INT J COMPUT COMMUN, ISSN 1841-9836
Vol.7 (2012), No. 3 (September), pp. 473-481

Optimal Bitstream Adaptation for Scalable Video Based On Two-Dimensional Rate and Quality Models

J. Hou, S. Wan

Junhui Hou, Shuai Wan
Northwestern Polytechnical University
School of Electronics and Information
Xi'an 710129, China
E-mail: houjunhuihn@gmail.com
swan@nwpu.edu.cn

Abstract: In this paper, a two-dimensional (2D) rate model is proposed considering the joint impact of spatial (i.e., the frame size) and SNR (i.e., the quantization step) resolutions on the overall rate-distortion performance. A related 2D quality model is then proposed in terms of perceptual quality. Then the two proposed models are applied to scalable video to address the problem of optimal bitstream adaptation. Experimental results show that the proposed rate and quality models fit the actual data very well, with high coefficients of determination and small relative root mean square errors. Moreover, given the bandwidth constraint and required display resolution of the end users, the optimal combination of SNR and spatial layers that provides the highest perceptual quality can be achieved using the proposed models.
Keywords: 2D rate model, 2D perceptual quality model, scalable video, bitstream adaptation.

1 Introduction

Recent multimedia applications are featured by various resolutions designed for a variety of devices with different computational and display capabilities. These devices range from cell phones and PDA's with small screens and restricted processing power to high-end work stations with high-definition displays. The related video services or applications are connected to different types of networks with various bandwidth limitation and loss characteristics. A highly attractive approach to address the vast heterogeneity is known as scalable video, which allows for spatial, temporal, and SNR scalabilities [1]. In Scalable Video Coding (SVC), the video signal can be encoded into a Base Layer (BL) and one or more Enhancement Layers (ELs), with each enhancement layer improving the resolution (either temporally or spatially) or the fidelity of the video sequence. As a result, certain parts of the scalable bitstream can be removed for adaptation to various capabilities of end users as well as varying network conditions.

At the network proxy or gateway, a bitstream adaptor is usually employed to extract the bitstream to meet particular constraints, e.g., targeted bit-rates and/or spatial or temporal resolutions. For a given set of constraints, the solution can be varieties of resolution combinations, leading to different visual qualities. The challenging problem of bitstream adaptation is therefore how to determine the combination of the spatial resolution (i.e., the frame size (s)), temporal resolution (i.e., the frame rate (t)) and SNR (Signal-to-Noise Ratio) resolution (i.e., the quantization step (q)) to be used for bitstream extraction under a given targeted bit-rate to maximize the resulting quality.

Many efforts have been devoted to bitstream adaptation for scalable video. A basic and content independent extractor is provided in the reference software of the Joint Scalable Video Model (2) [2]. In [3], an alternative extraction method is proposed based on rate-distortion optimization. This technique utilizes the concept of quality layers and improves the performance of the JSVM basic extractor by arranging the priority of layers based on their contributions to the global improvement in quality. A more

efficient method for extraction is proposed in [4], using an accurately and efficiently estimation of the quality degradation resulting from discarding an arbitrary number of Network Abstraction Layer (NAL) units from multiple layers taking drift into account. However, the methods in [3] and [4] are executed only within a single resolution, e.g., the SNR plane. In [5], an effective method is proposed to quickly solve the problem of spatial resolution selection based on an analysis on the content information. However, the Peak-Signal-to-Noise Ratio (PSNR) is used as the distortion criterion, which does not correlate well with the perceptual quality especially with regard to spatial scalability. In [6], two-dimensional (2D) rate and perceptual quality models in terms of the frame rate and the quantization step are built, and then the two models are applied to optimal bitstream extraction in SVC. However, the spatial resolution is not considered in [6] and the parameters in the quality model are difficult to obtain. In [7], the video quality under different spatial, temporal and SNR combinations is quantitatively and perceptually assessed, based on which an efficient adaptation algorithm is proposed. However, there is lack of a rate model to estimate the related bit-rate. On the other hand, performance improvement can be achieved by resorting to network-related technologies, such as using a priority mechanism [8], or self optimization of networked communications [9] presents a model for self optimization of network communications in order to improve cluster performance by shortening the data transfer time.

In this paper, 2D rate and quality models are proposed for optimal bitstream adaptation for scalable video under given bandwidth constrains and required display resolutions at the end user. Assuming that the frame rate is determined, the two 2D models are applied to extract bitstream to achieve the optimal combination of spatial and SNR resolutions.

The rest of the paper is organized as follows: Section 2 presents the 2D rate and perceptual quality models considering the impact of spatial and SNR resolutions. Their application in constrained scalable video adaptation is introduced in Section 3. Section 4 presents the experimental results. Section 5 concludes this paper and discusses future directions.

2 Two-Dimensional Rate and Perceptual Quality Models

In this section, the impact of the spatial and SNR resolutions on the bit-rate and the perceptual quality is analyzed, based on which a 2D rate model and a 2D perceptual quality model are respectively derived.

2.1 Two-Dimensional Rate Model

Considering SNR and temporal scalabilities, we have proposed an analytical 2D rate model for H.264/SVC [10]. In this paper, this model is extended to the spatial domain where a product of a power function of the quantization step q and a power function of the spatial resolution index s are used, given as

$$R(q, s) = cq^{\alpha} s^{\gamma}, \tag{1}$$

where α and γ are both content-dependent model parameters. The values of α and γ characterize how fast the bit-rate reduces with the increase of q and how fast the bit-rate increases with the refinement of the spatial resolution, respectively. Usually a sequence with richer texture has larger absolute values of α and γ. Here s is computed through dividing the frame size of the current spatial resolution by the frame size of the lowest spatial resolution. In order to evaluate the model accuracy, sequences, with QCIF, CIF and 4CIF resolutions were encoded into 3 dyadic spatial layers using JSVM9.19.7 [11], respectively. Each spatial layer contained 5 quality layers. The base quality layer in a lower spatial layer was used to perform inter-layer prediction to avoid drifting at the decoder. The GOP (Group of Picture) size was set to 1 to avoid the effect of temporal scalability. 120 frames were encoded for each sequence. The difference of the quantization parameter (QP) between adjacent quality layers and adjacent spatial layers

were set to 5 and 6, respectively, following [12]. The QP of the base quality layer of the lowest spatial level was set to 38. The model parameters were obtained by minimizing the Root Mean Square Error (RMSE) between the actual and predicted bit-rates. The actual values and those predicted using (1) are plotted in Figure 1. It is clear that the proposed 2D rate model fits the actual data very well. Table 1 gives the used parameters and the model accuracy in terms of the RRMSE ($RRMSE = RMSE/R_{max}$, where R_{max} denotes the maximum bit-rate in the actual data) and the Coefficient of Determined (CoD), defined as:

$$CoD = 1 - \frac{\sum_i \left(X_i - \widehat{X_i}\right)^2}{\sum_i \left(X_i - \overline{X}\right)^2}, \tag{2}$$

where X_i and $\widehat{X_i}$ are the actual and the predicted values of the bit-rate, respectively, and $\overline{X}$ is the mean of all actual bit-rates shown in Figure 1. It is once again demonstrated that the proposed rate model is very accurate in prediction, where high CoD and small RRMSE values can be observed for all tested sequences. Specifically, the average CoD and RRMSE are 0.9892 and 2.81%, respectively. And as expected, the “City" sequence with the richest texture among the tested sequences has the largest absolute values of α and γ.

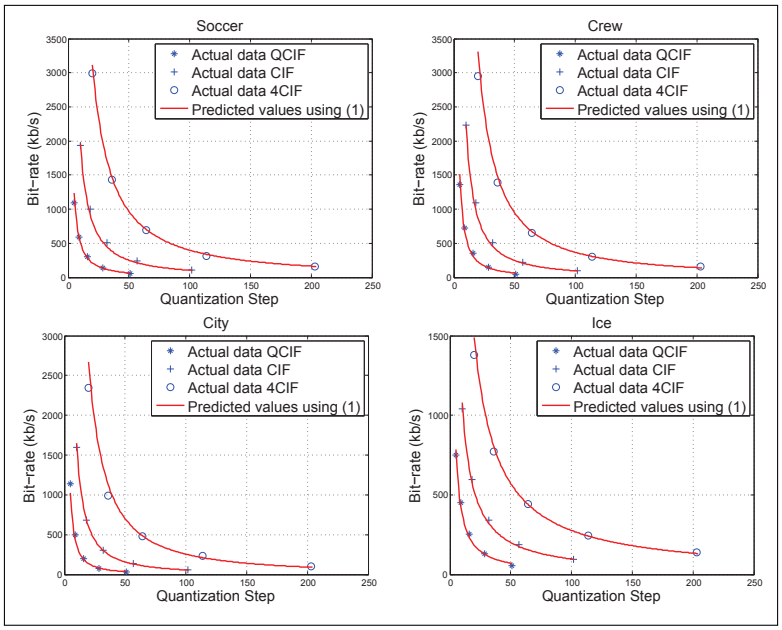

Figure 1: Actual bit-rates and predicted values using (1).

Table 1: The values of parameters in (1) and model accuracy

	Soccer	Crew	City	Ice	Ave.
$c^{\times 10^3}$	9.67	13.58	10.76	4.23	
α	-1.276	-1.365	-1.462	-1.048	
γ	0.970	0.965	1.078	0.756	
CoD	0.9958	0.9860	0.9792	0.9929	0.9892
RRMSE	1.71%	3.31%	3.97%	2.27%	2.81%

2.2 Two-Dimensional Quality Model

It has been widely acknowledged that the quality metrics of the PSNR and the Mean Square Error (MSE) do not correlate well with the perceptual quality. On the other hand, the subjective quality can be well captured by the Mean Opinion Scores (MOS) and Video Quality Metric (VQM) [13], at the cost of high complexity in testing and computations. Trading off between the complexity and the consistency with the human perception, the Structural Similarity (SSIM) [14] is used as the quality measure in this paper.

The SSIM measures the structural similarity as well as the luminance and contrast similarity between two images block by block. In this paper, the SSIM values have been measured with regard to different combinations of spatial and SNR resolutions, where the layers of lower spatial resolutions were up-sampled to 4CIF using a set of 6-taps filters provided by the JSVM. According to empirical observations, a logarithmic function in terms of the spatial resolution index and the quantization step is used to model the perceptual quality regarding different spatial and SNR resolutions, which is expressed as

$$QM_{ssim}(q, s) = a_0 + a_1 \ln q + a_2 \ln s + a_3 \ln q \ln s, \tag{3}$$

where a_0, a_1, a_2 and a_3 are all content-dependent model parameters. Here the second and third terms indicates the impact of the SNR and the spatial resolution on perceptual quality, respectively. The forth term models the joint impact of the SNR and the spatial resolution. The model parameters can be derived easily by minimizing the RMSE between the actual and predicted values. The actual and predicted qualities are shown in Figure 2. Table 2 lists the used parameters and the model accuracy in terms of the CoD and RRMSE. It can be concluded from Figure 2 and Table 2 that the proposed 2D perceptual quality model is very accurate in prediction with high CoD and small RRMSE values for all tested sequences.

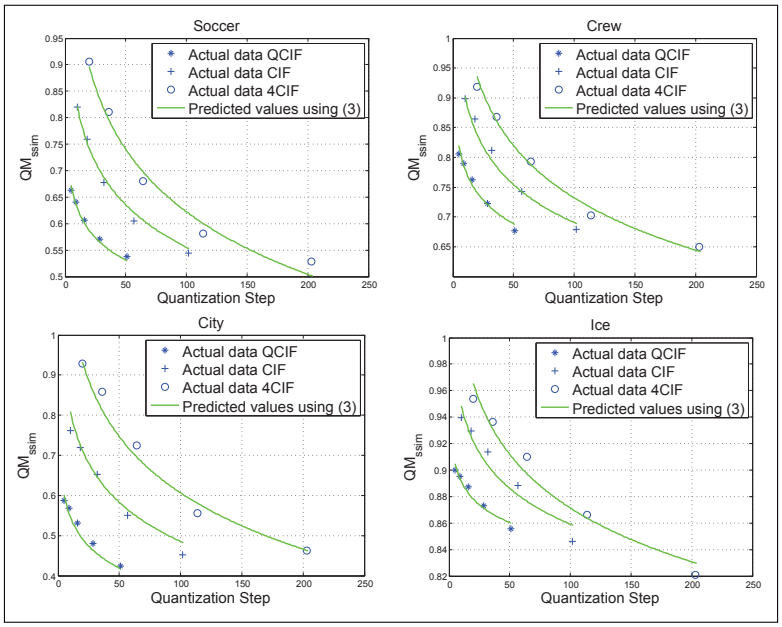

Figure 2: Actual qualities and predicted values using (3).

Table 2: The values of parameters (3) in and model accuracy

	Soccer	Crew	City	Ice	Ave.
a_0	0.7709	0.9112	0.7268	0.9395	
a_1	-0.0614	-0.0569	-0.0791	-0.0193	
a_2	0.2292	0.1454	0.2918	0.0739	
a_3	-0.0393	-0.0251	-0.0444	-0.0141	
CoD	0.9871	0.9842	0.9754	0.9561	0.9757
RRMSE	1.39%	1.11%	2.49%	0.8%	1.45%

3 Optimal Bitstream Adaptation for Scalable Video Using Proposed Models

The proposed models are applied to constrained bitstream adaptation for scalable video. Figure 3 provides a systematical view of the adaptation problem. For each video, a single full-resolution scalable bitstream is available at a server, where the bitstream will be adapted at a network proxy or gateway according to the user channel conditions and viewing preferences (i.e., the displayed spatial resolution). When a user requests the video from the server, the adaptor (at the proxy) will determine an appropriate bit-rate R_t for extraction based on the channel condition. Based on R_t and the user's settings of viewing preference (embedded in the user profile and sent to the adaptor), the adaptor determines the optimal set of spatial and SNR layers to extract, so as to provide the best perceptual quality.

For a given targeted bit-rate R_t and the required display spatial resolution, the adaptation problem can be formulated as the following constrained optimization problem:

$$\begin{aligned} \text{Determine } q, s \text{ to } & \text{ maximize } \ QM_{ssim}(q, s) \\ & \text{subject to } \ R(q, s) \le R_t \\ & \mathbf{U}(s)|s < S, \end{aligned} \tag{4}$$

where R_t and S denote the targeted bit-rate and the required display spatial resolution index, respectively. By $\mathbf{U}(s)|s < S$ it is indicated that up-sampling is executed if the extracted spatial resolution is less than the required display spatial resolution.

Assume that both the spatial resolution and the quantization step may take on any effective value. By setting $R(q, s) = R_t$, it can be obtained that

$$q = \sqrt[\alpha]{\frac{R_t}{cs^\gamma}}, \tag{5}$$

which describes the feasible q for a given s, to satisfy the rate constraint R_t. Substituting (5) into (3) yields

$$QM_{ssim}(s) = -\frac{a_3\gamma(\ln s)^2}{\alpha} + (a_3 \ln R_t/c + \alpha a_2 - a_1\gamma)\frac{\ln s}{\alpha} + a_0 + \frac{a_1 \ln R_t/c}{\alpha}. \tag{6}$$

Equation 6 is the achievable quality with different spatial resolutions under the targeted bit-rate R_t. Clearly, this function has a unique maximum, which can be derived by setting its first order derivative with respect to s to be zero. This yields

$$s = e^{(a_3 \ln (R_t/c) + \alpha a_2 - a_1\gamma)/2a_3\gamma}. \tag{7}$$

For any given R_t and S, we can solve (7) numerically to determine the optimal spatial resolution. Then using (5) and (6) the optimal quantization step can be determined, and the corresponding quality can be maximized. The parameters for the rate model, i.e., c, α and γ, can be easily derived from the bit-rates corresponding to several different (q, s)combinations using least square fitting. The quality model parameters, i.e., a_0, a_1, a_2 and a_3, can be derived using the least square fitting at the encoder, and then embedded in the header field of the video stream. Based on the simulations, only several bytes are required to represent those parameters which can be neglected compared to the actual video stream payload.

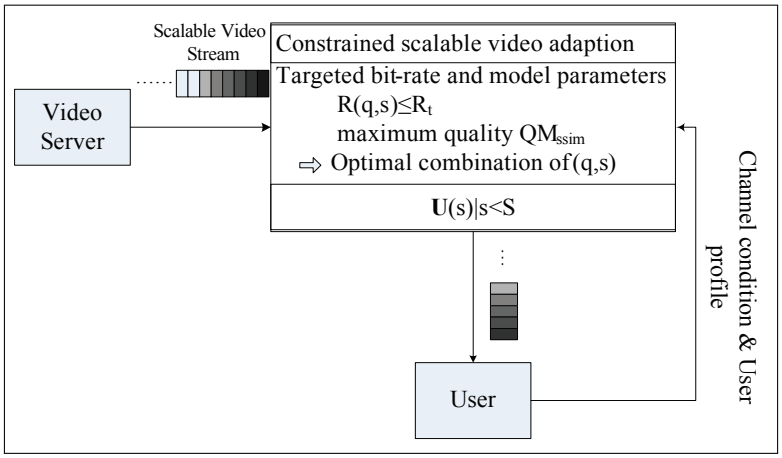

Figure 3: Constrained scalable video adaptation.

4 Experimental Results

The experimental results are presented in this section to evaluate the performance of the proposed extraction method. Firstly, assuming that the spatial resolution can be any positive values, and then the practical case where spatial resolutions to be discrete is considered.

4.1 Optimal Solutions Assuming q and s Taking Continuous Values

Assuming that both the spatial resolution and the quantization step can take continuous values. Figure 4 shows the optimal spatial resolution, quantization step and quality as functions of the targeted bit-rate R_t. As expected, as the targeted bit-rate increases, the optimal s increases while the optimal q reduces, and the achievable best quality continuously improves. Notice that the optimal s increases more rapidly for the “City" sequence than for the other sequences because of its richer texture. The up-sampling introduces more severe quality decrease than the quantization step. Therefore, under the bit-rate constraint, a larger spatial resolution with a larger quantization step is a better choice.

4.2 Optimal Solutions Under Dyadic Spatial Resolution Scalability

The H.264/SVC includes three profiles [15], i.e., the “Scalable Baseline" profile, the “Scalable High" profile, and the “Scalable High Intra" profile. While the latter two profiles support full spatial SVC scalability, the Scalable Baseline profile imposes some constraints to enable simplified application scenarios. For example, dyadic spatial scalability is provided in the baseline profile, where the scaling ratio of the width and height between adjacent spatial layers is equal to 2. From a practical point of view, it will be interesting to see the optimal combination of the spatial resolution and quantization step for different

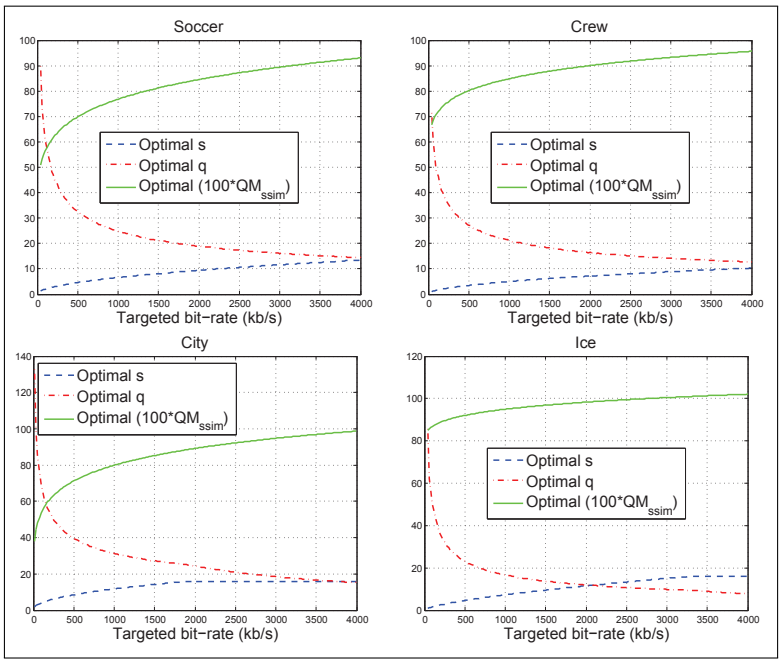

Figure 4: Optimal quantization step, spatial resolution index and the corresponding quality versus the targeted bit-rate by assuming the quantization step and the spatial resolution to be continuous.

targeted bit-rates under this SVC structure. To obtain the optimal solution for this SVC structure, we first determine the optimal s using (7), and then choose two spatial resolutions up and down around the value from the candidates. Finally, compute the quality using (6) corresponding to the two spatial resolutions and choose the spatial resolution that leads to a better quality.

The experimental results are shown in Figure 5. Because the spatial resolution can only increase in a discrete step, the optimal quantization step does not decrease monotonically with the bit-rate. Whenever the optimal s jumps to the next higher value, the optimal q first increases to meet the rate constraint, and then decreases while the optimal s is held constant, as the rate increases. Consistent with the previous results in Figure 4, for the “City" sequence with richer texture, the optimal s is 16 (corresponding to 4CIF) at a low bit-rates, whereas for other sequences, the optimal s stays 4 (corresponding to CIF) even at high bit-rates.

In practice, the SVC encoder with quality scalability does not allow the quantization step to change continuously. The finest granularity in quality scalability is a decrement of QP by 1 with each additional quality layer. This means that the quantization step reduces by a factor of $2^{-1/6}$ with each additional layer. In practice, much coarser granularity is typically used, with a decrement of QP by 3 to 6 typically [11]. When we limit the values of q to be discrete in addition to allow only dyadic spatial resolutions, a rate constraint cannot be always exactly met. However, one may still obtain the optimal q and s for any given constrains using the proposed scheme by estimating the bit-rate and quality of each combination in the finite set of feasible values for q and s.

5 Conclusions and Future Works

In this paper, a 2D rate model and a 2D quality model have been proposed, based on which a model-driven method for optimal bitstream adaptation is developed. Experimental results have demonstrated the accuracy of the two models. Using the proposed extraction method, the optimal combination of quality and spatial layers can be determined, providing the highest perceptual quality for a given bandwidth

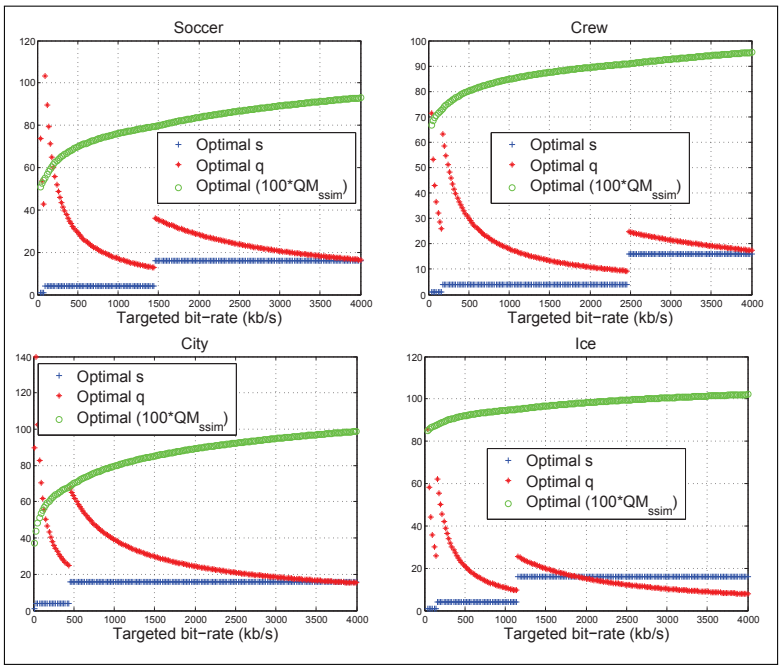

Figure 5: Optimal quantization step, spatial resolution index and the corresponding quality versus the targeted bit-rate by assuming that the q varies continuously and the spatial resolution takes QCIF/CIF/4CIF.

constraint and required display frame rate of the end user.

Future work may include an extension of the proposed models to three-dimension, taking temporal scalability into account. Moreover, the proposed models can be applied to other applications, e.g., advanced multidimensional rate control for video coding.

Acknowledgement

This work was supported by the National Science Foundation of China (60902052, 60902081), the Doctoral Fund of Ministry of Education of China (No.20096102120032), and the NPU Foundation for Fundamental Research (JC201038).

Bibliography

[1] H. Schwarz, D. Marpe, T. Wiegand, Overview of The Scalable Video Coding Extension of The H.264/AVC Standard, *IEEE Trans. Circuits Syst. Video Technol.*, Vol. 17, No. 9, pp.1103-1120, 2007.

[2] J. Reichel, H. Schwarz, and M. Wien, Joint Scalable Video Model 11 (JSVM 11), *Joint Video Team, Doc. JVT-X202*, 2007.

[3] I. Amonou, N. Cammas, S. Kervadec, S. Pateux, Optimized Rate Distortion Extraction With Quality Layers in The Scalable Extension of H.264/AVC, *IEEE Trans. Circuits Syst. Video Technol.*, Vol. 17, No. 9, pp.1186-1193, 2007.

[4] Ehsan Maani, Aggelos K. Katsaggelos, Optimized Bit Extraction Using Distortion Modeling in the Scalable Extension of H.264/AVC, *IEEE Trans. Image Process.*, Vol. 18, No. 9, pp.2022-2029, 2009.

[5] Yu Wang, Lap-Pui Chau, Kim-Hui Yap, Spatial Resolution Decision in Scalable Bitstream Extraction for Network and Receiver Aware Adaptation, *Proceedings of the 2008 IEEE International Conference on Multimedia and Expo*, pp.577-580, 2008.

[6] Y. Wang, Z. Ma, Y.-F. Qu, Modeling Rate and Perceptual Quality of Scalable Video as Function of Quantization and Frame Rate and Its Application in Scalable Video Adaptation, *Proceedings of the IEEE 17th Packet Video Workshop*, pp.1-9, 2009.

[7] Guangtao Zhai, Jianfei Cai, Weisi Lin, Xiaokang Yang, Wenjun Zhang, Three Dimensional Scalable Video Adaptation via User-End Perceptual Quality Assessment, *IEEE Trans. Broadcasting*, Vol.54, No.3, pp.719-728, 2008.

[8] A. Rahim, Z.S. Khan, F.B. Muhaya, M. Sher, M.K. Khan, Information Sharing in Vehicular AdHoc Network, *Int. J. of Computers, Communications and Control*, 5(5):892-899, 2010.

[9] A. Rusan, C.-M. Amarandei, A New Model for Cluster Communications Optimization, *Int. J. of Computers, Communications and Control*, 5(5):910-918, 2010.

[10] Junhui Hou, Shuai Wan, Fuzheng Yang, "Frame Rate Adaptive Rate Model for Video Rate Control," *Proceedings of the 2010 IEEE International Conference on Multimedia Communication*, pp.226-229, 2010.

[11] H.264 SVC Reference Software (JSVM 9.19.7) and Manual CVS sever, JVT, 2010 [Online]. Available: garcon.ient.rwth-aachen.de.

[12] Xiang Li, Peter Amon, Andreas Hutter, Andr Kaup, Performance Analysis of Inter-Layer Prediction in Scalable Video Coding Extension of H.264/AVC, *IEEE Trans. Broadcasting*, Vol. 57, No. 1, pp66-74, 2011.

[13] M. Pinson, S. Wolf. A New Standardized Method for Objectively Measuring Video Quality, *IEEE Transactions on Broadcasting*, Vol. 50, No.3, pp.312-322, 2004.

[14] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, Image Quality Assessment: From Error Visibility to Structural Similarity, *IEEE Trans. Image Process.*, Vol. 13, No. 4, pp.600-612, 2004.

[15] T. Wiegand, G. J. Sullivan, J. Reichel, H. Schwarz, M. Wien, Eds., Amendment 3 to ITU-T Rec. H.264 (2005) | ISO/IEC 14496-10:2005, Scalable Video Coding, 2007.