Quality modeling for the Medium Grain Scalability option of H.264/SVC

Michele Cesari, Lorenzo Favalli, Marco Folli
Universitá degli studi di Pavia
Via Ferrata 1
27100 Pavia, Italy
name.surname@unipv.it

ABSTRACT

Rate[1] vs. quality is a crucial trade off not only for efficient video coding and transmission but also for adaptive transmission strategies in wireless networks and/or congestion-prone networks. Scalable coders are well suited to tackle the time-varying capacities of these environments. In this paper we propose a semi-analytical model suitable for the medium grain scalable option of the H.264 standard and discuss the parameters influencing its performance. Results show it can effectively be used to represent the expected rate for different quality layers and thus its applicability to algorithms for resource optimization.

Keywords

H.264/SVC, Rate-Distortion models, Medium Grain Scalability

1. INTRODUCTION

The remarkable expansion of networks and multimedia applications, has made the video streaming services accessible in the wireless networks, peer to peer systems, and internet services. Due to the characteristics of these networks, the heterogeneous devices and the possible congestion due to traffic but also to error prone transmissions and capacity restrictions derived from user activity or mobility, it has been necessary to come up with flexible transmission systems, capable to adapt the video flow to the instantaneously available bandwidth. Scalable video coding offers a very efficient solution to this problem, allowing to receive the video at lower bitrates by accepting a progressive quality reduction. How quality is affected by stream truncation, is the subject of several works trying to define appropriate rate-distortion models which allow to obtain an estimation of the relationship between bitrate and distortion, using some video source statistics [1]. This allows to establish a decision policy within network distribution nodes, permitting to reach a compromise between the quality and the bitrate of the video when congestion occurs (e.g. [2][3]). In H.264/SVC, the most recent type of scalability introduced is the so called Medium Granular Scalability (MGS) [4], in which the enhancement layer packets are made in a way that they can be truncated at some given discrete points, and each supplemental information allows to increase the overall quality of the video. This is somehow in between the Fine Granular Scalability (FGS), in which the enhancement layer can be truncated arbitrarily, and the Coarse Grain Scalability which permits an increment of the overall quality of the video only in case the enhancement layer is entirely received and cannot be truncated in any way. Since it is possible to obtain several different rates by truncating the enhancement layer, it is interesting to somehow predict the loss of quality of each truncation. For this purpose, it is possible to use the Rate Distortion (R-D) theory to predict the obtained quality from some given rates. There are several ways to estimate a R-D curve, one of this is to build a mathematical formulation of the problem in order to correctly evaluate the distortion at a given rate (analytical model). Several different R-D models have been proposed in the past. In [7][8], authors have shown that classical models developed for non-scalable coders such as those presented in [5] and [6] cannot be extended to work with scalable-coded streams as they lose their accuracy. This is intrinsically due to the different statistical distribution of the *discrete cosine transform* (DCT) coefficients. While the residual DCT in the enhancement layer can be accurately represented by a mixture distribution (e.g. mixture Laplacian), this is not true for the coefficients in the base layer which is better described using a single distribution (Gaussian, Laplacian).

Dai et al., in [9], presented an analytical model, called Square Root Model, based on the statistical proprieties of the DCT coefficients. From this model, they derived a semi-analytical model, which can be used to evaluate the Rate-Quality performances from two empirical samples and from the source variance. The starting point of this article is to extend the semi-analytical model in [9] to estimate, without a high computational complexity, the Rate-Quality curves in case of video coded using the Medium Grain Scalability. In section 2, a description of the rate distortion models is presented. In section 3 we describe how to use the semi-analytical model for the MGS sequences. Finally, some results are reported in section 4.

[1]This work partially supported by Italian Ministry of research (MUR) under grants FIRB-RBIN043TKY "Software and Communication Platforms for High-Performance Collaborative Grid"

2. OVERVIEW OF RATE-DISTORTION MODELING

In this section we discuss some of the R-D models found in the literature with particular reference to the model proposed in [9] and applied to MPEG4 FGS since this is the starting point of our adaptation to H.264/SVC MGS.

The fundamental relationship between rate and distortion is that given by Shannon

$$D = \sigma^2 2^{-2R} \tag{1}$$

where D is the distortion measured as mean square error (MSE) and σ^2 is the source variance, assuming the source follows a gaussian distribution. This formula can be adapted to non-gaussian input distributions and becomes

$$D = \gamma\epsilon^2\sigma^2 2^{-2R} \tag{2}$$

where γ is the correlation coefficient and ϵ^2 is a scaling factor related to the actual distribution and is equal to 1.4 for a gaussian source, 1.2 for a laplacian one and is 1 for uniformly distributed source values. Distortion only depends on the statistical properties of the video source [9]. Based on formula 2, in [6] a model has been presented that relates both the bitrate and the distortion to the quantization step. Under the hypothesis of a uniform quantization step this model can be expressed as

$$D(\Delta) = \frac{\Delta^2}{\beta}, \quad R(\Delta) = \frac{1}{2} log_2(\frac{\epsilon^2\beta\sigma^2}{\Delta^2}) \tag{3}$$

where $\beta = 12$ for small values of Δ. Extension to larger values of Δ is obtained only by empirically adjusting the parameter β starting from either source statistics or measured values [6].

Another model for non-scalable coding schemes has been presented by Chiang et al. in [5]. In this work the source is by hypothesis laplacian with distribution $p(x) = \frac{\lambda}{2}e^{-\lambda|x|}$ and distortion D is expressed in terms of Mean Absolute Difference (MAD). Under these assumptions, the R-D relationship is expressed as

$$R = log\left(\frac{1}{\alpha D}\right) \tag{4}$$

which may be rewritten introducing the Taylor expansion of formula 4 to represent an operational model defined by

$$R = aD^{-1} + bD^{-2} \tag{5}$$

In this formula, parameters a and b must be derived using empirical data of the R-D curve. It has been shown ([7]) that these models are not longer accurate when applied to sequences encoded exploiting scalability. This happens for two main reasons. First of all, in the case of non-scalable coding techniques, it is assumed that the source's statistics can be represented by a single distribution. This is no longer true for the enhancement layer that can be better described using a mixture of distributions: since the enhancement layer only contains the *quality improvements*, the number of null DCT coefficients after quantization is significantly higher than in the base layer and the distribution shows a pronounced peak near zero [8].

A second reason is related to different use of the concept of quantization between base and enhancement layers. In the case of non-scalable coders, quantization step is mainly ruled by the $Qstep$ parameter, i.e., the scaling factor applied to DCT coefficients during quantization. This is different from what happens in the enhancement layer(s) of a scalable coder, where the quantization step is the element that determines the quality improvements with respect to the previous layer(s). As an example, in a FGS scalable coder based on *bitplane coding*, the quantization step Δ is given as

$$\Delta = 2^{(z-n)} \tag{6}$$

where z is the overall number of bitplanes and n is the index of the last received bitplane.

While additional layers are generated, the quantization step decreases and quality improves till all layers are available and $\Delta = 1$ corresponding to maximum quality. Based on these considerations, Dai et al. in [9] introduce an analytical model for scalable streams named *Square Root Model*. The hypothesis is that the source can be described by a Mixture Laplacian Distribution

$$p(x) = q\frac{\lambda_0}{2}e^{-\lambda_0|x|} + (1-q)\frac{\lambda_1}{2}e^{-\lambda_1|x|} \tag{7}$$

where x is the residual of the DCT, q is the probability that the value belongs to either component of the distribution and λ_0 and λ_1 are the $s$hape parameters of the two Laplacian distributions. Following [8], we have

$$D(\Delta) \approx \frac{2\alpha\xi}{1 - e^{b\Delta}} \tag{8}$$

where ξ is obtained from:

$$\xi = e^{b(\Delta-1)}(\frac{(\Delta-1)^2}{b} - \frac{2(\Delta-1)}{b^2} + \frac{2}{b^3}) - \frac{2}{b^3} \tag{9}$$

Introducing 8 in the following expression

$$PSNR = 10log_{10}(\frac{255^2}{D}) \tag{10}$$

and using a polynomial approximation of grade 2 for bit-

plane z a simple model for the PSNR is obtained given as

$$PSNR(z) = g_1 z^2 + g_2 z + g_3 \quad (11)$$

where g_1, g_2, e g_3 are arbitrary constants. In [9] authors also prove a simple model of the corresponding bit rate

$$R(z) = a_1 z^2 + a_2 z + a_3 \quad (12)$$

where a_1, a_2, e a_3 are parameters that must be estimated from empirical data.

By combining equations 12 and 10 it is possible to define a semi-analytical square root model (SQRT) as

$$PSNR(R) = AR + B\sqrt{R} + C \quad (13)$$

where A and B must be estimated via curve fitting having at least two empirical samples, and $C = 10log_{10}(\frac{255^2}{\sigma^2})$ where σ^2 is the source variance. In the quality domain this is a generalization of the classical definition 1. It has been shown in [9][7] that this model, tested on sequences coded using the MPEG4 FGS coder, is accurate and gives better results than those obtained using more traditional models.

3. PROPOSED SCHEME

Starting from the SQRT model described in the previous section, in this paragraph we adapt it to estimate the quality of a H.264/SVC MGS sequence from its bitrate. The key of the present work is the definition of variance of a source which is necessary to determine the parameter C of equation 13.

The most common definition of *source* usually refers to the DCT coefficients generated by the coder. This information is unfortunately available only almost at the end of the coding chain. This can be considered to be too late for an appropriate use of the model which would ideally require to be used to determine the quality of the video (given the rate) before the coding process since this estimate would allow to modify the coding parameters.

The idea to be exploited is then to assume as the source the raw sequence of luminance pixels and use a *complexity* measure of the sequence calculated before coding rather than the distribution of the DCT coefficients.

3.1 Quantization Scheme

Assumed that the main source of distortion in a coded video sequence is given by the quantization error [1] we first review the quantization process in the medium grain scalability (MGS) option of the H.264 standard.

In a scalable coder such as the FGS coder of the MPEG4 standard following the *bitplane coding* approach, the coded DCT coefficients are transmitted bit-by-bit from the *Most Significant Bitplane* (MSB), to the *Least Significant Bitplane* (LSB) with a quantization step given by equation 6. In H.264/SVC a different scheme is used named *Progressive Refinement Slice* and the quantization step is given by the difference among the slices of DCT coefficients.

In our work we assume that the quantization scheme is fixed so that formula 6 still holds when the bitplanes z are replaced by the slices s of DCT coefficients

$$\Delta = 2^{(s-n)} \quad (14)$$

In 14, s is the maximum number of slices as determined from the MGS vectors and n is the index of the last slice received.

3.2 H.264/SVC scalability

Before introducing the model we briefly review the concept of scalability in H.264/SVC. MGS scalability allows to choose a partitioning of a block of 4x4 DCT coefficients as specified by so called *MGSVectors*. To each packet of a slice is then assigned a quality identifier (Q_{id}) which represents a priority inside the bitstream. This type of scalability may lead to a number of layers that varies from coarse grain scalability to fine grain scalability.

This type of scalability can be combined with a time scalability to increase the flexibility of the scheme. Time scalability is obtained assigning different priorities to predicted pictures so that the number of temporal layers is a function of the size of the group of pictures (GOP) so that for each quality layer we may have $n = log_2(GOP)$ temporal layers. Similarly to Q_{id}s, we now introduce the temporal identifiers (T_{id}) for each temporal layer.

Figure 1 shows a Rate-Quality curve for the sequence Football determined empirically.

Looking at the empirical curve, we may note that the points corresponding to the quality layers exhibit a different behavior than the temporal layers. From this observation (as well as from those from other sequences coded) we derive the suggestion that the modeling process should be performed in two steps:

1. a first step to determine the Q_{id}s
2. a second step to match the T_{id}s given the Q_{id}s found in the first step

Since the two steps refer to two different sources of scalability, after several experiments, we decided to perform the two steps using different complexity measures in each step as they refer to different properties of the sequence: spatial information and temporal information. Since we foresee a scenario in which these models must provide "on the fly" information to scheduling algorithms, a further constraints is to find suitable complexity measures that can be easily extracted from data already available at the coder or that can be easily derived. For the spatial complexity we have selected the PSNR of the base layer, which is coded at a

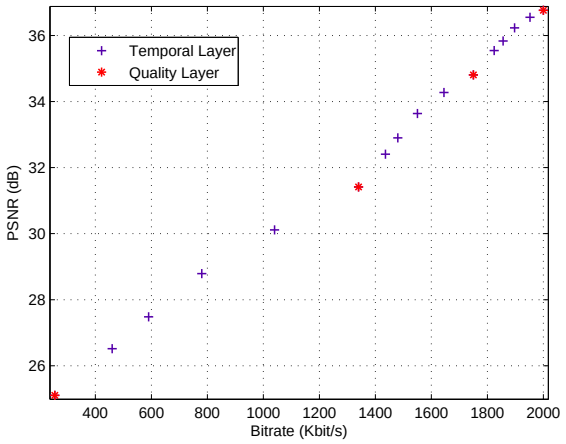

Figure 1: Empirical Rate-Quality curve of Football with 4 Quality layers and 4 Temporal Layer between 200Kbit/s and 2Mbit/s

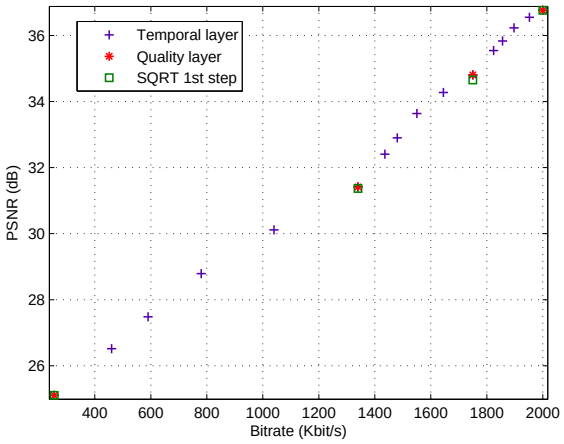

Figure 2: Estimation of quality layers by SQRT model 1st step

fixed rate, since this measure can easily be used to represent how difficult is the encoding process and consequently is an indirect measure of its complexity. This data is anyway required for the curve fitting procedure. For the time complexity, we have used the the simple differences among pixels in different frames and then determine the variance of this measure which is the used to determine the parameter C of the SQRT model as $C = 10log_{10}(\frac{255^2}{\sigma^2})$ where σ^2 is the time variance just defined.

Finally, we assume that two points are available from real measurements: namely the one corresponding to the base layer and the other to the full sequence. Since two empirical points are available, it is easy to implement also another model: the *linear* model, simply obtained with linear interpolation between the two available points. Since the Rate-Quality curves are built with a two step algorithm (first the points of the quality layers, then the points that belong to the temporal layers), it is possible to create a new model by hybridize the SQRT model and the lineal model. We then obtain a *semi-linear* model, in which the first step is made using the SQRT model, and the second step with the linear model.

Sequence	T. Variance [dB]	PSNR@200Kbit/s
Football	20	25.1
Mobile	22	27.4
Soccer	22	31.4
Harbour	25.2	27.9
Tempete	26.5	30.2
Foreman	27.5	35.2
Crew	28.8	33.2
News	34	39.9
Mother and D.	36.8	41.2
Waterfall	37	35.6

Table 1: Spatial and temporal complexity values.

Table 1 reports the complexity for ten well known sequences. It may be noted that lower values correspond to complex sequences. Furthermore, time complexity and PSNR appear to be independent one another.

The value of the complexity changes the convexity of the curve. For this reason we decided to extend our experiments to a large set of sequences directly captured from real TV broadcasts and compared the two-steps SQRT described so far with linear approximation, poly-line and hybrid models. Results are given in the next section.

4. RESULTS

We now present the results obtained by applying the two-steps modeling approach described in the previous section and discuss the precision of the model. At the end we will see that some modifications will be required depending on the characteristics of the sequence as characterized by means of the parameters specified above.

The first round of tests has been conducted coding 90 frames of a large set of sequences given in table 2: this phase has also allowed us to gain some experience on the MGS coding process. The simulation was conducted with to the following coding parameters:

Sequence	1st Step Error[dB]
Football	0.0497
Mobile	0.0935
Harbour	0.1420
Tempete	0.1211
Crew	0.1995
News	0.1962
Mother and D.	0.1043
Waterfall	0.0754

Table 2: 1st Step SQRT Errors

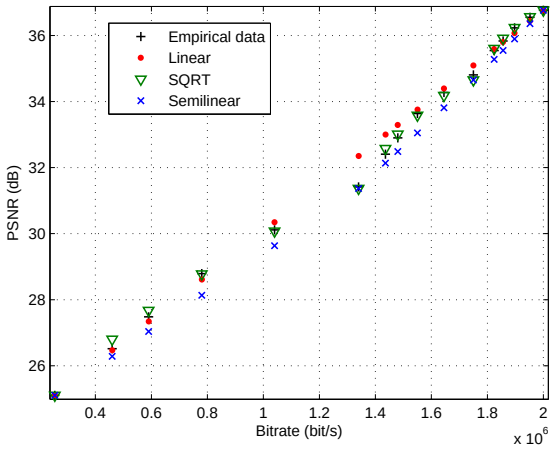

Figure 3: SQRT, Linear and Semilinear Models applied to Football sequence

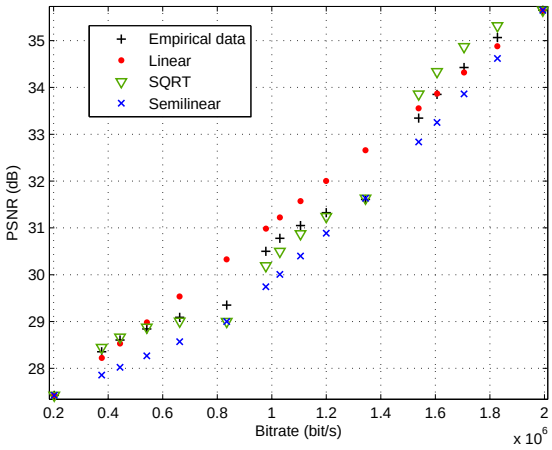

Figure 4: SQRT, Linear and Semilinear Models applied to Mobile sequence

- base layer fixed at 200 Kbit/s
- full stream 2 Mbit/s
- rate constrained coding setting
- MGS coding vector composed of three sectors following a 4-4-8 split of the 16 DCT coefficients
- GOP = 16 frames

We have selected sequences of 90 frames in order to avoid scene changes or high complexity variances, so we can evaluate the parameter C as the mean of the C value of each single frame. This assumption allows to reduce the complexity of the model, as we can evaluate only one time the C value.

These setting allow four quality layers with a slight emphasis on the lower frequencies and four temporal layers for each quality layer for a total of sixteen layers as shown in figure 1. In figure 2 we show an example of how the first step is performed in order to find the value of the quality id layers. For the sake of comparison, also the *linear* model and the *semilinear* model are used to evaluate the rate-quality curves in order to evaluate which method gives the best performances in every different situations. Parameters A and B in formula 13 have been estimated by curve fitting using the *Non Linear Least Square Data Fitting* algorithm in Matlab. Estimation of the quality layers is given in table 2 and can be considered as being very good and the error never exceeds 0.2 dBs. On the contrary, the SQRT model doesn't provide similar good results when applied in the second stage to describe the points corresponding to the temporal layers. For this reason, in table 3 we compare results with those obtained with a linear and a semi-linear model. The semi-linear model in particular provides better results for some sequences. To better understand if the precision of the different models can be related to some parameters characterizing the specific sequence, we have extended the test to a large number of sequences captured from normal broadcasting transmissions.

Sequence	SQRT	Semilinear	Linear
Football	0.0827	0.3011	0.2213
Mobile	0.1993	0.3873	0.3415
Harbour	0.1733	0.4381	0.28
Tempete	0.1764	0.3870	0.3828
Crew	0.3475	0.1173	0.2122
News	0.4137	0.2354	0.7259
Mother and D.	0.2950	0.1328	0.32
Waterfall	0.2451	0.1204	0.1902

Table 3: 2st Step Errors of SQRT, Semilinear and Linear models

Given that the parameters A and B are already determined in the first stage of the algorithm, we focus our attention on the value of parameter C: we recall that this parameter represents the temporal complexity measured as the luminance variance between two frames and can be derived *before* coding the sequence. Results for this extensive campaign are provided in table 4. We see that for sequences with a value of C smaller than 15 or larger than 45 neither of the models is accurate. All other values of C can be divided in different ranges to determine which of the models is best suited for that sequence. The SQRT is preferable when $18 < C < 27.5$ (examples in figures 3 and 4) while for $15 < C < 18$ and $27.5 < C < 45$ (examples in figures 5 and 6) the semi-linear model should be preferred.

C<15	Inaccurate
15<C<18	Semilinear
18<C<27.5	SQRT Model
27.5<C<45	Semilinear
C>45	Inaccurate

Table 4: Models working range

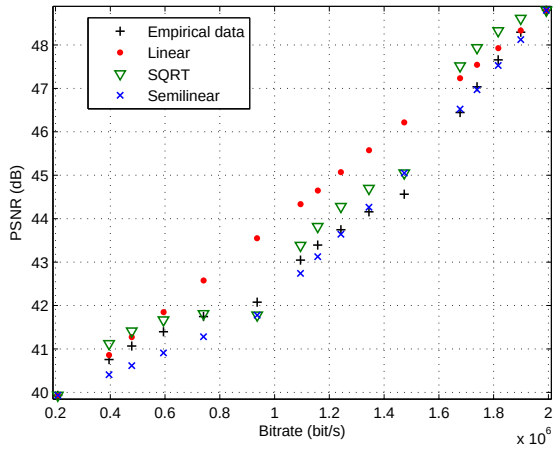

Figure 5: SQRT, Linear and Semilinear Models applied to News sequence

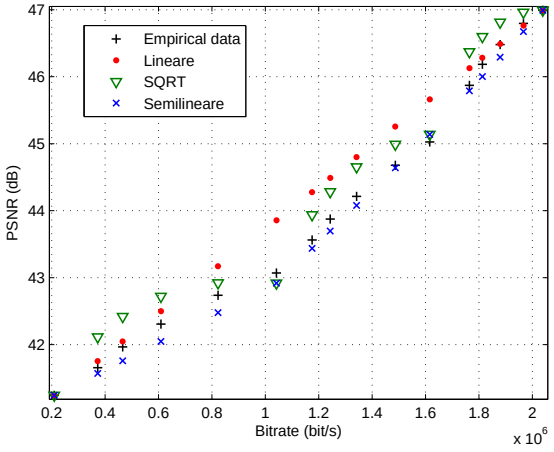

Figure 6: SQRT, Linear and Semilinear Models applied to Mother and daughter sequence

It must be noted that extreme values of C have been obtained using non-realistic sequences made of gaussian noise and "still" sequences, since these values can be obtained only with *particular* characteristic of the video sequence, that are difficult to be find in the common available sequences.

5. CONCLUSIONS

In this work we have implemented a simple quality model to estimate the PSNR in sequences coded using H.264/SVC MGS once the rate is given. To be implemented, the model only requires knowledge of the two extreme points of the enhancement part of the stream (the quality of the base layer and of the full stream) and of a measure of the temporal complexity calculated over the raw luminance sequence. As a first step, the intermediate quality points (2 in our case) can be derived via curve fitting. In a second step the temporal layers are derived. Results show that the proposed approach can be accurate although best results are obtained changing the type of model used. Fortunately, the choice can be made based upon a parameter that can be easily derived from the sequence before coding. This parameter must not be computed every frame but could be considered constant for a whole "scene".

6. REFERENCES

[1] H. M. Hang, J. J. Chen, "Source Model for Transform Video Coder and Its Applications Part I: Fundamental Theory," *IEEE Transaction Circuits and Systems for Video Technology* vol. 7, no. 2, April 1997.

[2] E. Setton, X. Zhu, and B. Girod, "Congestion-optimized scheduling of video over wireless ad-hoc networks,"Circuits and Systems, 2005. ISCAS 2005. IEEE International Symposium on, vol. 4, pp. 3531-3534, Kobe (JP), May 2005.

[3] M. Roitzsch, M. Pohlack, "Video quality and system resources: Scheduling two opponents," *Elsevier J. Vis. Commun. Image R. 19, 2008, pp. 473-488.*

[4] H. Schwarz, D. Marpe, T. Wiegand, "Overview of the Scalable Video Coding Extension of the H.264/AVC Standard" *IEEE Transaction Circuits and Systems for Video Technology*, vol. 17, no. 9, September 2007.

[5] T. Chiang, Y. Q. Zhang, "A new rate control scheme using quadratic distortion model," *IEEE Transaction Circuits Systems for Video Technology*, vol. 7, no. 1, pp. 246-250, February 1997.

[6] H. M. Hang, J. J. Chen, "Source model for transform video coder and its applications part I: fundamental theory," *IEEE Transaction Circuits Systems for Video Technology*, vol. 7, no. 2, pp. 197-211, April 1997.

[7] C. H. Hsu, M. Hefeeda, "On the accuracy and complexity of rate-distortion models for fine-grained scalable video sequences," Technical report TR 2006-12, Simon Fraser University

[8] M. Dai, D. Loguinov, H. Radha, "Statistical analysis and distortion modeling of Mpeg-4 FGS," *IEEE ICIP*, pp. 301-304, September 2003.

[9] M. Dai, D. Loguinov, H. Radha, "Rate-distortion analysis and quality control in scalable Internet streaming," *IEEE Transaction on Multimedia* vol. 8, no. 6, pp. 1135-1146, December 2006.