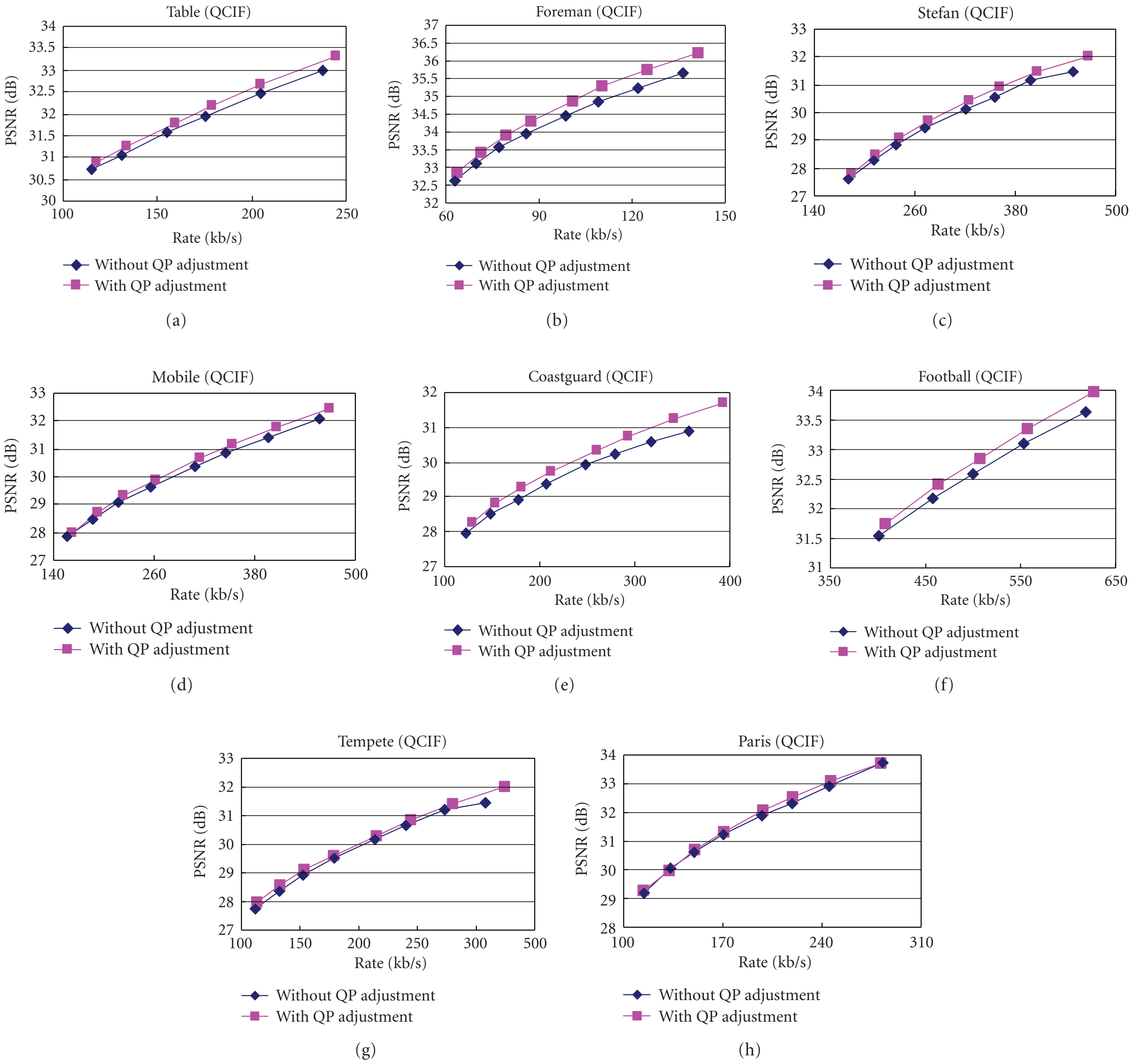


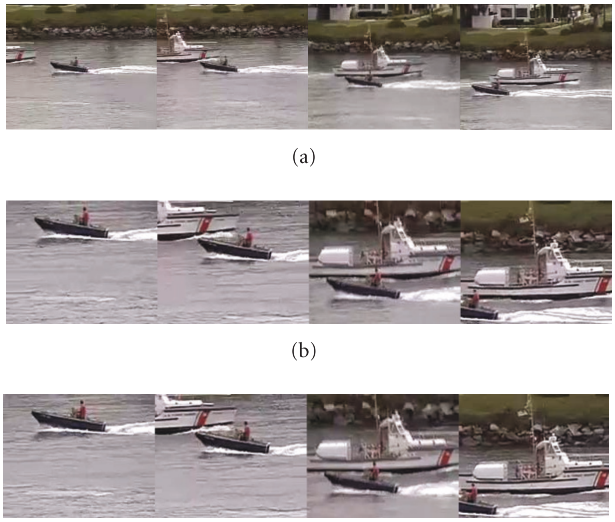
Figure 7: Transcoding performance comparison between with and without QP adjustment.

computational resource at proxy servers. The adaptation framework helps mobile users to gain better visual perception experiences when enjoying video browsing over wireless channels. When generating bitstreams, a visual attention model is utilized to detect the most informative regions in each frame, which is referred to as attention objects. Then the information of attention objects including positions and attention values will be encoded, encapsulated with the proposed SEI structure, and embedded into bitstreams. The attention information will then be used in the adaptation system to generate a bitstream of attention areas in each frame to adapt to the display sizes of mobile devices. More importantly, we have developed an innovative attention-biased QP adjustment scheme based on the detected attention information to accomplish the bit allocation between attention areas and overall frames. In this way, we can achieve a balance between the encoding performance and transcoding performance.

The contributions of this research lie in three important aspects. First, the shift of the complexity from proxy to video generation server enables the proxy to provide better real-time applications since there is no need to generate the ROI at the proxy. Second, the design of encapsulation of attention information with proposed SEI structure enables the

Table 2: Subjective testing results.

Sequence	Score of Question 1	Score of Question 2
Table	3.67	2.5
Foreman	4	3
Stefan	4	2.67
Mobile	2.33	2.33
Coastguard	3.67	2.83
Football	3.5	2.67
Tempete	3.33	2.5
Paris	2.33	2.5
Average	3.35	2.63



(a)

(b)

(c)

Figure 8: Subjective quality comparison (a) downsizing, (b) without QP adjustment, (c) with QP adjustment.

embedding of the side information into the standard compliant H.264 compressed video. Third, the embedded attention information has been utilized for adaptive QP adjustment to improve the video transcoding performance while maintaining the overall encoding performance for high-end PC users.

Extensive experiments have been carried out to demonstrate that both subjective and objective quality improvements have been noticeable comparing with the passive approach we have developed in our earlier study [10]. The improvements are significant when comparing with the simple downsizing method. However, for video sequences with low motion, our QP adjustment algorithm has resulted in little improvement. In our future research, we will explore new bit allocation method to improve the quality of attention areas while maintaining the coding performance for those low-motion video sequences.

ACKNOWLEDGMENT

This work is supported by NSFC General Program under Contract no. 60672161, 863 Program under Contract no. 2006AA01Z317, and NSFC Key Program under Contract no. 60632040.

REFERENCES

[1] J.-G. Kim, Y. Wang, S.-F. Chang, and H.-M. Kim, "An optimal framework of video adaptation and its application to rate adaptation transcoding," *ETRI Journal*, vol. 27, no. 4, pp. 341–354, 2005.

[2] 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.

[3] J. Xin, C.-W. Lin, and M.-T. Sun, "Digital video transcoding," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 84–97, 2005.

[4] 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.

[5] A. Sinha, G. Agarwal, and A. Anbu, "Region-of-interest based compressed domain video transcoding scheme," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '04)*, vol. 3, pp. 161–164, Montreal, Canada, May 2004.

[6] G. Agarwal, A. Anbu, and A. Sinha, "A fast algorithm to find the region-of-interest in the compressed MPEG domain," in *Proceedings of the International Conference on Multimedia and Expo (ICME '03)*, vol. 2, pp. 133–136, Baltimore, Md, USA, July 2003.

[7] A. Vetro, H. Sun, and Y. Wang, "Object-based transcoding for adaptable video content delivery," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 3, pp. 387–401, 2001.

[8] K. B. Shimoga, "Region of interest based video image transcoding for heterogeneous client displays," in *Proceedings of the 12th International Packetvideo Workshop (PV '02)*, Pittsburgh, Pa, USA, April 2002.

[9] "Draft ITU-T recommendation and final draft international standard of joint video specification (ITU-T Rec. H.264/ISO/IEC 14496-10 AVC," in Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, JVT-GO50, 2003.

[10] Y. Wang, H. Li, X. Fan, and C. W. Chen, "An attention based spatial adaptation scheme for H.264 videos on mobiles," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 20, no. 4, pp. 565–584, 2006, special issue on Intelligent Mobile and Embedded Systems.

[11] L.-Q. Chen, X. Xie, X. Fan, W.-Y. Ma, H.-J. Zhang, and H.-Q. Zhou, "A visual attention model for adapting images on small displays," *Multimedia Systems*, vol. 9, no. 4, pp. 353–364, 2003.

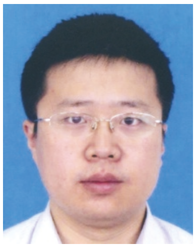
[12] Y.-F. Ma and H.-J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," in *Proceedings of the 11th ACM International Multimedia Conference (MM '03)*, pp. 374–381, Berkeley, Calif, USA, November 2003.

[13] X.-S. Hua, X.-R. Chen, L. Wenying, and H.-J. Zhang, "Automatic location of text in video frames," in *Proceedings of the ACM International Multimedia Information Retrieval Conference (MIR '01)*, pp. 24–27, Ottawa, Canada, October 2001.

[14] X. Fan, X. Xie, H.-Q. Zhou, and W.-Y. Ma, "Looking into video frames on small displays," in *Proceedings of the 11th ACM International Multimedia Conference (MM '03)*, pp. 247–250, Berkeley, Calif, USA, November 2003.

[15] "JVT reference software official version," Image Processing Homepage, http://bs.hhi.de/∼suehring/tml/.

Houqiang Li received the B.S., M.S., and Ph.D. degrees in 1992, 1997, and 2000, respectively, all from the Department of Electronic Engineering and Information Science (EEIS), University of Science and Technology of China (USTC), Hefei, China. From November 2000 to November 2002, he did postdoctoral research in Signal Detection Lab, USTC. Since December 2002, he has been on the faculty of the Department of EEIS, USTC, where he is currently an Associate Professor. His current research interests include image and video coding, image processing, and computer vision.



Yi Wang received the BE degree from the Electronic Engineering and Information Science (EEIS) Department, University of Science and Technology of China (USTC), in 2002. Currently, he is working toward the Ph.D. degree in the EEIS Department of USTC. He worked as a Research Intern at Microsoft Research Asia from 2005 to 2006. His research interests include image and video compression, video transmission, and video adaptation techniques.

Chang Wen Chen received the B.S. degree from the University of Science and Technology of China (USTC), in 1983, the M.S.E.E. degree from the University of Southern California in 1986, and the Ph.D. degree from the University of Illinois at Urbana-Champaign, in 1992. He is Allen S. Henry Distinguished Professor in the Department of Electrical and Computer Engineering Florida Institute of Technology, since July 2003. He is also Grand Master Chair Professor of the USTC since 2006. Previously, he was on the Faculty of Electrical and Computer Engineering at the University of Missouri-Columbia, from 1996 to 2003, and at the University of Rochester, from 1992 to 1996. From September 2000 to October 2002, he served as the Head of the Interactive Media Group at the David Sarnoff Research Laboratories in Princeton, NJ. He has also consulted with Kodak Research Labs, Microsoft Research, Mitsubishi Electric Research Labs, and NASA. He has been the Editor-in-Chief for IEEE Trans. Circuits and Systems for Video Technology since January 2006. He has been an Editor for several IEEE Transactions and international journals. He was elected an IEEE Fellow for his contributions in digital image and video processing, analysis, and communications and an SPIE Fellow for his contributions in electronic imaging and visual communications.