

An Attention Based Spatial Adaptation Scheme for H.264 Videos on Mobiles

Yi Wang¹, Xin Fan¹, Houqiang Li¹, Zhengkai Liu¹, Mingjing Li²

¹*Dept. of EEIS, Univ. of Sci. and Tech. of China, Hefei, 230027, P.R. China*

²*Microsoft Research Asia, 5F Sigma Center, No. 49, Zhichun Road, Beijing, 100080, P.R. China*

¹{wy1979, van}@mail.ustc.edu.cn, {lihq, zhengkai}@ustc.edu.cn, ²mjli@microsoft.com

Abstract

When browsing videos in mobile devices, people often feel that resolution greatly affects their perceptual experience in the limited screen size. In this paper, an attention based spatial video adaptation scheme is proposed to overcome display constraints by producing the region of interest. According to the size of the target display, we automatically detect and crop the informative region in each frame to generate a smooth sequence. To avoid costly fully encoding operations, we employ a set of transcoding techniques based on the H.264 standard. Experimental results show that this approach not only improves the perceptual quality but also saves the bandwidth and computation, especially for the videos which are not well edited.

1. Introduction

Mobile devices are becoming more and more pervasive in our daily lives. Besides the traditional communicating manners of voice and text, images and videos are broadly utilized for communication and amusement. Especially with the increasing popularity of embedded cameras, people can not only enjoy videos anytime, but also share the scene where he is standing with others. When browsing videos in mobile devices, people often feel that the display resolution is still a critical factor affecting the browsing experiences although the limitations of network bandwidth and computation power are greatly improved recently. There are strong needs for mobile users to access videos in better quality.

Although there has been much effort on video adaptation for mobiles, majority of these approaches focus on bit rate reduction and speed-up delivery. In recent years, some video transcoding schemes [14] are proposed to achieve the spatial resolution reduction. These traditional spatial transcoding schemes are usually based on spatial downsizing by the factor of two or other integer factors. Considering perceptual results, excessive resolution

reduction may cause much loss of desired information. Accordingly, authors of [4] propose an automatic modeling approach to crop a most informative region within the original image and assign the cropped region a proper scaling ratio according to a certain display size. In a similar way, an attention based video adaptation solution [5] is established for better perceptual results in small displays. However, they adopt the pixel-domain approach that original videos are decoded and then fully encoded after processing. It is often very costly for a large amount of process in practice. Therefore, we address this problem by a spatial video transcoding approach. It is more efficient than fully encoding the generated sequence and only with the tiny impact of quality. Compared to the previous video transcoding solutions to spatial resolution reduction, we only manipulate the attention-getting region, which provides better perceptual experiences and saves the bandwidth at the same time.

In view of efficient coding performance and promising applicability of the latest H.264/AVC standard [1] in mobile video terminals, our system is specifically performed for videos in the H.264 format. Our approach is to first determine most informative regions in decoded frames by an attention based modeling method. Since organizing these cropped regions directly may cause the jittery results, we adjust positions of the cropped regions by a virtual camera control [5][11] to smooth the outputs. Subsequently, the original video sequence is transcoded according to the selected regions through processes of motion vector adjusting, mode decision and drifting error removing. Since only regions of interest are delivered to the user, there are considerable improvements on bitstream reduction and computational complexity. On the other hand, users also gain the satisfying browsing experiences. The improvements will be demonstrated in the experiments in Section 5.

2. Background of H.264 video coding standard

The newest video coding standard, H.264, has gained more and more attention recently, mainly due to its high

coding efficiency and minor increase in decoder complexity. The H.264 standard achieves much higher coding efficiency than previous video coding standards by using a collection of new encoding tools, however, at a cost of significant increase in encoding complexity. Among all modules in the encoder, motion estimation (ME) and mode decision contribute most of the complexity. H.264 supports various block sizes for ME. It uses tree-structured hierarchical macroblock (MB) partitions. Inter-coded 16x16 pixel MB can be broken into MB partitions of sizes 16x8, 8x16, or 8x8. 8x8 partitions are also known as sub-macroblock. Sub-macroblocks can be further broken into sub-macroblock partitions of sizes 8x4, 4x8, and 4x4. To simplify our description, we will call these block types (or mode) as 16x16, 16x8, 8x16, 8x8, 8x4, 4x8, 4x4, where 8x8, 8x4, 4x8 and 4x4 is a partition of sub-macroblock 8x8.

Assuming that we have M block types and the search range for each block type is the same and equal to $\pm W$, this will imply that we need to examine $M \times (2W+1)^2$ positions compared to only $(2W+1)^2$ positions for a single block type. If N reference frames are used in ME, the checked positions will be $N \times M \times (2W+1)^2$. It is obvious that ME in H.264 is much more complicated than that in previous standards. Therefore, significant efforts were placed in reducing the computational cost of ME while at the same time retaining quality. Many algorithms have been proposed such as 3-step search (3SS) [6], New 3SS (NTSS) [9], Predictive Motion Vector Field Adaptive Search Technique (PMVFAST) [13], Enhanced Predictive Zonal Search (EPZS) [12], etc. These algorithms have much reduced the encoding complexity with slight loss of the quality.

In recently years more attention has been paid on fast mode decision for its more significant contribution to the complexity reduction. Some algorithms [7][8][15] have been proposed to speed up the process of mode decision. In general, these algorithms all reduce complexity by eliminating some block types in ME, though the omitting strategies differ. These algorithms can decrease the number of checked block types in ME, but there are still several block types needed to be checked. We notice that motion information is available in transcoding process. Since motion information can reflect the activity of each MB, it is possible to further reduce complexity by utilizing motion information. In our transcoding scheme, we decide block type for each MB directly according to the original motion information without ME for each mode. It will be shown in Section 6 that this method reduces the complexity greatly.

3. System architecture

As shown in Figure 1, our system consists of three main modules: *decoder*, *attention area extractor* and *transcoder*. The module of *decoder* is to decode the high-resolution

(HR) bitstream of H.264. Decoded information will be transmitted to the module of *attention area extractor* and *transcoder*. The module of *attention area detector* includes several sub-modules: motion, face, text and saliency detectors to reveal the attention objects and the combiner to output smooth attention areas for the following *transcoder*. Under the constraint of video coding standard, the size of the attention area is specified according to a certain display resolution on mobiles in our system. Based on the output areas, the last module, *transcoder*, will produce the low-resolution (LR) bitstream conformed to H.264 standard. It is composed of three sub-modules: mode decision, motion vectors adjusting and drifting error compensation. We will give details of each module respectively in Section 4.

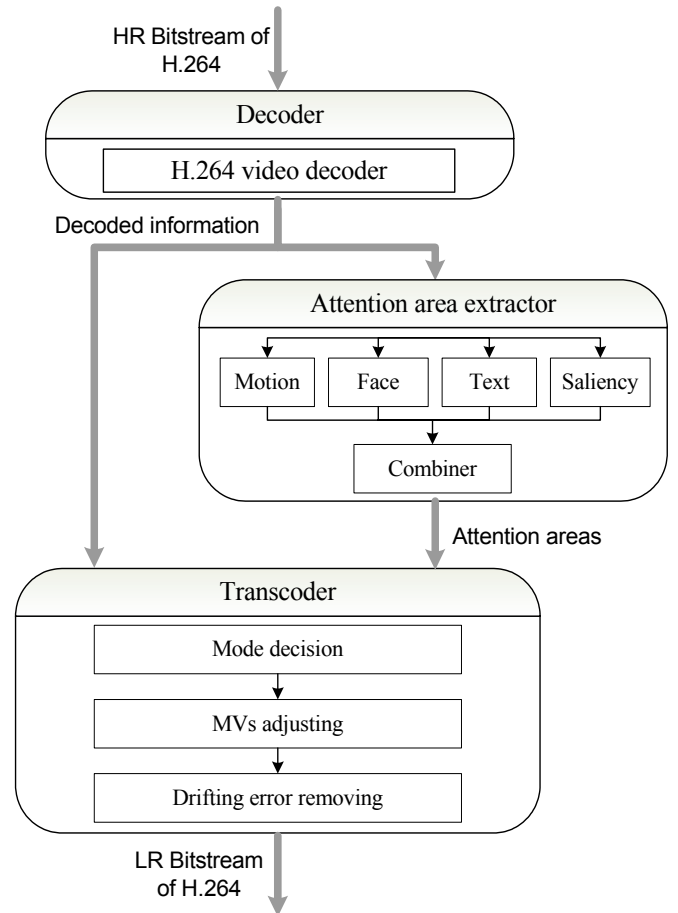


Figure 1. Flowchart of the video adaptation scheme

4. Visual attention modeling

In this section, we adopt a visual attention model [4] to reveal the region attracting the user's attention in each decoded video frame.

4.1 Visual attention model

As shown in Definition 1, a set of information carriers – attention objects (AOs) [5] are defined in our method.

$$Definition1: \{AO_i\} = \{(SR_i, AV_i, MPS_i)\}, 1 \leq i \leq N \quad (1)$$

Each AO owns three attributions: SR , AV and MPS . SR is referred 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 sub-sampling during the reduction of display size.

4.2 Automatic Attention-based Modeling

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

It is different from the modeling of static pictures [4] that moving parts in a video are usually noticeable. In our implementation, video sequences are stored in H.264 format and the approximate motion information can be measured by the Motion Vector Field (MVF) of a frame:

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

where $(dx_{i,j}, dy_{i,j})$ denote two components of 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:

$$AV_{motion} = Area_{motion} \times W_{motion}^{intensity} \times W_{motion}^{coherence} \quad (3)$$

For early stages of attention processing are often deployed by ensemble of low-level features such as contrast, orientation, and intensity etc., an improved algorithm of saliency detection [10] is performed to extract salient AOs . In additional, faces and texts often carry the semantic information users expect and can be detected with much accuracy currently. Therefore, these three kinds of objects are also taken into account in our model and their attributions are evaluated similar to [5].

4.3 Combination of Detection Results

As the definition in [4][5], the size of Region-of-Interest (ROI) in a frame can be arbitrary. In our scheme, the size are constrained to a set of specific sizes, for example, 352x288 pixels for CIF, 176x144 pixels for QCIF etc., for being accordant with video coding standard. We define the attention area as a rectangle, whose size can be adaptively adjusted to a near value in the specific sizes according to different display sizes of mobiles. 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 [4] is utilized to crop the rectangular ROI in a frame. If the size of ROI is not fit to the near specific size, we will clip or expand the edges from the center of ROI to make it equal to the constrained size. In order to avoid the jittery results caused by producing these regions directly, the technique of virtual camera control [5][11] is adopted in our system to adjust positions of the cropped regions. The output of adjusted cropped regions is a smooth sequence. Such a rectangle region after adjustment is so-called attention area in our scheme.

5. Video transcoding based on the attention model

To generate a new bit-stream, a direct method is to re-encode each new frame of the sequence, which is referred to as cascade of decoder and encoder scheme. But the computational complexity is too high to be accepted in our system. Therefore, we propose an effective transcoding approach to address this problem. The cascade scheme is chosen as the basic structure. Figure 2 illustrates the proposed transcoding architecture.

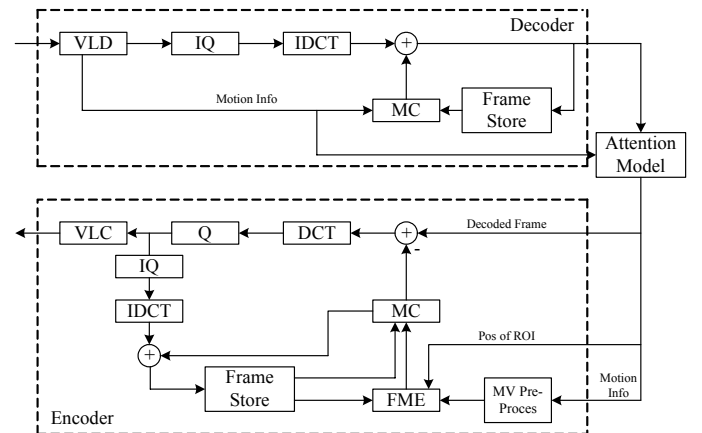


Figure 2. Block diagram of the transcoding system

Firstly, video information, e.g. motion and reconstructed frames, is obtained by decoding the incoming bit-stream.

We detect the attention area and crop it from the reconstructed frames to generate the new video. Then the encoder produces the output bit-stream for the new video. Since full-scale ME and full modes decision will result in unaccepted complexity, several techniques are adopted in our scheme to reduce encoding complexity, including motion reuse, simple mode decision. To improve quality performance, residue is re-computed to remove drifting error. Additionally, MV adjusting is employed to solve errors caused by cropping ROI from frames. Most of these techniques are included in two modules, FME (Fast Motion Estimation) and MV Pre-process in Figure 2.

5.1 Motion Vector Adjusting

MV adjusting is performed in the stage of MV Pre-Process as is shown in Figure 2 to solve errors caused by the difference between coordinates of ROI in two adjacent frames.

In Figure 3, we use a standard sequence, foreman, as testing data to show the comparative PSNR of with and without MV adjusting. It is illustrated that without MV adjusting, there would be a noticeable performance loss in transcoding process.

To present clearly, we first give some denotations:

- 1) F_1 and F_2 : two adjacent frames in incoming bitstream
- 2) $f_1(X_1, Y_1)$: ROI in F_1 , where (X_1, Y_1) is the origin coordinate in F_1
- 3) $f_2(X_2, Y_2)$: ROI in F_2 , where (X_2, Y_2) is the origin coordinate in F_2
- 4) MB_o : one MB in F_2
- 5) (MV_x, MV_y) : motion vector of MB_o
- 6) MB_p : prediction of MB_o in F_1
- 7) (X_{mbo}, Y_{mbo}) : coordinate of MB_o in F_1
- 8) (X_{mbp}, Y_{mbp}) : coordinate of MB_p in F_2
- 9) (X'_{mbo}, Y'_{mbo}) : coordinate of MB_o in f_1
- 10) (X'_{mbp}, Y'_{mbp}) : coordinate of MB_p in f_2

According to the definition of MV, we have the relation:

$$X_{mbp} = X_{mbo} + MV_x; Y_{mbp} = Y_{mbo} + MV_y \quad (4)$$

When cropping f_1 and f_2 from F_1 and F_2 , we can compute new coordinates for MB_o and MB_p

$$X'_{mbo} = X_{mbo} - X_2; Y'_{mbo} = Y_{mbo} - Y_2 \quad (5)$$

$$X'_{mbp} = X_{mbp} - X_1; Y'_{mbp} = Y_{mbp} - Y_1$$

If (X_1, Y_1) and (X_2, Y_2) are not identical, the relation between coordinates of MB_o and MB_p will be not met, i.e. ,

$$X'_{mbp} \neq X'_{mbo} + MV_x; Y'_{mbp} \neq Y'_{mbo} + MV_y \quad (6)$$

If we don't adjust MV for MB_o , the performance will loss greatly. So all MVs of MBs in f_2 must be modified by:

$$MV_x' = MV_x + (X_2 - X_1), MV_y' = MV_y + (Y_2 - Y_1) \quad (7)$$

where MV_x and MV_y are original MVs; MV_x' and MV_y' are modified MVs.

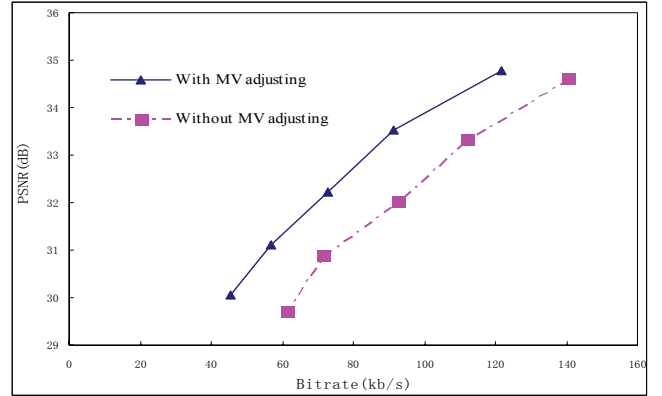


Figure 3. PSNR with and without motion vector adjusting in a standard sequence

5.2 Fast mode decision

As mentioned in Section 2, in H.264 coding standard, ME and mode decision contribute most of the complexity. To decrease heavy load of the transcoder, we propose a novel mode decision method in our transcoding scheme. It is not needed to re-searching all modes by utilizing the available motion information. Due to omitting the process of mode decision, computational complexity can be reduced greatly.

Since each MB is composed of sixteen blocks of sizes 4x4 and each block has its own MV, a MB has sixteen MVs. We observe part of MVs may be with the same value each other. Accordingly, a mode decision algorithm is presented as follows:

- 1) Group those MVs with the same value in a MB into several classes.
- 2) Count the number of MV in each class and find the class with the max number of MV. This MV is denoted as MV_{max} and the number is N_{max} .
- 3) If N_{max} is bigger than eight, the MB is set to be block type 16x16 and MVs of all blocks are replaced by MV_{max} .
- 4) If N_{max} is smaller than or equal to eight, further mode decision is needed. If the above eight blocks of 4x4 and the bottom eight blocks have the same MV, respectively. Then the mode is 16x8. Similarly, if the left blocks and right blocks have the same MV respectively. Then the mode is 8x16. If the above two cases are not satisfied. The mode is P8x8.
- 5) If decided mode is 16x8 or 8x16, the mode will be assigned to the MB and MVs will not be changed.
- 6) If decided mode is P8x8, the MB will be assigned the 8x8 mode. Four MVs of block with size 8x8 are chosen as candidate MVs.

However, simple MV reuse may introduce considerable quality degradation. To improve the performance, MV refinement is performed around the original MVs. The search range is 4. This will increase little complexity, but a significant quality gains can be achieved.

5.3 Drifting error removing

Figure 4 shows open-loop and close-loop architectures. Open-loop architecture is a simple solution since no frame memory is required and there is no need for an inverse Discrete Cosine Transform (DCT). However, it is subject to drifting error. In video coding, drift error refers to the continuous decrease in picture quality when a group of motion-compensated (MC) interframe pictures are decoded. In general, it is due to the mismatch between predictive and residual components [14]. The performance of close-loop architectures is better than that of open-loop architectures, but close-loop architectures have more complexity.

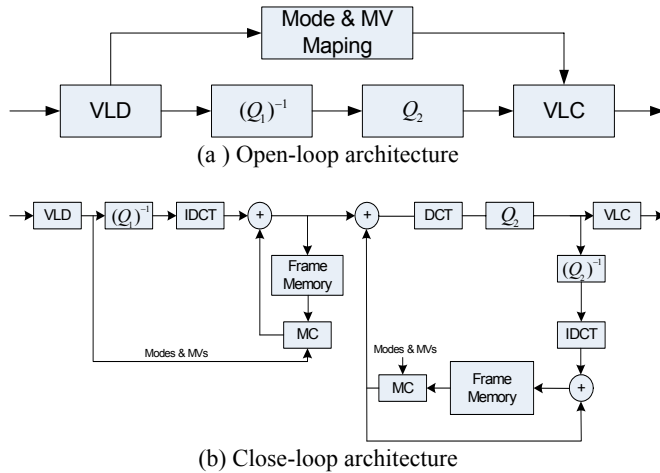


Figure 4. Open-loop and Close-loop architectures of transcoder

In our proposed scheme, close-loop architecture is adopted for several reasons. Firstly, the goal of our work is to improve users' perceptual experiences, while drift error from the open-loop solution would degrade perceptual results obviously. Secondly, the development of hardware is fast for the close-loop solution. For special chips have been designed for DCT, the complexity of DCT is not a handicap currently. Thirdly, previous standards such as MPEG-1, MPEG-2, MPEG-4 and H.263 make use of the 8x8 DCT as the basic transform. Since float computing is performed in these standards, the complexity is relatively high. In H.264 standard, an integer transform which operates on 4x4 blocks of residual data is adopted. All operations can be carried out with integer arithmetic. The core part of the transform is multiply-free, i.e. it only

requires additions and shifts operations. More details about integer transform can be found in [2]. Therefore, the complexity of transform of H.264 is much lower than that in the previous standards. It can be concluded that with the close-loop architecture, drifting error will be avoided and the quality can be guaranteed.

In our transcoding scheme, the process of ME has been bypassed with the utilization of the original motion information. Thus the complexity can be reduced mostly. At the same time, by making full use of the original decoding information, we can gain the similar display result and more favorable perceptual quality than the original reconstructed sequence. An example of the output results is shown in Figure 5.

6. Experiments

We have implemented a prototype based on JM61e [3], the reference software of H.264 and done several experiments to evaluate the effectiveness of our scheme. Our test bed was a Dell PC with P4 2.0G CPU, 512M memory and MS Windows XP Professional system. There were three kinds of videos tested in our experiment: standard sequences (foreman, coastguard, news, akiyo and mobile), home videos from personal collections and TV program videos from CNN news. For each kind of video, five segments were chosen with equal durations and specified intervals for testing.

6.1 Subjective Testing

In order to evaluate the perceptual impression of the output video stream, a controlled user study experiment was carried out. We invited eleven volunteers who have no knowledge of our project. Two subjective assessment questions were given to the users and the average scores are shown in Table 1.

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)

2. Do you think the visual quality of the result is acceptable for small displays? (3-Good, 2-Fair, 1-Poor)

Table 1. Subjective testing results

Video type	Standard	Home	TV	Average
Score of question 1	3.52	3.71	3.28	3.50
Score of question 2	2.57	2.71	2.29	2.52

From the user study experiments, we can learn that our scheme achieved acceptable display results. Perceptual results of home videos are improved more than other kinds of videos.

6.2 Objective Testing

On the condition of appreciating user experience, we also measured the bitrate reduction in our method. The results are shown in Table 2. The original videos were compressed using JM61e [3] with the Quantization Parameter 28. The original bit-rates ranged from 261.91 to 1323.87 kb/s in standard sequences, from 420.8 to 547.9 kb/s in the home videos and from 246.9 to 319.5 kb/s in TV program videos. The average bitrate saving is shown in Table 2.

Table 2. Bitrate reduction comparison

Video type	Standard	Home	TV
Average bitrate saving	71.8%	72.8%	59.6%

In other experiment, we adopted decoding speed (frames decoded per second) to evaluate the reduction of complexity and the results are shown in Table 3.

Table 3. Decoding speed comparison

Video type	Standard	Home	TV
Original speed (frame/s)	14.9	16.4	13.0
Current speed (frame/s)	54.1	77.4	60.6

From the above results of subjective and objective experiments, we can learn that the transcoded home videos gain a better performance than other types of videos. One possible reason is that home videos are not well-edited and they are more suitable for further editing by producing the smoothed attention areas as is performed in the proposed solution. In addition, since home videos often only focus on a few objects, the redundant information can be left out without much impairment of perceptual results.

To reveal the performance of our transcoding scheme, we compared it with the fully encoding scheme in three aspects, PSNR, rate and encoding time. In the comparative experiment, five standard sequences, foreman, coastguard, news, akiyo and mobile were tested.

As shown in Figure 6, the scheme we proposed has very similar encoding performance (PSNR loss less than 0.2dB) with fully encoding scheme while the computational complexity was reduced by about 80%.

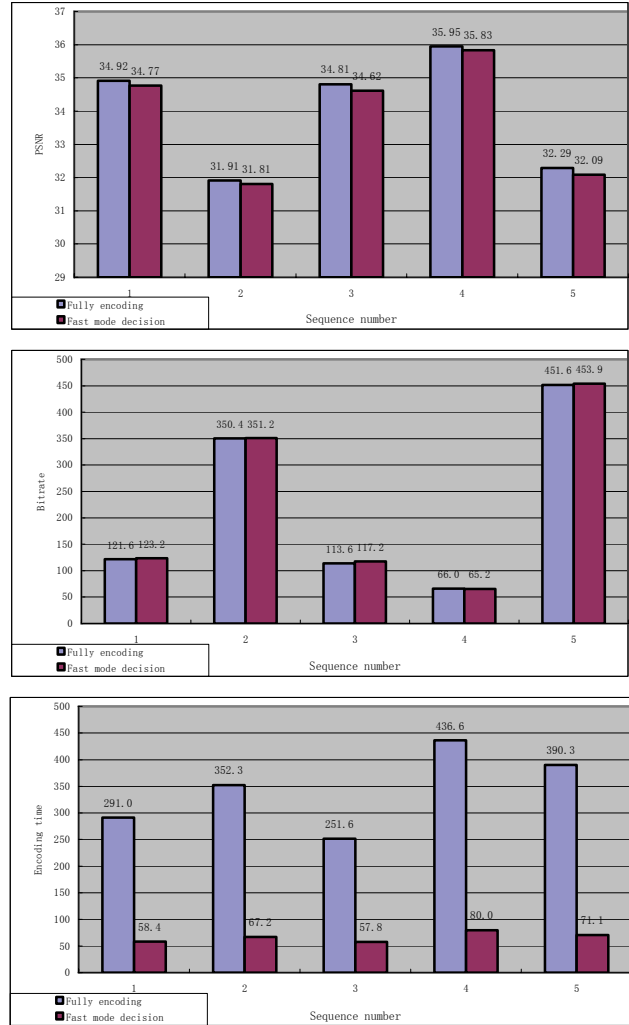


Figure 6. Performance comparison of fully encoding scheme and our scheme

7. Conclusions

In this paper, we proposed a novel solution for mobile users to enjoy videos over wireless channel. A visual attention model is utilized to detect the most informative regions and a new transcoding method is employed to produce perceptually improved bitstream conformed to H.264 standard. In our scheme, users can gain most information of the original sequence with better experience. Additionally, the transcoded video is at a lower bitrate and decoding computation is much reduced. Moreover, our approach can be easily applied to video transcoding proxies.

8. ACKNOWLEDGEMENT

The work of Yi Wang, Houqiang Li, and Zhengkai Liu are supported by NSFC under contract No. 60333020 and open fund of MOE-Microsoft Key Laboratory of

9. References

- [1] "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.
- [2] "H.264/MPEG-4 Part 10: Transform & Quantization", *H.264 tutorial*, <http://www.vodex.com>.
- [3] "JVT reference software official version," *Image Processing Homepage*, <http://bs.hhi.de/~suehring/tml/>.
- [4] 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", *ACM Multimedia Systems Journal*, Springer-Verlag, Vol.9, No.4, pp. 353-364, 2003.
- [5] X. Fan, X. Xie, H.-Q. Zhou and W.-Y. Ma, "Looking into Video Frames on Small Displays," *Proceedings of the eleventh ACM international conference on Multimedia*, pp. 247-250, Berkeley, CA, USA, Nov. 2003.
- [6] T. Koga, K. Iinuma, A. Hirano, Y. Iijima, and T. Ishiguro, "Motion-compensated interframe coding for video conferencing," *Proceedings of NTC 81*, pp. C9.6.1-9.6.5, New Orleans, LA, Nov./Dec. 1981.
- [7] C.-H. Kuo, M. Shen and C.-C. J. Kuo, "Fast Inter-prediction mode decision and motion search for H.264," *Proceedings of IEEE International Conference on Multimedia and Expo 2004*, pp.663-666, Taipei, Taiwan, Jun. 2004.
- [8] Jeyun Lee, and Byeungwoo Jeon, "Fast mode decision for H.264," *Proceedings of IEEE International Conference on Multimedia and Expo 2004*, pp.1131-1134, Taipei, Taiwan, Jun. 2004.
- [9] Renxiang Li, Bing Zeng, and Ming L. Liou, "A New Three-Step Search Algorithm for Block Motion Estimation," *IEEE Trans. Circuits Syst. Video Technol.* Vol. 4, No.4. Aug. 1994.
- [10] Y.-F. Ma and H.-J. Zhang, "Contrast-Based Image Attention Analysis by Using Fuzzy Growing," *Proceedings of the eleventh ACM international conference on Multimedia*, pp. 374-381, Berkeley, CA, USA, Nov. 2003.
- [11] X. Sun, J. Foote, D. Kimber, and B.S. Manjunath, "Panoramic video capturing and compressed domain virtual camera control," *Proceedings of the nine ACM international conference on Multimedia*, pp. 229-238, Ottawa, Canada, Sep. 2001.
- [12] A.M. Tourapis, "Enhanced Predictive Zonal Search for Single and Multiple Frame Motion Estimation," *Proceedings of Visual Communications and Image Processing 2002*, pp. 1069-1079, San Jose, CA, Jan. 2002.
- [13] A.M. Tourapis, O.C.Au, and M.L.Liou, "Highly efficient predictive zonal algorithms for fast block-matching motion estimation," *IEEE Trans. Circuits Syst. Video Technol.* Vol. 12, No.10, pp. 934-947, Oct. 2002.
- [14] A. Vetro, C. Christopoulos, H. Sun, "An overview of video transcoding architectures and techniques," *IEEE Signal Processing Magazine*, Vol.20, No.2, p18-29, Mar. 2003.
- [15] P.Yin, H.-Y.C. Tourapis, A.M. Tourapis, and J. Boyce, "Fast mode decision and motion estimation for JVT/H.264," *Proceedings of International Conference on Image Processing 2003*, pp.853-856, Barcelona, Spain, Sep. 2003
- [16] P. Yin, A. Vetro, B. Lui, and H. Sun, "Drift compensation for reduced spatial resolution transcoding," *IEEE Trans. Circuits Syst. Video Technol.* Vol. 12, pp. 1009-1020, Nov. 2002.



Figure 5. Comparison of the original video sequence (the above) and the transcoded video sequence (the below)