

Video Streaming to Mobile Handheld Devices: Challenges in Decoding, Adaptation, and Browsing

Chang Wen Chen¹ and Houqiang Li²

¹ Department of Electrical & Computer Eng., Florida Institute of Technology,
150 W. University Blvd., Melbourne, FL 32901 USA

² Dept. of EEIS, University of Science and Technology of China,
Hefei, 230027, P.R. China

cchen@fit.edu, lihq@ustc.edu.cn

Abstract. Growing popularity and richer functionality of contemporary mobile handheld devices such as PDAs and smart phones have enabled emerging video streaming applications to these devices via various wireless networks. However, these handheld devices are significantly constrained by limited resources available to them for some sophisticated operations. Among these limitations, the available processing power has limited these devices to perform full scale decoding of the video compressed with H.264 standard [1], [2]. The typically small display size of the handheld devices has been imposing significant barriers for the devices users to enjoy high-resolution videos. This paper presents some challenges in design decoding algorithm for smart phones, in video adaptation to suit heterogeneous devices, in browsing videos via handheld devices by diverse mobile device users. Several approaches will be proposed to meet these challenges and to offer the mobile device users an enhanced experience in these emerging mobile video applications.

Keywords: H.264, video streaming, mobile devices, handheld devices, video decoding, video adaptation, video browsing, visual attention modeling.

1 Introduction

Recent advances in wireless communication and networking have enabled an entire new array of applications in multimedia over wireless networks beyond traditional applications in voice and text. Among these real-time multimedia applications, video streaming to mobile handheld devices has become feasible under various wireless environments. In particular, with the growing popularity of mobile devices, users can enjoy videos in almost anyplace and at anytime. For video streaming applications in the wired networking environment, the videos stored in the video server are generally encoded at high-resolution (HR) and high bitrate streaming is generally able to guarantee users' access and browsing experiences.

There are several fundamental difficulties in enjoying these videos with mobile devices over wireless networks. The first critical constraint is the necessity of the low power operations for all the mobile handheld devices driven by the battery. These battery driven devices have limited processing power and memory size that generally

prevent the implementation of computationally complex algorithms. The second critical constraint is that the wireless links for mobile handheld devices to access the video data are usually band-limited and error prone. Transmission of compressed video over such links presents significant challenges. The third essential 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 with current display size but with an enhanced viewing experience.

In this paper, we will present some challenges to overcome these fundamental constraints in order for the users of mobile handheld devices to enjoy a true universal multimedia access (UMA) experience under the pervasive media environment. UMA refers to the user experiences in which access and consumption of multimedia content over heterogeneous networks have been accomplished using diverse terminals in a seamless and transparent way [3]. We will then address the technical issues aiming at overcoming some fundamental constraints in video streaming to mobile devices. In particular, we will discuss the decoding of the compressed video on the smart phone, adaptation of video streams at the edge servers for heterogeneous mobile devices, and the browsing of streaming video by a diverse class of mobile device users.

The rest of this paper is organized as follows. Section 2 presents in some detail the challenges that stem from various physical constraints of the wireless networks or mobile devices. Section 3 introduces several key technologies that have recently been developed to address the challenges in order to provide enhanced UMA experiences to mobile device users. These technologies include video decoding, video adaptation, and video browsing for streaming to mobile devices. Section 4 concludes this paper with a summary and a discussion of future technology trends.

2 Challenges in Video Streaming to Mobile Handheld Devices

Mobile handheld devices are inherently resource constrained because they need to be portable and powered by battery. The first challenge for video streaming to the mobile devices is to design a suitable video encoding and decoding algorithm for mobile video applications. In general, portable, battery operated mobile devices pose the following particular constraints [4] that need to be carefully taken care of when implementing a video coding scheme:

- 1) Power (battery life)
- 2) Memory size
- 3) Cache size/memory bandwidth
- 4) CPU speed
- 5) Display Resolution
- 6) Bandwidth

As a result, we believe that the desirable attributes of a video coding scheme developed for mobile device applications should include:

- 1) High coding efficiency to meet the limited bandwidth constraints.
- 2) Balance of coding efficiency and complexity of implementation geared towards inexpensive consumer-level appliances.

- 3) Low cost implementation of decoder, considering the limited CPU power and the limited cache/memory.
- 4) Massive deployment in vast numbers of consumer-level mobile receivers
- 5) Coding tools included should provide significant coding gain relative to the cost for their decoder implementation.

Considering the current video streaming applications, the first step towards UMA through mobile devices is to enable seamless implementation of state-of-the-art video decoder on the mobile devices so as to allow the users to decode and display the received video bitstreams. This is particularly challenging when the cellular phone is used as the mobile device for video streaming and display. This is because the capacity left to implement video streaming is quite limited after the essential services of voice calling are implemented on the phone.

The second challenge for video streaming to the mobile devices is to design some video adaptation schemes in order to meet the requirements of seamless services for an array of heterogeneous mobile devices having diverse functionality and capability differences. In particular, the display size of these devices can be quite different and the display resolution of these devices is usually much smaller than the resolution of original video. Video adaptation with reduced spatial resolution is therefore necessary in order to stream the appropriate amount of bitstreams to the mobile devices whose link capacity is constrained by the bandwidth of wireless access networks.

In recent years, some video transcoding schemes [5], [6] have been developed to achieve the spatial resolution reduction. These traditional spatial transcoding schemes are usually based on spatial downsizing by the factor of two or power of two. Considering perceptual results, excessive resolution reduction will cause significant loss of desired information. More recently, some researchers proposed 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 [7]. Similarly, an attention based video adaptation solution [8] was developed for better perceptual results in small displays. However, these schemes adopt the pixel-domain approach that the original videos are decoded and then fully encoded after processing. This is often very costly in terms of computational complexity and is not suitable for mobile devices in practice.

Traditionally, the video adaptation operations are implemented at the proxy server linking between the wired networks and the mobile devices. One such example will be the access point for wireless local area networks (WLAN). When the access point receives the spatial resolution and other parameters from the mobile devices, these parameters will be utilized to design an appropriate adaptation scheme for this particular device. Every time a different request may be received from heterogeneous mobile devices diverse, the video adaptation scheme will needed re-designed. This creates significant burden at the proxy server such as access point in the WLAN which may become the bottleneck in implementing the UMA environments [3].

This technical issue constitutes the third challenge for video streaming to mobile devices. It is the best interest of any UMA service provider to shift the computational burden from proxy server to the original server which in general enjoys significantly more resources than the proxy servers. The shift of some key video adaptation operations from proxy server to original video server can be made possible as long as the

desired video adaptation for heterogeneous mobile devices can be appropriately parameterized and these parameters can be embedded into the compressed video bit-streams to be utilized by the proxy servers.

3 Key Technologies for Seamless Video Streaming to Mobiles

In the following, we shall discuss several key technologies that have been developed recently to address the technical issues in providing seamless video streaming services to mobile handheld devices. We will start with the most essential operations for such service in terms of reduced complexity video decoding on the mobile devices. We will then examine the video adaptation schemes that are necessary for streaming reduced spatial resolution video for small display size of the mobile devices. Finally, we will present an innovative approach that is able to shift the burden of extensive video adaptations at the proxy server to the original server to allow light operations at the proxy servers so as to enable a seamless video streaming from the server to proxy and to the mobile users.

3.1 Reduced Complexity Video Decoding for Mobile Devices

We consider implementing the video coding scheme on the mobiles to be compliance with the current H.264/AVC standard [1], [2]. H.264/AVC is the powerful and state-of-the-art video compression standard that has recently been developed by the ITU-T/ISO/IEC Joint Video Team consisting of experts from ITU-T's Video Coding Expert Group (VCEG) and ISO/IEC's Moving Picture Experts Group (MPEG) [10]. The H.264/AVC standard represents a delicate balance between coding gain, implementation complexity, and costs based on state of VLSI design technology.

To develop a reduced complexity video coding scheme based on H.264/AVC standard, we need to investigate each component of the video encoding and decoding in terms of its contribution to coding efficiency against its induced implementation complexity. Since this reduced complexity video decoding scheme is designed for local playback on mobile device, we focus on those coding modules that are adopted by both encoder and decoder. This is because these modules have greatest impact on both coding efficiency at the encoder and the implementation complexity at the decoder. However, when a module is included only at the encoder, we shall adopt a high coding efficiency coding technique even though the module is high in implementation complexity. Such coding module will not impact the mobile devices since these modules will not be implemented at the decoder.

The implementation has been focused on several key modules in video coding. One important common feature of these implementations is that the key components have been enforced to be compliance with the H.264/AVC standard. We shall use Nokia 6600 as an example to illustrate the principles behind the design. However, we the principles of implementation enable the potential widespread adoption of proposed scheme on other mobile devices from different manufacturers.

We have considered the following key modules for the potential reduction of implementation complexity:

- 1) Inter Mode
- 2) Intra Mode
- 3) Quarter-sample-accurate motion compensation
- 4) Rate Distortion Optimization
- 5) Transform and quantization [11]
- 6) Interpolation
- 7) Entropy coding [12]
- 8) In-the-loop deblocking filter [13]
- 9) Frame drop

For each of these modules, we investigate how the reduction of complexity can be accomplished with minimum loss of coding efficiency. When certain reduction causes substantial loss in coding efficiency, such reduction will not be carried out. The following table lists the modules we considered with a comparison between the H.264 Reference Software JM76 [14] and the proposed scheme. Notice that many modules remain unchanged because the gain in complexity reduction is out weighted by the loss in coding efficiency. On average, the loss of coding efficiency is under 0.5dB in PSNR caused by complexity reduction. However, the complexity reduced codec can now be implemented on the Nokia 6600 for decoding H.264 video bitstreams. The following table summarizes the results and implementation details can be found in [15].

Module	H.264 Ref. SW JM76	The Proposed Scheme
Inter Mode	seven	seven
Intra Mode	nine for Intra4x4 and four for macroblock	five for Intra4x4
Motion Vector	quarter-accuracy	quarter-accuracy
RDO	enabled	enabled
Transform	4x4 integer transform,	4x4 integer transform
Quantization	52 steps	52 steps
Interpolation	6-tap for half-pix and bilinear for quarter-pix	4-tap for half-pix and bilinear for quarter-pix
Entropy Coding	CAVLC	Exp-Golomb
Loop filter	enabled	enabled
Reference frame	two	two

3.2 Attention-Based Video Adaptation for Mobile Devices

As we indicated in Section 1, there have been numerous research activities in spatial resolution reduction for displaying video at mobile devices with small display size. Majority of traditional simple transcoding techniques have been on spatial downsizing by the factor of two or power of two [5], [6]. Considering the perceptual results, excessive resolution reduction will cause significant loss of detailed information. Therefore, for video streaming to mobile devices, a better way to solve this problem is to identify the most informative regions within the original video frame and crop an appropriate portion of the video frame containing the informative region for the small display at the mobile devices.

Once again, we consider the H.264/AVC standard as the default video coding scheme for these applications because of its efficient coding performance and its widespread adoption by the applications community. A typical approach for video adaptation would be to generate a new bitstream with the direct method. In this case, the scheme re-encodes each new frame of the sequence after full decoding of the received data. However, full-scale motion estimation and full modes decision will result in unaccepted complexity at the mobile devices. Therefore, we need to develop several techniques in order to reduce encoding complexity, including techniques in motion reuse and simple mode decision [16], [17].

One major issue needs to be resolved is how to identify the informative region within a video frame for the generation of reduced resolution video streaming and display at the mobile devices. The approach we developed is to determine the most informative region in the decoded frames by an attention based modeling. As shown in Definition 1, a set of information carriers – attention objects (*AOs*) [7] are defined in our method.

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

Each *AO* owns three attributes: *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.

Accordingly, three attributes of *AOs* will be measured by an automatic modeling method. We have taken into account four types of attention objects in the proposed model: motion objects, face objects, text objects and saliency objects. Since we study video streaming, the modeling is different from that of the static pictures [8] in that motion information with in a video frame carries significant information. For H.264 video format, the Motion Vector Field (MVF) of a frame is able to measure the motion information well.

After we extract the attention areas from the video frames, we are faced another technical challenge in generating the reduced resolution smooth video frames from the extracted attention areas. A straightforward approach is to directly present the focus regions to the user. However, this naïve approach will cause jitters in the video sequences since such frames will be discontinuous after cropping. We notice that in some tracking systems [18], virtual camera control is introduced to smooth the tracking process. Therefore, we assume that there is also a virtual video camera steered to pan and zoom in the video frames. Two types of focuses are introduced to implement virtual camera control: Camera Focus (CF) and Target Focus (TF). Camera Focus represents the focus displayed to the users and Target Focus represents the destination focus that can either be assigned manually or determined automatically. Corresponding display ratios of the two types of focus regions are defined as Camera Ratio (CR) and Target Ratio (TR), respectively. The Euclidean distance between CF and TF is denoted as Δd and the difference of CR and TR is defined as Δr . The direct focus shifting is substituted for a smooth following process from the current focus region to the target focus region, with a set of pan and zoom operations. As a result of a

sequence of fixation, following, and shifting operations, the output of adjusted cropped regions will be a smooth sequence. Such a rectangle region after adjustment will be noted as the attention area in video adaptation. The detailed implementation on attention information-based video adaptation for mobile devices can be found in [19]. Yet, there are still open issues in video adaptation [20].

3.3 Embedded Intelligent Video Adaptation for Browsing Via Mobiles

Although the attention area extraction-based video adaptation is able to generate reduced resolution video for mobile devices with small display, this region-of-interest (ROI) based transcoding at proxy 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 areas 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 shall become a significant burden for either edge server or proxy. The second shortcoming of the ROI-based transcoding is the need to perform ROI detection for heterogeneous users every time a user requests the video browsing service. Such repeated operations will sometimes overwhelm the proxy server.

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 compressed video to be extracted by the proxy server, the burden of the computational complexity can be shifted from the transcoding at proxy server to encoding at the original server. It is this analysis of shifting the operations from proxy server to original server that motivates us to design an attention-information-based video adaptation for browsing video via mobile devices.

This intelligent design of video spatial adaptation is based on the assumption that the videos stored in the original video server are usually off-line generated and the computational complexity is not an issue with off-line 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 shift the attention detection operations from the transcoding process to the encoding process, then, we will be able to shift the complexity from proxy to the video server.

Based on the above analysis, we have recently developed an intelligent spatial adaptation framework 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 be detected from the video simultaneously. Then the bitstreams and the attention information will be stored together in the video server. When the server offers services for mobile devices users, the adaptation system placed on a proxy or server will perform adaptation manipulation on the original high resolution video by making use of the attention information in the video to meet the display constraint of the mobile devices.

The separation of attention information detection and adaptation operation has two benefits. First, since the detection process need be performed only once and the detected attention information can be used for heterogeneous users, the workload of this new adaptation system will be reduced greatly while the system still remains

flexible. This intelligent adaptation system only needs to complete the function in transforming high resolution videos into low resolution videos for mobile devices. This will greatly facilitates the potential 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 pre-detected attention information.

Since original sequences are unavailable in the transcoding or video adaptation process, reconstructed frames after decoding have to be used as the input video to transcoding. The better quality the reconstructed original video frames, the higher performance the transcoder will be able to achieve. In browsing video via mobile devices, the new video sequence generated by the transcoding or video adaptation suitable for the target mobile device will consist of mostly attention areas in each frame. Therefore, if we can adjust the bit allocation strategy in the original encoding for attention and non-attention areas, we shall be able to improve the quality of low resolution adapted 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 non-attention areas to attention areas when we encode the original video.

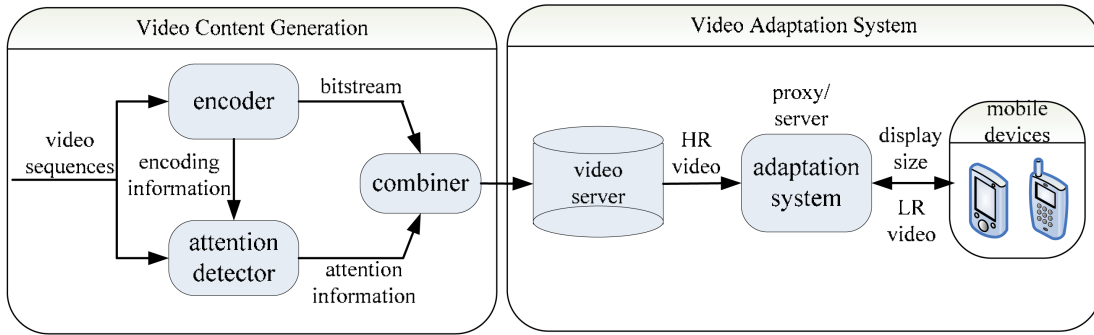


Fig. 1. The structure of attention-information-based spatial adaptation framework

We have implemented several strategies in motion estimation and mode decision to achieve a high performance transcoding [21], [22], [23]. We tested the proposed intelligent video adaptation with several standard video sequences at a variety of bitrates in which the original CIF bitstreams are transcoded into QCIF bitstreams containing attention areas. Comparing with the simple video adaptation scheme we presented in Section 3.2, this embedded intelligent video adaptation is able to obtain R-D performance improvement at all bitrates and the improvement can be as high as 0.5dB. The details of this intelligent video browsing scheme can be found in [24].

4 Summary and Future Technology Trends

In this paper, we have presented several fundamental constraints in order for the users of mobile handheld devices to enjoy true universal multimedia access experiences under the pervasive media environment. These fundamental constraints are originated from the inherent physical limitations of either wireless access links or mobile device

operations. We then addressed several technical issues aiming at overcoming some fundamental constraints in order to provide seamless video streaming to mobile devices. In particular, we discussed the decoding of the compressed video on the smart phone, adaptation of video streams at the edge servers for heterogeneous mobile devices, and the browsing of streaming video by a diverse class of mobile device users via the wireless access links.

One recent emerging mobile video application is the mobile Internet TV (IPTV) which has attracted significant attention from both research community and service providers. In this application, the consumers are expecting seamless streaming of broadcasting video contents to the mobile receiving devices. Although the content creators may produce some special programs targeting only mobile reception devices with low spatial resolution, many broadcasting video programs need to be adapted from high resolution originals produced for high end users to low resolution videos for reception by mobile devices. In this case, all three technical approaches presented in this paper are applicable to mobile IPTV applications.

Another future technology trend is in the application of video adaptation in peer-to-peer (P2P) video streaming to and from mobile devices. Video adaptation becomes necessary when the video data stored in your peers have significant resolution mismatch because simple streaming of original data cannot be implemented. In this case, we will need to integrate the video adaptation with protocol designs for the seamless P2P streaming of video data to and from mobile device peers.

Finally, the visual attention-based video adaptation has been based on the basic assumption that human attention tends to remain similar when viewing entertainment video. This assumption is not applicable to special circumstances in which we seek unique visual information within a video frame that cannot be easily characterized by generic human visual attention attributes. Special design is needed for these unique cases, including surveillance and tactical video for security and defense applications.

Acknowledgment. This work has been supported by a grant from the National Natural Science Foundation of China (No. 60572067). The authors wish to thank Mr. Yi Wang for his contributions to this project.

References

1. Wiegand, T., Sullivan, G.J., Bjontegaard, G., Luthra, A.: Overview of the H.264/AVC video coding standard. *IEEE Trans. Circuits Syst. Video Technol* 13, 560–576 (2003)
2. Tamhankar, A., Rao, K.R.: An overview of H.264/MPEG-4 part 10., Communications and Signal Processing Track, Texas Systems Day, November 15, 2003, Dallas Hall, SMU (2003)
3. Kim, J.-G., Wang, Y., Chang, S.-F., Kim, H.-M.: An optimal framework of video adaptation and its application to rate adaptation transcoding. *ETRI Journal* 27(4)(2005)
4. Mobile Profile Proposal, ISO/IEC, JTC1/SC29/WG11 and ITU-T SG16 Q.6, document JVT-C161, 3rd meeting, Fairfax, VA, USA, 6-10 March 2002 (2002)
5. Vetro, A., Christopoulos, C., Sun, H.: An overview of video transcoding architectures and techniques. *IEEE Signal Processing Magazine* 20(2), 18–29 (2003)
6. Xin, J., Lin, C.-W., Sun, M.-T.: Digital Video Transcoding, In: *Proceedings of the IEEE*, 93(1) (January 2005)

7. Fan, X., Xie, X., Zhou, H.-Q., Ma, W.-Y.: Looking into video frames on small displays. In: Proceedings of the eleventh ACM international conference on Multimedia November 2003, Berkeley, CA, USA, pp. 247–250 (2003)
8. Chen, L.Q., Xie, X., Fan, X., Ma, W.Y., Zhang, H.J., Zhou, H.Q.: A visual attention model for adapting images on small displays. *ACM Multimedia Systems Journal* 9(4), 353–364 (2003)
9. Hua, X.-S., Chen, X.-R., Liu, W., Zhang, H.-J.: Automatic Location of Text in Video Frames, In: Proceedings of the third International Workshop on Multimedia Information Retrieval, Ottawa, Canada, October 2001, pp. 126–129 (2001)
10. 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)
11. H.264/MPEG-4 Part 10: Transform & Quantization, H.264 tutorial <http://www.vodex.com>
12. Marpe, D., Schwarz, H., Wiegand, T.: Context-Based Adaptive Binary Arithmetic Coding in H. 264/AVC Video Compression Standard. In: *IEEE Trans. Circuits Syst. Video Technol* 13, 620–636 (2003)
13. List, P., Joch, A., Lainema, J., Bjontegaard, g., Karczewicz, M.: Adaptive deblocking filter. *IEEE Trans. Circuits Syst. Video Technol* 13, 614–619 (2003)
14. JVT reference software official version, Image Processing Homepage <http://bs.hhi.de/suehring/tml/>
15. Wang, Y., Li, H., Chen, C.W.: A novel video coding scheme for mobile devices. *Proc. Int. Conf. Mobile and Ubiquitous Multimedia* October 2004, College Park, MD (2004)
16. Kuo, C.-H., Shen, M., Kuo, C.-C.J.: Fast Inter-prediction mode decision and motion search for H.264,” *Proceedings of IEEE International Conference on Multimedia and Expo 2004*, Taipei, Taiwan, pp. 663–666 (June 2004)
17. Lee, J., Jeon, B.: Fast mode decision for H.264. In: *Proceedings of IEEE International Conference on Multimedia and Expo 2004*, Taipei, Taiwan June 2004, pp.1131–1134 (2004)
18. Sun, X., Foote, J., Kimber, D., Manjunath, B.S.: Panoramic video capturing and compressed domain virtual camera control. In: *Proceedings of the ninth ACM international conference on Multimedia*, Ottawa, Canada, September2001 pp. 229–238 (2001)
19. Wang, Y., Li, H.Q., Fan, X., Chen, C.W.: An attention based spatial adaptation scheme for H.264 videos over mobiles. *Int. Journal Pattern Recognition & Artificial Intelligence special issue on Intelligent Mobile & Embedded Systems* 20(4), 565–584 (2006)
20. Chang, S.-F., Vetro, A. : Video adaptation: Concepts, technologies, and open issues, In: *Proc. IEEE*, 93(1) 148–158 (2005)
21. Tourapis, A.M., Au, O.C., Liou, M.L.: Highly efficient predictive zonal algorithms for fast block-matching motion estimation. *IEEE Trans. Circuits Syst. Video Technol.* 12(10), 934–947 (2002)
22. Yin, P., Tourapis, H.-Y.C., Tourapis, A.M., Boyce, J.: Fast mode decision and motion estimation for JVT/H.264. In: *Proceedings of International Conference on Image Processing 2003*, Barcelona, Spain, September 2003 pp. 853–856 (2003)
23. Shen, B.: Submacroblock motion compensation for fast down-scale transcoding of compressed video. *IEEE Trans. Circuits Syst. Video Technol.* 15(10), 1291–1302 (2005)
24. Wang, Y., Li, H., Chen, C.W.: An attention-information-based spatial adaptation framework for browsing videos via mobile devices. *EURASIP Journal of Advances in Signal Processing* (Accepted for publication) (2007)